European Journal of Taxonomy 934: 93–186
https://doi.org/10.5852/ejt.2024.934.2529

ISSN 2118-9773
www.europeanjournaloftaxonomy.eu
2024 · Cardini A.

Monograph

urn:lsid:zoobank.org:pub:603008D3-E60F-4A15-9DD7-C752FFC08C4C

A practical, step-by-step, guide to taxonomic comparisons using Procrustes geometric morphometrics and user-friendly software (part B): group comparisons

Andrea CARDINI

Dipartimento di Scienze Chimiche e Geologiche, Università di Modena e Reggio Emilia,
Via Campi, 103 - 41125 Modena, Italy.
School of Anatomy, Physiology and Human Biology, The University of Western Australia,
35 Stirling Highway, Crawley WA 6009, Australia.
E-mails: alcardini@gmail.com, andrea.cardini@unimore.it
urn:lsid:zoobank.org:author:A63AF653-521E-4EE0-9623-7B9273DF7D0A

Table of contents

Abstract 93
Introduction 94
Sexual dimorphism (SDM) 94
Interspecific comparisons 96
Material and methods 99
Methods, results and discussion subdivided by study question 99
B1) Sexual dimorphism within species 99
B2) Sexual dimorphism in relation to species differences 107
B3) Pairwise tests of species mean differences 122
B4) Species discriminant analysis (DA) 125
B5) Summary and visualization of species shape differences 132
B6) Relationship between shape and size within and across species 145
B7) Species comparisons of the magnitude of size and shape variance 159
Conclusions 164
Guidelines: aims, readership, limitations and further readings 164
Beyond free user-friendly software: why learning R 166
Main steps and results, in brief 166
What is new? From a broader assessment of measurement error to the sensitivity of results to sampling 169
Final remarks on GMM in taxonomy 170
Acknowledgements 170
References 171
Appendix B: Checklist of the main steps 184

Abstract. In this second part of the study, using a ‘clean’ dataset without very low precision landmarks and outliers, I describe how to compare mandibular size and shape using Procrustes methods in adult North American marmots. After demonstrating that sex differences are negligible, females and males are pooled together with specimens of unknown sex and species are compared using a battery of tests, that estimate both statistical significance and effect size. The importance of allometric variation and its potential effect on shape differences is also explored. Finally, to provide potential clues on founder effects, I compare the magnitude of variance in mandibular size and shape between the Vancouver Island marmot (VAN) and the hoary marmot, its sister species on the mainland. In almost all main analyses, I explore the sensitivity of results to heterogeneous sample size and small samples using subsamples and randomized selection experiments. For both size and shape, I find a degree of overlap among species variation but, with very few exceptions, mean interspecific differences are well supported in all analyses. Shape, in particular, is an accurate predictor of taxonomic affiliation. Allometry in adults, however, explains a modest amount of within-species shape change. Yet, there is a degree of divergence in allometric trajectories that seems consistent with subgeneric separation. VAN is the most distinctive species for mandibular shape and mandibular morphology suggests a long history of reduced variation in this insular population. Geometric morphometrics (GMM) is a powerful tool to aid taxonomic research. Regardless of the effectiveness of this family of methods and the apparent robustness of results obtained with GMM, however, large samples and careful measurements remain essential for accuracy. Even with excellent data, morphometrics is important, but its findings must be corroborated with an integrative approach that combines multiple lines of evidence to taxonomic assessment. The analytical protocol I suggest is described in detail, with a summary checklist, in the Appendix, not to miss important steps. All the analyses can be replicated using the entire dataset, which is freely available online. Beginners may follow all the steps, whereas more experienced researchers can focus on one specific aspect and read only the relevant chapter. There are limitations, but the protocol is flexible and easy to improve or implement using a programming language such as R.

Keywords. Allometry, ANOVA, discriminant analysis, founder effect, interspecific variation, island population, mandible, North American marmots, group differences, sexual dimorphism.

Cardini A. 2024. A practical, step-by-step, guide to taxonomic comparisons using Procrustes geometric morphometrics and user-friendly software (part B): group comparisons. *European Journal of Taxonomy* 934: 93–186. https://doi.org/10.5852/ejt.2024.934.2529

Introduction

This brief section concludes the main introduction of part A. I focus, now, on the details of the study outline for the most important analyses, which are the group comparisons. As in A, after the Introduction and before some general considerations in the Conclusions, the methods, results and discussion specific to each research question are organized in separate 'chapters'. I refer the reader to part A, and its Appendix, for more information on introductory topics, such as the study background; samples, digital images, and the landmark configuration; power and sample size; statistical testing and common assumptions in group comparisons. As in A, most of the time I will be talking about species differences and interspecific analyses, but the reasoning is similar in within-species analyses (e.g., geographic populations) or in comparisons including different taxonomic levels (species, subspecies etc.), as it may happen when evolutionary boundaries are fuzzy and taxonomic status uncertain (Zachos 2016).

Sexual dimorphism (SDM)

SDM is not a main aim in a taxonomic study. However, the assessment of SDM is necessary to decide whether or not to pool the data regardless of sex in the interspecific comparisons. Age variability is also important to consider, but, in this, as in most taxonomic studies, I assume that the researcher is using adults (see part A for the distinction between absolute and biological age). An age class other than adults can be used, if age discrimination in young is accurate and specimens are available. In fact, being able to run a taxonomic comparison at multiple levels (age groups, ontogenetic trajectories etc.) is desirable, as it provides evidence that better represents the life history of a lineage. Yet, most of the time this is difficult, because adequate samples, accurately representing different age classes, are very hard to obtain. In marmots, as typical in mammals, ontogenetic changes are large, which makes controlling for age a compulsory choice to avoid a strong confounding factor when taxa are compared. However, if in doubt, as with sex, within-species age differences in size and shape can also be tested using age as a factor in a one-way analysis of variance (ANOVA or, if multivariate, MANOVA[1]). In the ANOVA, age can also be analysed together with taxon (e.g., Klenovšek & Kryštufek 2013) and sex (e.g., González *et al.* 2002), although this makes the multifactorial design (two- or three-way) definitely more complex. In all instances, based on the ANOVA results, a researcher will decide whether or not to pool within-species subgroups (age, sex etc.).

The assessment of the magnitude and significance of SDM is part of the preliminary analyses. It is another type of comparison of group mean differences, which mostly employs the same methods as for the interspecific comparisons. Also, whether SDM is small enough to be considered negligible, it is better verified in relation to between species variation (Neff & Marcus 1980). For instance, if patterns of SDM are similar across species and the differences negligible in relation to interspecific variation, pooling sexes in taxonomic comparisons is unlikely to lead to inaccurate results, even if SDM is significant. Pooling females and males in gonochoric animals, if done appropriately, has the advantage of increasing sample size and, therefore, statistical power.

As with other analyses (part A), unless specified otherwise, all main tests of SDM (B1–2) and species differences (B2-3-4) will be done in parallel on centroid size (CS) and shape. For the assessment of SDM, specifically I will test:

B1) *SDM within species.* Sex differences are tested first one species at a time. This is complimentary to B2, which examines SDM in relation to interspecific differences. The statistical model is simpler in B1 than in B2. However, in B1 the same hypothesis is tested multiple times, potentially inflating the rate of type I errors (i.e., false positives, where differences are claimed that are not present). Thus, B1 requires a cautious interpretation (see Methods and Discussion). B1 is somewhat optional, since SDM is also tested in B2. However, B1 provides details that are not in B2 (e.g., whether all species show SDM or just some) and might help especially with unbalanced samples (i.e., heterogeneous sample sizes) that make multifactorial ANOVAs more difficult to interpret and potentially less accurate.

B2) *SDM in relation to species differences.* The second and main test of SDM is a species by sex two-way analysis of variance (ANOVA, for size, and, because it is multivariate, MANOVA for shape), testing the overall significance of species and sex, but also the interaction of these two factors. The interaction represents the "dependence of the effect of one factor on the level of another factor" (Sokal & Rohlf 2009: 195). More precisely, the species by sex interaction is assessing if SDM is similar (non-significant interaction) in magnitude and direction across all marmot species. For instance, for size, if there is SDM, with males larger than females, a non-significant interaction indicates that, on average, a male mandible is always larger than a female one by approximately the same amount (e.g., ~ 10% larger) in all species. In contrast, if SDM varies depending on the species (significant interaction), pooling sexes in interspecific tests leads to likely inaccuracies, as sex and species differences might become mixed up in the comparisons.

[1] In GMM, shape analyses are always multivariate and, thus, the ANOVA becomes a MANOVA. In part A, I made a very limited use of the word MANOVA, because the multivariate ANOVA was employed only in the context of shape ME using MorphoJ, where it is simply called *Procrustes ANOVA*. However, as briefly said in A1, MorphoJ's ANOVA is using all shape variables together and it is, therefore, a MANOVA. In this paper, for brevity, I will follow MorphoJ's convention and use ANOVA as a general term, when I refer to both univariate and multivariate analyses run in parallel on CS or shape variables. In contrast, if a description is specific to shape, I will be using the term MANOVA. Regardless of this convention; however, I stress that all shape analyses are and must be multivariate. Shape is by definition multivariate and analysing shape variables one at a time is almost always meaningless (Rohlf 1998; Adams *et al.* 2011).

Interspecific comparisons

The hierarchical ANOVA in part A and the two-way ANOVA in B2 have already tested the overall significance of interspecific differences. However, the main aim of those ANOVAs was assessing ME (measurement error) and SDM, respectively. If ME is shown to be negligible (A1), after SDM is tested (B1–2) all further analyses can exclusively focus on taxonomic differences. Because in marmots SDM is negligible (see ME results in part A and results of sections B1 and B2 in this paper), interspecific analyses will use pooled-sex samples. If SDM was large and significant, in contrast, the analyses would be the same, but should be run separately for females and males. All main tests will be done using all species and individuals (outliers excluded). However, to explore the sensitivity of results to small and heterogeneous samples, as well as to potentially unfavourable p / N ratios (with p being the number of variables and N the sample size), I will also replicate several analyses using subsamples or subsets of variables, as explained in the corresponding subsections.

Comparisons in this study follow an established design in taxonomic research using morphometrics. The rationale of the protocol is simple:

a) decide the taxonomic level of the comparison (species, in my case);

b) find comparable, homogeneous groups within each taxon (e.g., adults or, when SDM is large, adults of the same sex);

c) test their overall differences using an ANOVA (MANOVA, for shape) or equivalent methods;

d) explore, describe and summarize the magnitude and patterns of these differences.

To this basic design, researchers may add further steps to investigate specific issues. Sometimes, one might want to explore the relationship between shape and size (allometry (Klingenberg 1998, 2016, 2022)) to better understand its impact on taxonomic differences. Although less common than allometric analyses, comparisons of the amount of phenotypic variability (also called 'within-species disparity') in size and shape across taxa are also interesting in taxonomy. This type of disparity analysis may provide clues on population bottlenecks that have reduced genetic and, as a consequence, phenotypic variance. Disparity analyses have long been used in macroevolutionary studies, and especially in palaeontology, to quantify the magnitude of phenotypic evolutionary divergence (Foote 1997). Using the North American marmot mandibles, I will exemplify both allometric analyses and the test of differences in the magnitude of variance in size or shape between taxa.

As in part A, all shape analyses must be multivariate and use the entire block of shape variables together (all shape coordinates or all PCs of the Procrustes shape coordinates). Readers might find papers, especially from the nineties and early 2000, where analyses are performed also on landmark shape coordinates one at a time or on partial warps/uniform components one by one or in subsets. This, as well as testing shape PCs one at a time, is a mistake. For 'historical reasons' (Rohlf 2015), however, this type of univariate analysis of partial warps is still part of the output of several of the programs of the TPS Series. None of these tests is meaningful, as explained in Rohlf (1998, 2015) and in simpler terms in several of my own papers (Viscosi & Cardini 2011; Cardini 2020a; Cardini & Verderame 2022). Even just separating uniform and non-uniform shape variation rarely makes sense in biology. Uniform changes are "transformation for which parallel lines remain parallel", such as, for example, squares transformed into parallelograms or "cubes into parallelopipeds" (Marcus *et al.* 1993: 532). Non-uniform variation, quantified by a full set of partial warps, in contrast, concerns localized differences, that vary in magnitude and pattern within a structure. I am not aware of any biological application where a sound, convincing explanation for using subsets of shape variables (with the exception of PCs – see below) is given. Thus, in the software (and in publications), one should not consider results of per-landmark analyses or tests of partial warps one at a time and the like. In contrast, the focus must be on fully multivariate tests using all shape variables. These tests will produce identical outcomes using either all Procrustes shape coordinates or all partial warps, including the uniform component, or all shape PCs. Later, however, I will suggest to preferentially use PCs (all of them, if possible) especially in programs such as PAST, which are not specific to GMM and may have problems with matrices with highly collinear and, thus, redundant variables, like the Procrustes shape coordinates.

As an exception to the general rule of including all shape information in the statistical analysis of Procrustes shape data, dimensionality reduction using a subset of the first PCs of shape can also be justified sometimes. Analyses in a subspace of the total Procrustes shape space are usually done when p/N is large. However, prior to the test, a researcher should demonstrate that the chosen subspace (i.e., that specific number of first PCs) captures the large majority of shape variation relevant to the study question. This is, nonetheless, a compromise that does not exclude the possibility of losing a small but interesting amount of information 'hidden' in the discarded higher order PCs. The need of reducing shape dimensionality is, in itself, an indication that p/N is large and, therefore, potentially problematic (Cardini *et al.* 2019; Rohlf 2021). On the relationship between dimensionality, PCs eigenvalues (i.e., variances) and informative content in geometric morphometric data, readers can find a stimulating, but rather technical discussion in the context of information theory in O'Keefe e*t al.* (2022).

Beginners must be aware also that the term 'relative warp' is still common in GMM in place of the simpler 'principal component'. Almost all the time in biological research, despite differences in computational details, a relative warp analysis (Rohlf 1993) is simply equivalent to a PCA on the Procrustes shape coordinates (Zelditch *et al.* 2004). In rare cases, a relative warp analysis becomes a modified PCA that puts more or less weight on large or small scale changes, but this weighting is based on bending energy, a non-biological model used in morphometrics as "a metaphor borrowed ... from the mechanics of thin metal plates" (https://www.sbmorphometrics.org/glossary/gloss1.html). As with analyses of subsets of partial warps, it is hard to provide a justification for this type of rescaling in biological applications. Thus, since the proliferation of redundant terminology creates confusion, I urge to completely avoid terms such as 'relative warp' or 'relative warp analysis' and opt for the simpler PCA.

With the caveats provided in these last two paragraphs (i.e., avoiding univariate or bivariate analyses of shape variables etc., and replacing redundant terminology), the design I follow to test taxonomic differences is largely taken from a series classical studies (Rohlf *et al.* 1996; Corti & Rohlf 2001; Frost *et al.* 2003), which inspired many other taxonomic papers in GMM (e.g., Cardini 2003, 2022; Amaral *et al.* 2009; Bogdanović *et al.* 2009; Cardini & Elton 2009; Schutz *et al.* 2009; Ivanović *et al.* 2009; Gidaszewski *et al.* 2009; Machado & Hingst-Zaher 2009; Herler *et al.* 2010; Elton *et al.* 2010; Berns & Adams 2013; Salvidio *et al.* 2015; de Moura Bubadué *et al.* 2016; Meloro *et al.* 2017). Specifically, the analyses I will perform on the marmot mandible Procrustes data are:

B3) *Pairwise tests of species mean differences.* Pairwise tests are complimentary to an ANOVA testing species differences. I do not perform the ANOVA, because a two-way species by sex ANOVA has already been done (B2). Pairwise comparisons, after simultaneously testing species with the ANOVA, represent a type of 'post-hoc' test, whose aim is to find which pairs of species specifically differ and how much. It could be that all species differ, only some or even just one specific pair. As with B1, because the same hypothesis is tested multiple times, post-hoc tests should be interpreted with caution, as they might inflate type I errors (i.e., the rate of 'false positives', in which differences are claimed which are in fact absent or negligible). In general, P values must be used wisely and always together with the corresponding estimates of effect size. In this study, as in A, the magnitude of the effect being tested is estimated with R square (Rsq), which is the amount of variance accounted for by a factor. Rsq, also known as the coefficient of determination, has some potential disadvantages (see A3 and B1), but it is easy to compute and interpret both for univariate and multivariate data.

B4) *Species discriminant analysis (DA).* This analysis is also complimentary and in most respects equivalent (the significance test is identical) to a multivariate ANOVA, but has a different focus. Its purpose is to estimate species classification accuracy using a set of predictors: the larger the accuracy, the larger the differences. A DA must always be cross-validated (see Methods). Sometimes, this method is also called canonical variates analysis (CVA). A CVA can be used to draw scatterplots that maximize group differences (Sneath & Sokal 1973; Neff & Marcus 1980). However, for reasons I explain later, to this aim I will be using an alternative ordination method, the between group PCA (bgPCA – see Rohlf 2021, and references therein). Although, in theory, a DA/CVA could be done using a single predictor, such as CS, it is regarded as a multivariate technique and implemented in most software so that one must have at least two variables to run the analysis. In GMM applied to taxonomic research, there is usually more interest in group predictions based on multivariate shape. CS, in contrast, is less frequently employed to predict groups. Indeed, size might be a poor predictor of taxonomic affiliation, because it is univariate (thus, contains less information) and often considered prone to change and homoplasy (e.g., Maurer *et al.* 1992; Marroig & Cheverud 2005; Millien 2006). Nonetheless, there seems to be little evidence, for now, to robustly claim that size varies more easily than shape because of plasticity, adaptation or both. Thus, whether size is less informative in taxonomic research cannot be assumed a priori. The importance of size differences should not be overlooked and a careful analysis of size is potentially as interesting as shape analysis. Therefore, to explore how well mandibular size differences predict species affiliation, I use, as explained in the Methods, an expedient to circumvent the limitation of the software that restricts a DA/CVA to multivariate data.

B5) *Summary and visualization of species shape differences*. If interspecific differences are found, it is important to understand what are the patterns of similarity (i.e., who is more similar to whom and how much). For univariate size, a graphical summary of variation is easily done using box-plots, which have already been introduced in part A. For shape, morphometricians employ mainly ordination methods (i.e., summaries based on scatterplots, such as PCAs and CVAs) and, somewhat less frequently, phenograms (i.e., tree-like representations of similarity). I will provide examples of shape ordinations using both all individuals and species, as well as the species mean shapes. Mean shapes will be also used to build a phenogram whose information is complimentary to the ordination. Graphical summaries of variation in Procrustes shape will be accompanied, as customary in GMM, by shape diagrams (Klingenberg 2013) that aid the interpretation of group differences, once these have been analytically demonstrated (using tests) and summarized (ordinations and phenograms).

B6) *Relationship between shape and size within and across species.* The evidence is not strong, as mentioned, but size is often considered more evolutionary labile than shape (Grossnickle 2020). Because differences in the size of a structure tend to change its proportions, a researcher might want to know if shape variation is influenced by size differences and whether this happens in similar ways across all taxa (Emerson & Bramble 1993; Gayon 2000). The covariation between shape and size is called allometry and, when it happens within a species in the same age group (adults, in my case), it is called static allometry (Klingenberg 1998). Allometry, if present, can be compared among species or statistically controlled for (see below). The comparison of interspecific allometric trajectories (evolutionary allometry) is in itself a potential source of taxonomic information, as closely related groups are expected to show less divergence in allometric patterns. I will first test allometry within species using a series of multivariate regressions of shape on size (Klingenberg 2016, 2022). Later, I will test it again, simultaneously in all species, using a multivariate analysis of covariance (MANCOVA). The MANCOVA tests the overall significance of static allometries but, more importantly, provides information on whether allometric trajectories are similar across species. Similar allometric trajectories imply that, within each marmot species, the changes in proportions of the different mandibular regions occur in a comparable fashion. For instance, if, say, larger individuals of hoary marmots tend to have longer angular processes, and interspecific allometric trajectories are similar, also in the other species larger marmots will typically have longer mandibular angles. If one can demonstrate that interspecific allometries are similar in direction (i.e., statistically parallel), the MANCOVA can be repeated with a slightly different design (see B6 Methods) and used to test whether species differ in shape even when the effect of size on shape is statistically minimized. Thus, controlling for allometry in this way (i.e., 'size-correcting' shape), before testing groups, allows to investigate whether shape differences, if present, are purely related to interspecific changes in size.

B7) *Species comparisons of the magnitude of size and shape variance.* This last type of analysis is less common in taxonomic research, but can be informative and especially interesting in specific cases. One example is when one group consists of a small population with a limited geographic range. This could be an island population or any other type of small, isolated population, including endangered taxa and populations surviving in fragmented habitats. In these instances, if genetic data are not yet available, a researcher could use morphology as a proxy to preliminarily investigate the occurrence of a reduced phenotypic variability as an indication of genetic bottlenecks. Because the Vancouver Island marmot (VAN) survives in a tiny insular population, it is interesting to test whether variance in mandibular size and shape is similar to that of its continental sister species, the hoary marmot. The same question could be asked for the Olympic marmot, which has a larger population but is isolated on a peninsula. Unfortunately, the sample of the Olympic marmot is too small for accurate inferences. Thus, I will exemplify the comparison of variances only for VAN, but will also replicate the tests in subsamples of hoary marmots, and other marmot species, to explore the impact of sampling error on estimates of variance.

Material and methods

The information on the study samples is in part A (main text and table 2). However, as a reminder to aid readers, in Table 1, showing the tests of SDM one species at a time (B1), I replicated the list of scientific and common names, as well as the corresponding abbreviations. The methods of data collection are detailed in A, whereas in part B I describe only the techniques specific to group comparisons. As in A, in parallel with the theory, there will be tips to implement a specific analysis using one (or more) of the user-friendly programs I adopted for this study. Table 1 of part A can be used for suggestions on R packages that allow equivalent analyses. The convention of using italics for examples of commands to run the analyses in PAST, MorphoJ or the TPS Series is the same as in A. File extensions too are named as in A (e.g., *.nts or NTS). All analyses in B, however, are done using only the 12 landmarks configuration with the 445 specimens left after excluding the potential outliers (see outlier detection in A). Alpha, the significance threshold, is 0.005 (see A3 Methods).

Methods, results and discussion subdivided by study question

B1) Sexual dimorphism within species

Methods (B1)

Testing sample mean differences in size and shape with only two groups, as in the case of SDM, is straightforward. There are several options in terms of software. The test statistics and models might differ, but results are usually similar for 'reasonable' data. As in part A, I informally use the attribute 'reasonable' for data with a small p / N ratio (more individuals than variables) and sample sizes that are not too highly heterogeneous (i.e., N may differ, but differences between groups are not very large). Sample size in tests of SDM (both in B1 and B2), however, could be smaller. This is because, as it happens in the marmot dataset, there might be individuals of unknown sex. These will have to be excluded from the tests of sex differences. In MorphoJ, to subset specimens, one might use different options from the menu *Preliminaries* (*Include or Exclude Observations* or *Subdivide Dataset By* using, in both instances, an appropriate classifier). In PAST, one can use subsets built in MorphoJ (and, if necessary, edited in a spreadsheet) or, for shape, simply unselect unsexed specimens before running an

Table 1. Within-species 10 000 permutation tests for sex mean differences in CS or shape, performed in MorphoJ. As in part A, in this and other tables, species names are abbreviated using the first three letters of the scientific names (shown again here, together with the common names, to aid readers); significant P values (P < 0.005) are in italic; the most relevant results for the Discussion are emphasized with a light grey background.

Data	Scientific	Common	Acronym	P	Rsq
CS	*M.* (*M.*) *broweri*	Alaskan	bro	0.2731	20.5%
	M. (*P.*) *caligata*	hoary	cal	0.3629	1.1%
	M. (*P.*) *flaviventris*	yellow-bellied	fla	*0.0039*	6.0%
	M. (*M.*) *monax*	woodchuck	mon	0.2543	1.5%
	M. (*P.*) *olympus*	Olympic	oly	0.0368	32.8%
	M. (*P.*) *vancouverensis*	Vancouver Isl. marmot	van	0.0463	20.8%
shape	*M.* (*M.*) *broweri*	Alaskan	bro	0.3177	15.9%
	M. (*P.*) *caligata*	hoary	cal	0.0193	2.8%
	M. (*P.*) *flaviventris*	yellow-bellied	fla	0.0212	1.7%
	M. (*M.*) *monax*	woodchuck	mon	0.0704	2.0%
	M. (*P.*) *olympus*	Olympic	oly	0.6619	6.4%
	M. (*P.*) *vancouverensis*	Vancouver Isl. marmot	van	0.7045	4.0%

SDM test. Re-superimposing shape data in analyses of subsamples is an option, but generally makes a negligible difference in the results (see A2)[2].

Group mean differences in CS can be tested using *Statistics, F and t test (two samples)* in PAST. With only two groups, F and t are equivalent test statistics, as F equals t squared. However, if sample variance is not homogeneous, the t-test can be computed with a correction for heteroscedasticity. Any introductory statistical manual has descriptions of these tests (e.g., Moore & McCabe 2005; Howell 2013). The test is performed in PAST using both parametric and resampling methods. The result window in PAST also provides the Levene's test for the assumption of homogeneity of variance. As all other analyses in the menu *Statistics* of PAST, the F and t test for two independent samples is specific to univariate data. However, PAST offers equivalent tests for multivariate data (*Multivar, Discriminant/Hotelling* or *Multivar, Two-group permutation*, for respectively parametric and resampling methods) and there is also the possibility of testing homoscedasticity by assessing the statistical equivalence of the covariance matrices of two multivariate samples (*Multivar, Box's M*). For a brief discussion on the assessment of common statistical assumptions, however, I refer the reader to the Appendix A.

[2] Users, however, have to be careful if analyses are done after reducing the data dimensionality with a PCA. Dimensionality reduction is an operation that, as stressed multiple times, bears some risks and must be done rigorously and only when strictly necessary. Now, let us say, as a made-up example, that the researcher demonstrated that 10 PCs adequately summarize shape variation in the 445 marmot specimens. If he/she, later, needs to analyse a subsample, for instance for testing SDM in the Olympic marmot, this cannot be done by re-cycling the scores of the 10 PCs, obtained in a PCA of all 445 individuals, to analyse the 14 specimens of *M. olympus*. If for the Olympic marmot (or any other subsample of the total sample) one needs to reduce dimensionality, a new PCA will have to be done on the Procrustes shape coordinates in this sample and a number of the first PCs computed, that is adequate for the specific case. Thus, the user might find not only that less (or, even if unlikely, more) PCs are necessary for *M. olympus,* but also that inevitably the PCs scores are different from those obtained in the total sample and the sub-sample specific set of PCs has a better correspondence to the Procrustes shape distances (i.e., it is a more accurate summary) of this species.

Using PAST for testing the mean differences of two samples is convenient and fast. Another option, which I favoured in this study, is regressing in MorphoJ size or shape onto a dummy variable (a covariate, in MorphoJ's jargon) coding females as -1 (or 0) and males as 1 (or vice versa). The test, already briefly mentioned in part A, employs Rsq (the variance explained by SDM) as a test statistics and estimates its P value with permutations (10 000 in my analysis). The advantage of using the regression approach in MorphoJ for testing group mean differences is that one obtains the estimate of the size of the effect being tested (the Rsq), as well as the corresponding P value. The test is simple and does not require normally distributed data because it is using a resampling method. Also, even if homoscedasticity is not tested, MorphoJ's regression provides the equivalent of a univariate two-group jitter-plot in the *Graphics* window, which can, as a crude approximation, help to spot large differences in variance between females and males. The multivariate extension of the SDM regression test in MorphoJ is also very simple. As with univariate data, this procedure is analogous to testing the absolute mean difference between females and males using their Euclidean distance (i.e., the length of a straight line between the means in a univariate or multivariate space). The regression is specified always in the same way (MorphoJ's menu *Covariation, Regression*), but the user selects, as dependent variables, CS for size and the Procrustes coordinates for shape. The independent variable is the sex dummy covariate in both cases.

The regressions have to be done within each species. Tests in small samples are inevitably less accurate and powerful. Regardless of the type of test, testing SDM one species at a time potentially inflates type I error rates. Caution in the interpretation of results and/or a correction for multiple tests are usually enough to avoid serious issues (Armstrong 2014; Krzywinski & Altman 2014). I will not use a correction, however. If I did, a sequential Bonferroni is a simple option, but has pros and cons and there are alternatives (Verhoeven *et al.* 2005). I have chosen a conservative alpha (0.005) which partly protects against false positives. Furthermore, I carefully interpret P values in relation to Rsq and sample size, as well as using graphical summaries to complement the tests. Yet, as discussed in part A, Rsq also must be interpreted with care, because it is a biased estimator, which tends to be inflated (i.e., larger than true) in small samples.

In terms of graphical summaries for the within-species tests of SDM, the 'two-group jitterplot' in the *Graphics, Scores* window of MorphoJ gives a preliminary idea of the amount of overlap or separation of females and males. For CS, this plot is approximately equivalent to a jitter plot in PAST (see below), but, for shape, it only shows the component of multivariate shape variation that covaries with sex. Thus, the visual inspection of these multivariate regression scores, in MorphoJ, may suggest more separation than real in the full multivariate space.

Instead of using MorphoJ's regression plot, for CS I prefer to draw box and jitter-plots, subdivided by species and sex, in PAST. Box and jitter-plots show the distribution of CS, but also the medians, quartiles and range, so that one can inspect overlap or separation between sexes across all species. In part A (see 'outliers: univariate size'), I have already described this type of univariate plots, how to obtain them in PAST and how to combine the two separate plots into a single one using a photo-editor.

For shape, summary scatterplots of multivariate variation can be helpful to explore sex differences. With a negligible SDM, within-species ordinations, such as a PCA, should show large overlap in the distribution of females and males along the main axes of shape variation (PC1, PC2, PC3 etc.). With an appreciable SDM, in contrast, sex differences in adults of monomorphic species generally dominate PC1, which should, therefore, suggest some degree of separation between females and males, plotted using different symbols and/or colours. The within-species PCA scatterplots are easily computed in PAST by selecting the Procrustes shape coordinates of a species, before using the command *Multivar, Principal components*, checking the *Var-covar* and *Disregard Groups* boxes. Females and males will have to be, first, marked using different colours, as explained in part A. The same type of scatterplot can be done in MorphoJ, after splitting the sample by species. If all the species datasets are selected in the MorphoJ's project tree, the PCA is performed simultaneously for all species (see footnote 17 in the section 'Outliers: multivariate shape' in the Methods of A2). For exploring SDM in shape ordinations, specimens of unknown sex must be excluded, because they affect the computation of the axes on which individuals are projected and, thus, potentially confound sex differences.

Ordinations, which help to maximize group separation, instead of total variance regardless of groups as in a PCA, are also appropriate to explore SDM in shape. I will present an example using a bgPCA, but I defer a more detailed explanation of this method to the chapters B4-B5, where the DA/CVA and bgPCA are used more extensively. In the example, I will be focusing only on the three species with the largest samples (the woodchuck, hoary and yellow-bellied marmot), because samples of females and males in the other species are too small for robust results in ordinations.

Results (B1)

Testing mean size and shape of females and males one species at a time suggests that sex differences range from modest to negligible (Table 1). Only CS in yellow-bellied marmots is significant ($P < 0.005$). Olympic marmots and VAN, for CS, and hoary and yellow-bellied marmots, for shape, are just below a conventional 0.05 threshold[3]. The average SDM Rsq is 14% for CS and 6% for shape. However, these averages are likely biased by the inflated Rsq of the three species with smaller samples of sexed individuals (Alaskan, Olympic marmots and VAN, with an averaged Rsq of 25% for CS and 9% for shape). If only species with large samples of females and males are considered (i.e., woodchucks, hoary and yellow-bellied marmots), CS and shape average Rsq drop respectively to 3% and 2%. Thus, overall, within-species tests of SDM indicate largely negligible sex differences in mandibular morphology.

Box and jitter-plots (Fig. 1a) of CS with separate sexes show an almost complete overlap in the range of mandibular size variation of females and males within each species. Males, however, tend to be slightly larger than females, which is particularly evident and supported for yellow-bellied marmots. The variability, and asymmetry, of the box-plots of the smallest samples (Alaskan, Olympic and Vancourver Island marmots) suggest a large amount of sampling error in these groups. In the other three species, which have much larger samples, in contrast, box-plots of females and males are roughly symmetric and approximately similar within each species in terms of the size of the box and length of the whiskers, although the lower whisker of the woodchuck is longer in females compared to males.

Simple within-species PCA scatterplots of shape (not shown) suggest large and, sometimes, almost complete overlap between females and males. The small SDM in shape, however, becomes even more evident if data are summarized using both species and sex as groups. This type of plot is appropriate also for complementing the results of B2, but it is shown here in order to have a 'counterpart' of the species by sex box and jitter-plots (Fig. 1a). Thus, Fig. 2 shows the two main axes of separation for species and sex in the woodchuck, hoary and yellow-bellied marmots using a bgPCA. In this figure, species are well separated, whereas females and males almost completely overlap within each species. Therefore, as with CS, mandibular shape SDM looks totally negligible compared to species differences not only in the statistical tests but also in the data plots.

Discussion (B1)

SDM in North American marmots mandibular size and shape is negligibly small. This is consistent with previous studies on mandibles (Cardini & Tongiorgi 2003; Cardini 2003; Nagorsen & Cardini

[3] As discussed in part A (power analysis), SDM in shape becomes significant using the 0.005 threshold in hoary and yellow-bellied marmots, when tested using parametric tests, such as those available in TPSRegr. Parametric tests are generally more powerful than resampling methods. However, even if significant, the choice of parametric methods would not change the conclusion that the effect of SDM is several times smaller than interspecific differences. For instance, for shape in yellow-bellied and hoary marmots, despite SDM $P < 0.05$ using permutation tests (Table 1) and < 0.005 with parametric tests (see Discussion in A3), the SDM Rsq is just 2-3% compared to ~ 20% for interspecific mean differences (see results of A1 and B2).

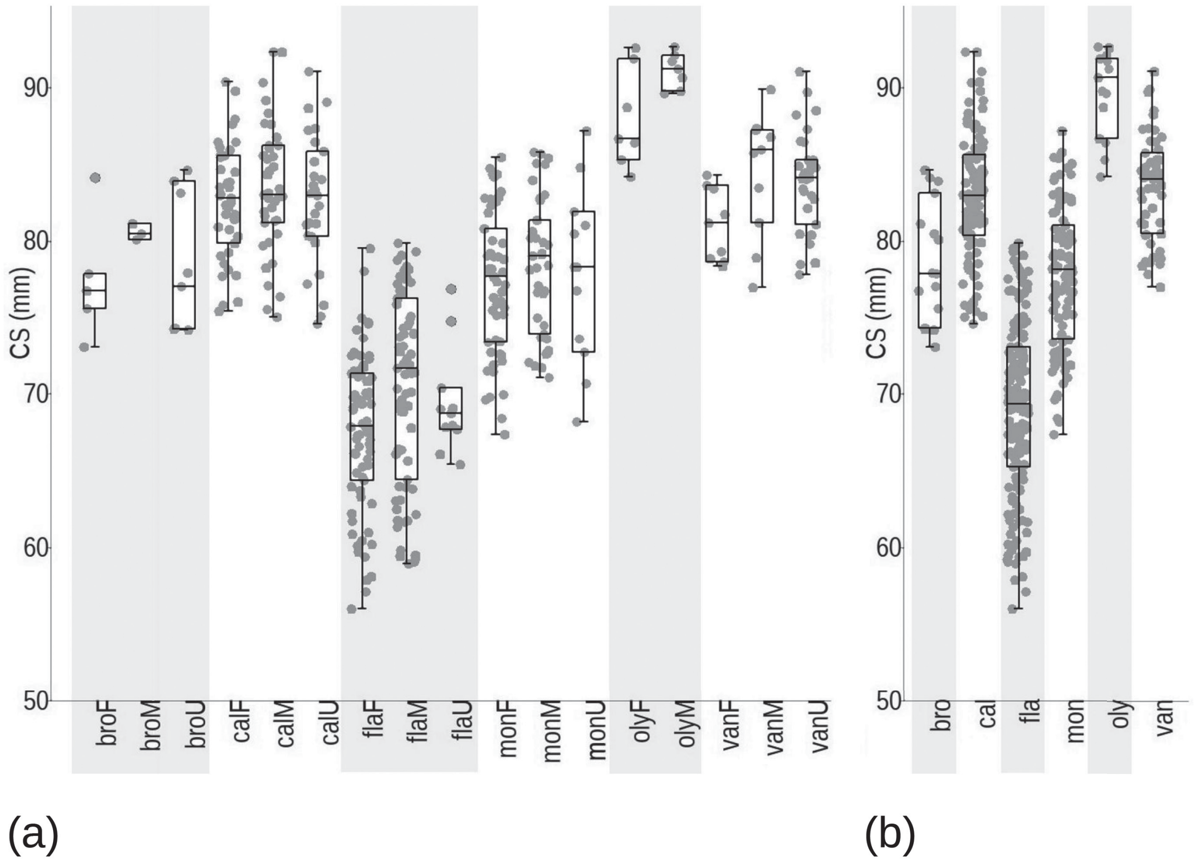

Fig. 1. Box and jitter-plots of CS, for each species. **a**. Separate plots for females, males and unknown individuals. **b**. Plots with pooled sexes. As in part A, as well as shown in Table 1, species names in all figures are abbreviated using the first three letters of the scientific name (e.g., *caligata* = cal) and F for female, M for male, and U for individuals of unknown sex.

2009). That average mandibular CS is just between 1% and 4% larger in males, compared to females, corresponds well with the observation of a generally very limited amount of male-biased SDM in marmots condylobasal and hind foot length, at most, on average respectively 7% and 3% longer in males (Matějů & Kratochvíl 2013). Also, as in other marmots and ground squirrels (Matějů & Kratochvíl 2013), there seems to be no clear evidence that this modest amount of size SDM follows the prediction of the Rensch's rule (Abouheif & Fairbairn 1997), that, when male is the larger sex, SDM should be more pronounced in larger species. For instance, if we focus on the largest, and thus more reliable, species samples, the estimate of the average male mandibular CS is 4% larger than in females in the yellow-bellied marmot, which is the only species to reach significance in the tests. However, the difference is only 1% in the woodchuck and hoary marmot, despite these species having a mandible on average ~ 10–20% larger compared to yellow-bellied marmots.

SDM Rsq are modest for CS and very small for shape, when the differences between females and males of the smallest species samples (Alaskan, Olympic and Vancourver Island marmots) are not considered. We know that tests in small samples are problematic, because estimates of means and variances are inaccurate and statistical power is low. As explained in A (Discussion on power, in A3, and Appendix A), sample mean differences tend to be overestimated when sample size is small. Thus, in Alaskan and Olympic marmots, as well as in VAN (where the total N is large, but most specimens are of unknown

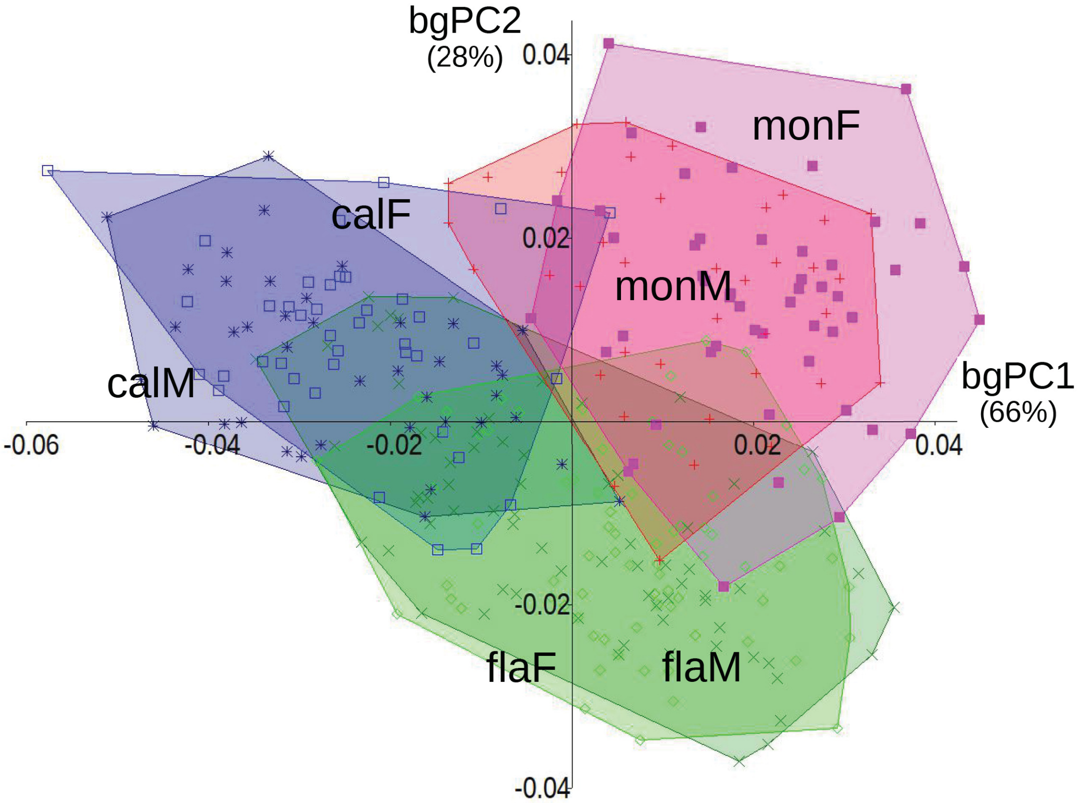

Fig. 2. Visualization of shape SDM in relation to interspecific differences using a bgPCA. In this, and other Figures, percentages of variance in the scatterplots of multivariate shape are shown in parentheses, below the label for the corresponding axis. On bgPC1–2, which together account for almost all between group variance (94%), there is a large overlap between females and males within each species, whereas, between species, the separation is clear.

sex), one cannot really be confidently about the estimates of mandibular SDM (non-significant in all of them). Yet, given the general similarities in reproductive biology, ecology and morphology of marmots, it is a reasonable expectation that SDM in the mandibles of these three species is of similar magnitude, and thus minimal, as in the other species with much larger samples. Indeed, some marmot species are considered monomorphic (Tafani *et al.* 2013) and macroevolutionary studies suggest that sex differences in the sciurids in general, and more specifically in the Marmotini tribe (to which marmots belong), are small or absent (Hayssen 2008). For instance, the marmotine female to male body length ratio is 0.993 and slopes and intercepts of interspecific regressions of body mass onto body length are also very similar in females and males (Hayssen 2008). Nevertheless, in the Marmotini, males tends to be slightly heavier (6%) than females (Hayssen 2008).

Why is SDM so small in marmots? In terms of reproductive strategy, marmots range from monogamous to moderately polygynous (Armitage 1999, 2000; Kyle *et al.* 2007). Even when polygynous, however, females of some species might have extra-pair matings (Goossens *et al.* 1998; Waterman *et al.* 2007; Maher & Duron 2010). In fact, in the only solitary marmot species, the woodchuck, multiple paternity is so common that its reproductive strategy can be described as promiscuous (Maher & Duron 2010). If there is monogamy or males have a limited control on female reproduction, sexual selection is unlikely to promote large differences in size (Ralls 1977; Lindenfors *et al.* 2007). In my study, however, yellow-bellied marmots represent a small exception, in that they show significantly larger mandibles in males, unlike other species. Their sample is the largest, and thus the test is more powerful, but hoary marmots and woodchucks have N only moderately smaller than yellow-bellied marmots and their CS SDM, which is even smaller in magnitude, is not significant (Table 1). The relatively larger SDM of yellow-bellied marmots is supported by average differences in hind foot and condylobasal length, which are both 7% longer in males of yellow-bellied marmots compared to a difference of just 0.4–4% between males and females in woodchucks and hoary marmots, respectively (Matějů & Kratochvíl 2013). The appreciable, even if modest, dimorphism in mandibular size of *M. flaviventris* seems, thus, genuine. SDM in shape, in contrast, is non-significant in all North American species. Thus, overall, the within-species tests of SDM provide a first line of evidence that mandibular SDM is mostly negligible and interspecific comparisons might be done with pooled-sex samples.

Assessing SDM one species at a time is the easiest approach, but has pros and cons. The main advantage is, besides simplicity, that it provides species-specific information. A disadvantage, however, is the potential inflation of type I errors, because the same hypothesis is tested over and over. More importantly, unlike in the species by sex ANOVA, SDM is not 'scaled' in relation to the magnitude of interspecific differences. Finally, for shape SDM, when tested one species at a time, there is no explicit comparison of the direction of average sex differences. With direction, in a univariate analysis, one simply means which sex is larger. In a multivariate analysis, however, it means comparing the species SDM variance covariance structure or, at least, visualizing shape change to assess differences and similarities in SDM across species. For the visualization, one can use a wireframe, 'lollipops' or thin-plate spline diagrams (Klingenberg 2013) to describe the differences between the female and male mean shapes of each species. There was no motivation for doing it at this initial stage of the analysis, however, because SDM in marmot mandibular shape is so small.

The disadvantages I have listed in the previous paragraph for the within-species tests of SDM (B1) are less concerning when these tests, one species at a time, are complimentary to the species by sex ANOVA (B2). Also, the limitations of this approach can be, to some extent, overcome. In the chapter on pairwise comparisons (B3), I will say more on how to mitigate against potentially inflated type I error rates in multiple tests. Furthermore, even when testing SDM one species at a time, we are not relying only on P values: Rsq are also carefully considered to assess the magnitude of SDM. And, to put these Rsq into context, one could calculate the Rsq of the interspecific comparisons using separate sexes, and then compare their magnitude with that of the SDM Rsq. I suggest, however, not to do it this way, because it is time-consuming, inelegant and less statistically powerful than the species by sex ANOVA (B2).

The ANOVA (B2) provides also a mean to compare patterns of shape SDM not only in terms of magnitude, but also direction. For instance, if SDM is significant, a researcher might want to know if, say, having on average a deeper and relatively shorter mandible in males, compared to females, in one species is also a main feature of SDM in other species. The visualization of species-specific shape change between sexes is, as said, one way to qualitatively explore the answer to this question. However, the direction of shape SDM can be more accurately quantified and compared. This is achieved by computing pairwise species angles between vectors of mean sex differences in shape. The vectors are simply the slope coefficients of the regressions of the Procrustes shape coordinates onto the sex dummy variable. The angle quantifies the similarity in regression slopes between two species: a small angle implies SDM vectors pointing in similar directions and, thus, congruence in the patterns of shape change between the two species.

Angles are easily computed in MorphoJ: first, one selects a species regression in the project tree – e.g., hoary marmot shape onto sex; then, one specifies the shape SDM regression of the second species – for instance, the yellow-bellied marmot – using the lower box of the *Comparison, Compare Vector Directions* window; finally, he/she clicks on *execute* to run the test. If the angle (*Results* window) is relatively close to zero and significantly smaller than expected from pairs of random vectors, the direction of shape

SDM is statistically similar. Otherwise, if power is adequate and both regressions were significant, but their angle is not, that means that the two species likely differ in the pattern of shape SDM.

However, because the test of angles is done pairwise for all species, there is, again, the risk of inflating the rate of type I errors. Also, assessing angles pairwise is more laborious than simultaneously testing species and sex differences in a two-way MANOVA (B2). Besides, as with other analyses, when N is small and heterogeneous, and maybe there is a large number of variables, estimates of angles may be inaccurate (Cardini & Elton 2007). For all these reasons, although the problem with p / N and small samples remains, I tend to prefer the species by sex MANOVA to assess the similarity of patterns of shape SDM. Nevertheless, exploring vector angles at least in the largest samples does provide useful information. For instance, one might check if angles support the findings of the MANOVA: with a negligible species by sex interaction (see B2), the expectation is that angles of shape SDM regression vectors, at least using the more accurate estimates of the largest samples, should all be relatively close to zero. Testing, or at least computing, angles of shape SDM vectors is also a kind of post-hoc comparisons for the species by sex interaction of the MANOVA, because it allows to detect which species, if any, diverges more from a common pattern of sex differences in shape.

Patterns of variation between sexes can also be examined using summary plots for univariate or multivariate data. I discuss shape, first and briefly, as the outcome of the graphics is unambiguous. The bgPCA ordination fully supports the tests of shape SDM showing a very small effect of sex, except when Rsq are strongly biased by small N. That mean sex differences are almost certainly overestimated in small samples had already been shown in Appendix A (Fig. A1). This is why I left out the three species with the smallest samples of females and males (i.e., VAN, the Alaskan and the Olympic marmot) in the shape summary scatterplots of species and sex (Fig. 2). Thus, focusing on more accurate results obtained in the three largest species samples, one can readily see (Fig. 2) that females and males overlap almost completely, but species are very well separated. In fact, one can appreciate the dominant effect of species differences also by considering the distribution of the between group variance in the scatterplot. In a bgPCA, the number of axes is set by the number of groups minus one (see B4–B5 for more on group ordinations). Thus, we should have five axes (three species by two sexes minus one) with non-zero variance. In contrast, it is as if there were almost only three groups and, thus, just two axes accounting for almost all (94%) between group shape variance. This happens because the effect of species (the three included in the ordination) dominates, with sex having such a vanishingly small impact on group structure that there is no sex separation on any bgPC (including bgPC3 and bgPC4, which are not shown). Indeed, in these three species, if the Procrustes distances between means, split by species and sex, are compared, interspecific distances turn out to be on average three times larger than within-species distances between means of females and males.

The box and jitter-plots of mandibular CS, with samples split by species and sex (Fig. 1a), also indicate overlap between sexes and a degree of separation among species. The interspecific separation is, for some species (e.g., woodchucks compared to yellow-bellied marmots or hoary marmots compared to VAN) less striking than in the ordination of shape. Yet, it is evident and more pronounced that the tiny sex differences within species. The box and jitter-plots, however, show something else, which is worth being discussed. There are three apparent outliers, detected with the outlier option of the box-plot in PAST (see methods in A2). It seems, at a first glance, worrying that none of them had been spotted in the preliminary search for outliers (also in A2). However, these ‘outliers’ are, in a sense, an artefact. All three are found in two of the smallest samples (female Alaskan marmots, N = 5, and unsexed yellow-bellied marmots, N = 12). When sexes are pooled, after demonstrating the negligible SDM in CS, all of them fall within the main range of variation for size in those two species. As explained in part A, outlier detection in small samples is hard or impossible and results from the application of methods based on rigid thresholds should be inspected with special care in small samples. In yellow-bellied marmots, the sample of individuals of unknown sex, where two apparent outliers where found, is not very small. However, it is very heterogeneous in a species where there is a subtle size SDM and a large range of variation. In this unsexed sample, 10 out of 12 individuals are below the median CS of the total yellow-bellied marmot sample and, thus, have relatively small mandibles. The remaining two unsexed individuals would be just above or below the third quartile (i.e., the upper side of the box) in a box-plot with all 156 yellow-bellied marmots. Having, by chance, such an asymmetric distribution in the unsexed sample of yellow-bellied marmots, with many small to medium size individuals and just a couple of fairly large ones, pushes these two large individuals above the threshold for outliers. The plot is, thus, misleading and, after pooling sexes (Fig. 1b), there is no reason to exclude these specimens. Large inaccuracies due to sampling error in the smallest samples also likely explain why, in both the Alaskan and Olympic marmot, the range of variation in male mandibular CS is tiny: N is very small (3 and 7, respectively); and some of these few specimens could have also been positively autocorrelated[4], if they were members of the same colony, and thus likely relatives, and maybe had been collected in the same year.

To summarize, the first step, exploring the potential impact of SDM on the taxonomic comparisons, indicates small sex differences, that are likely to be negligible in interspecific comparisons. The magnitude of SDM is likely to be similar, although appreciably larger in the mandibular size of yellow-bellied marmots, but probably spuriously larger in the three species (VAN, Alaskan and Olympic marmots) where few individuals are of known sex. The problem with small samples in tests of SDM had been anticipated by the prospective part of the power analysis (A3), that showed how testing SDM using small random subsamples of the largest species suggested serious issues with low statistical power. Randomized subsampling experiments (Appendix A) also demonstrated that mean shape differences between females and males are likely to be inflated, which is consistent with the large SDM Rsq found in VAN, Olympic and Alaskan marmots, all of them with within sex N ≤ 11. This is a first clear example of a well-known, almost obvious, problem, that affects many taxonomic studies and tends to be overlooked: with small differences, samples must be large for accuracy and power (Cardini *et al.* 2021).

B2) Sexual dimorphism in relation to species differences

Methods (B2)

The conventional GMM approach for testing sex differences across species in taxonomy is a species by sex ANOVA (e.g., Rohlf *et al.* 1996; Corti & Rohlf 2001; Frost *et al.* 2003). This approach was borrowed from multivariate traditional morphometrics (e.g., Campbell & Mahon 1974; Willig *et al.* 1986), but see also Neff & Marcus (1980) for an introduction). The species by sex ANOVA is a ‘two-way’ analysis, because there are two factors (species and sex). The ‘ME ANOVA’ was, instead, a three-way analysis using species and sex as main factors, but also taking into account duplicates by adding ‘individuals’ as a random factor. However, in the ‘ME ANOVA’, the interactions were assumed to be negligible (see Discussion). In contrast, in the species by sex ANOVA in part B, with each individual now represented by the average of its two digitizations[5], it becomes crucial to assess not only species and sex, but also

[4] I.e., observations that, like ‘pseudo-replicates’, are more similar than expected if truly independent.

[5] Please, note that, as explained in part A, I can average the duplicate digitizations of an individual using its raw coordinates, because they represent replicate digitizations of the same image. In contrast, if I had taken multiple images, repositioning each specimen in order to take into account also this source of measurement error, then, having demonstrated a negligible ME, I should average the Procrustes shape coordinates (not the raw ones!) and CS of each individual, because the raw coordinates have positional differences. Averaging is easily done in MorphoJ using *Preliminaries, Average Observations By* individual (the classifier uniquely identifying each specimen) and then selecting the appropriate variables in the window that will be opened. If needed for specific aims, one could later restore size in the averaged shape replicates by multiplying column-wise the average Procrustes shape coordinates of each individual by its averaged CS. For instance, this could be useful to save a new ‘clean’ dataset, without replicates, with a single file having both size and shape information that could be re-separated into CS and Procrustes shape coordinates by redoing the superimposition in another program (say, TPSSmall or TPSRegr).

their interaction. The interaction, as anticipated in the Introduction, provides an estimate of whether patterns of sex differences across species are statistically similar. If the interaction is significant, that indicates that SDM varies depending on the species; in this case, interspecific tests (B3 etc.) will have to be run separately for females and males. In contrast, with a non-significant test and small Rsq for the species by sex interaction, one can focus on the main effects of species and sex (see below), and use the results to decide if females and males can be pooled. I stress the importance of a small Rsq for the interaction, because one can have a non-significant interaction despite a large Rsq, especially when N is small and/or heterogeneous across groups. If N is small, power may be low and non-significance might simply reflect the large uncertainties due to poor sampling. This means that SDM may, in fact, vary depending on the species, but the effect is not detected by the ANOVA because of low power. Besides, with small samples, Rsq itself is inaccurate and potentially inflated, which makes it less reliable. Thus, in general, with small samples, one cannot accurately test sex differences and confidently assess whether patterns of SDM are similar or different across species. Exploratory graphical analyses, together with the relevant SDM information from the published literature, might be the best one can do to decide about sex differences when N is too small.

When sample size is large and the species by sex interaction is not significant, and only in this case, the ANOVA is repeated after leaving out of the model the interaction. Most statistical software (unlike, unfortunately, PAST – see below) has an option to do this. For brevity, I refer to this second analysis, in which only species and sex are tested, as the 'species plus sex' ANOVA. Before discussing how to interpret its results to decide if SDM is negligible in taxonomic comparisons, I suggest a consideration that rephrases the meaning of 'negligible interaction' in a way that can help to make it clearer. The ANOVA including the interaction treats the interaction as a third predictor, besides species and sex. In this sense, it can be said that the 'species plus sex' ANOVA is using a reduced model (two predictors: species and sex) compared to the full model (three predictors: species, sex and species by sex) of the first ANOVA. Which is the model that accounts, overall, for more variance in the dependent variable (in my case, CS or shape)? Inevitably, it is the full model that has the larger Rsq, because adding any predictor, even if merely by chance, will account for some variance in the dependent variables beyond the variance already explained by the other factors. With a negligible interaction, however, the expectation is that the full and reduced models have almost identical Rsq.

Thus, with a non-significant interaction, the 'species plus sex' ANOVA provides the information to decide how to use females and males in the taxonomic comparisons. If species is significant, but sex is not and its Rsq is small compared to the species Rsq, all further analyses (B3 and following ones) can be done regardless of sex. This means including both females and males, when species are compared. For instance, Rohlf *et al.* (1996: 350) first used a "two-way MANOVA (sex by sample) ... performed" on Procrustes shape data and then, based on its results, decided to pool sexes, "because only group differences [i.e., those among taxa] were statistically significant, [so that] the two-way MANOVA was collapsed to single classification design and analyzed using CVA".

There is at least another potential scenario to consider in the outcome of the 'species plus sex' ANOVA. What shall be done if, despite a negligible interaction, the Rsq of sex is approximately as large as the Rsq of species differences? Because SDM is large and significant, its effect on interspecific comparisons cannot be overlooked. Therefore, further analyses (B3 etc.) will be run in parallel using separate sexes[6]. There is an alternative, however. Because SDM is similar in all species (as shown by the negligible

[6] I am not considering a case of highly unbalanced sample sizes, where one sex is much more represented in most species than the other. In this instance, especially if it is known from the literature that there is SDM, a researcher might simply use only the most abundant sex (say, females) for the taxonomic comparisons and exclude the other (males, in my made-up example).

interaction, assuming N is adequately large and the result is reliable), data can be 'sex-corrected'[7]. I do not provide details on 'sex-corrections' [8]. There are slightly different ways of doing them, but the basic rationale is to statistically remove the mean differences between females and males before pooling sexes. For instance, within each species, a very simple approach could be to subtract the female mean from female sample and, then, add the resulting residuals to the male mean, thus producing a single dataset with males and 'masculinized' females.

The results of the size or shape two-way ANOVAs may not be clear cut. Thus, the decision on whether to pool sexes or not might require a degree of arbitrariness, unless one takes results at face value. This second option means, for instance, that a significant species by sex interaction becomes a rigid rule to separate sex in species comparisons, even when the interaction Rsq is very small compared to the variance accounted for by main factors. This is not unlikely to happen, when samples are large and, thus, statistical power high. For those who want some more flexibility, at the cost of less rigour, I made a decision 'tree' in Table 2. The tree provides a tentative guideline, but has to be used with a grain of salt, on a case by case basis and in relation to the available data (sample size, p/N ratios within groups etc.).

With multi-way ANOVAs, there is a practical issue that restricts the user's choice. The availability of free, user-friendly, software for ANOVAs with more than one factor is limited. This is especially evident for the multivariate analysis. PAST 2.17c can do two-way species by sex univariate ANOVAs, including the interaction, but has no equivalent for the multivariate case[9]. Because of the limitation in the software, I will be using different programs for the ANOVA of CS and the MANOVA of shape. Besides, samples will be smaller in the species by sex ANOVAs, because 89 specimens of unknown sex have to be excluded. For building the subsample of individuals of known sex, I used MorphoJ's *Preliminaries, Include or Exclude observations* and the sex classifier to exclude unsexed individuals.

[7] With a non-significant and small effect of the interaction, in fact, using pooled sexes should not affect results of interspecific tests when samples are perfectly balanced (same sample size in all species and sex groups). However, a perfectly balanced design is most unlikely in taxonomic analyses of morphological data. Besides, even in this unlikely scenario, pooling sexes when SDM is large makes summary plots and diagrams potentially misleading (Cardini 2020a).

[8] A 'sex-correction' is a solution to consider with caution. It is mainly useful when samples are relatively small or contain a large proportion of individuals of unknown sex, because it helps to maximize sample size and increase statistical power. Small samples, however, make the ANOVAs potentially inaccurate: statistical power is small; estimates of parameters (e.g., group means) are inaccurate (Cardini *et al.* 2021); statistical assumptions are hard to verify. Because the sex-correction is valid only as long as results of the ANOVA results are accurate, the situation where the correction is most useful is also the one where it can be misleading. With large samples, in contrast, the ANOVA is more reliable, but one has weaker reasons for sex-correcting data in the presence of a large SDM, since analyses can be done using separate samples for females and males. With separate sex analyses in large samples, power will be slightly lower but data are not statistically transformed, an operation which always implies a potential degree of inaccuracy. Besides, parallel tests in females and males also allow to check the congruence of results in different samples from the same taxa.

[9] In fact, with a perfectly balanced design (but only in this specific case), one could run a permutational ANOVA in PAST and also assess the interaction (*Multivar, Two-way NPMANOVA* using *Euclidean* distances). Users should be careful, because PAST allows to run the analysis even with unbalanced samples, but the help file is clear that this should not be done and, thus, results should not be trusted. Unfortunately, perfectly balanced samples are, as already said, rare in taxonomy. One could select random balanced subsamples, but N would be limited by the smallest species samples of either females or males (with a consequent loss of power and accuracy). For instance, with my marmot dataset (Table 1 in part A), even excluding the two smallest samples (the Alaskan and Olympic marmots), the sample size of random balanced subsamples would be set by N = 9, which is the number of females in VAN. That means excluding almost 80% of individuals and analyse a total of just 72 specimens (9 by 2 sexes by 4 species) instead of 334 using the total samples of the same four species. Excluding also VAN, the loss of power is much less pronounced for the remaining three species (woodchucks, yellow-bellied and hoary marmots). In that case, the total N of a perfectly balanced design using random subsamples would be 228 (with 38 specimens per sex per species) instead of 315 with unbalanced samples of the same three species. This implies leaving out slightly less than 30%, but restricts the analysis to just the three of the six species.

Table 2. Decision 'tree' for the species by sex ANOVA. In the full model (first ANOVA), the main focus is on the interaction; in the reduced (second ANOVA), it is on sex. Light grey background in the first ANOVA suggests when one might do the second one, after excluding the interaction; black background and uppercase emphasize, respectively, separate sex and pooled sex analyses.

1st or 2nd ANOVA	Factor †	P	Rsq*	Decision
1st: 'species + sex + species by sex' (full model)	interaction	significant	large	separate sex analyses
			small	uncertainty, but could try 2nd 'species + sex' ANOVA with no interaction
		not significant	large	likely problems with the data (small N, producing inaccurate parameters and low power): explore taxonomic differences graphically without testing?
			small	do 2nd ANOVA with no interaction
2nd: 'species + sex' (interaction excluded; reduced model)	sex	significant	large	separate sex analyses (or 'sex-correct' data)
			small	POOL SEXES (or 'sex-correct' data)
		not significant	large	likely problems (low power etc.) with the data: explore taxonomic differences graphically without testing?
			small	POOL SEXES

† Accurate results are more likely with large, fairly similar N across groups, and no evident violation of homoscedasticity.

* Decide if large or small by comparing: for the first ANOVA, the Rsq of the full model with the Rsq of the reduced model; for the second ANOVA, Rsq of sex with Rsq of species.

The resulting variables (raw coordinates or CS and shape coordinates) can later be exported as text file, manipulated in a spreadsheet and formatted as appropriate to be used in other programs (see the help files of the software and part A on data formats).

For CS, I run the species by sex ANOVA in PAST using the command *Statistics, Two-way ANOVA*. As clearly explained in the help file and easily replicated using the example file I provide in the SI, data must be formatted in three columns, with species first, followed by sex and finally CS. Both species and sex must be coded using integer numbers (e.g., 1 for Alaskan marmots, 2 for hoary marmots etc. and 1 for females and 2 for males or vice versa). PAST 2.17c, however, does not allow to repeat the ANOVA after excluding the interaction, when this effect is not significant. If the interaction is not significant and its Rsq is small, the tests for the main effects (species and sex) generally produce very similar results regardless of including or excluding the interaction. The Rsq for each effect must be calculated manually in PAST. This can be done by copying/pasting the ANOVA output in a spreadsheet and then dividing the sum of squares[10] (SS) of species, sex and species by sex, each by the total SS. The resulting Rsq should be treated as an approximation, however, because its computation depends on the type of SS used in the ANOVA (see below), which is not defined in PAST 2.17c[11]. At the time of this study, PAST 4 seems to share the limitations of 2.17c for the two-way ANOVA, but future versions will likely offer more options

[10] More accurately, SS are the sum of squared deviations from the mean (see also the Glossary in Appendix A).

[11] Results of the species by sex ANOVA in PAST 2.17c are very similar (but not identical) to those obtained in R using either type I or type II SS.

(Hammer, pers. comm.), including whether the interaction is tested or not and possibly also the choice of the model to partition variance among factors (i.e., the type of SS used in the ANOVA). For now, users may have to put up with the limitations, be careful and report precisely what was done (or could not be done). Alternatively, they may have to use a commercial program or R.

For shape, as anticipated, PAST has no option for a two-way MANOVA equivalent to the univariate analysis I described above. Besides commercial programs, there might be other free statistical software, I am not familiar with, in which to run the analysis. The two-way MANOVA can also be done in R (R Core Team 2023) using the package *Car* (Fox & Weisberg 2019) and parametric tests [12] or, using permutations and a different type of SS [13], the package *Vegan* (Oksanen *et al.* 2022). For 2D GMM data, the species by sex MANOVA can, however, be done in TPSRegr (Rohlf 2015). The analysis is more laborious, because it is based on a series of regressions onto dummy variables. The dummy variables numerically code groups (species and sex) and their interaction. Results are equivalent to a MANOVA using the default type III SS options (see below) of some of the main commercial programs (Howell 2013). I explain first how to build the files for TPSRegr and then how to run the two-way MANOVA. R users, and those using commercial software for two-way MANOVAs, can skip this part of the methods and read only the description of data plots in the last three paragraphs before the Results of B2.

In TPSRegr, files can be loaded in NTS format for both landmarks and groups. The format is simple and well explained in the help of any program of the TPS Series. As briefly outlined in the power analysis of part A, NTS is a text file that one can create in a text editor (e.g., Notepad++: https://notepad-plus-plus.org/). For our specific use, a matrix with observations in rows and variables in columns is pasted in the TXT file, whose extension is changed into *.nts. Before the matrix, one needs a line with four numbers. For the marmot raw landmark coordinates, for instance, this first line is: 1 356 24 0, where 1 indicates a rectangular matrix; 356 is the sample size N excluding outliers and individuals of unknown sex; 24 is p, which is the number of variables (i.e., 12 2D landmarks, each with its X and Y coordinates); zero is a code to tell the software that there are no missing data. If N and p are followed by an L (1 356L 24L 0), then one can have more information pasted between the first line and the those of the data matrix. Thus, on line two, there might be 356 names for the specimens, followed on line three by 24 names for the variables. For both, specimens and variables, names must be a single word without blanks, separated by the next name by blanks (or tabs or, in fact, even placed on different lines). The names of the specimens could be the same as the ID used in MorphoJ's TXT files, and those for the coordinates could be X1 Y1 X2 Y2 … X12 Y12.

Together with the NTS file with the landmark coordinates of the 356 individuals of known sex, a second NTS file is necessary to specify the independent variables, on which to regress shape data. This is the MANOVA 'design matrix', consisting of dummy variables for the same 356 individuals in exactly the same order as in the NTS file with the landmark coordinates. For convenience, as in the CS species by sex ANOVA in PAST, a desirable order has the species, one after the other, and, within each species, all females followed by all males (or vice versa). Building a design matrix in a spreadsheet is tedious, but easy. The design matrix has three blocks of dummy variables. The first is a single variable for sex, with females coded -1 and males 1 (or vice versa). The second block, for species, has g-1 dummy variables, where g is the number of taxa, which in my case is six species. This marmot 'species' block is, thus,

[12] Using type III SS as in TPSRegr, the MANOVA can be specified with the command: library(car); *Anova(lm(Y~species*sex, contrasts=list(species=contr.sum, sex=contr.sum)), type=3, test.statistic="Wilks")* with Y being the matrix of the PCs (with non-zero variance) of the Procrustes shape coordinates; if the interaction is not significant, *species*sex* is replaced by *species+sex* to replicate the MANOVA without interaction.

[13] Using type I SS (which is not used in TPSRegr and might not be the most appropriate for testing the species by sex interaction) and 10 000 permutations, the MANOVA is specified with the command: library(vegan); *adonis2(Y~species*sex, permutations=9999, method="euc").*

made of five variables, each contrasting one species (I used VAN, the last in the matrix, but the decision is arbitrary) with any of the remaining five. For example, for contrasting VAN with the Alaskan marmot, I coded the former as -1, the latter as 1, and all other species as zero. For VAN vs hoary marmots, VAN is unchanged (coded -1) and hoary marmots are coded 1, with the other four species coded as zero. The third contrast will be VAN vs yellow-bellied marmots, the fourth VAN vs woodchucks, and the fifth VAN vs Olympic marmots, with these last three dummy variables built using the same rationale as for the previous two. Finally, the third block codes the interaction (species by sex) using another five dummy variables. The 'interaction' block is the easiest to build, as it is obtained by multiplying the values of the sex dummy variable (first column) by the corresponding value of each of the five species dummy variables (columns 2, 3, 4, 5 and 6). From the spreadsheet, the variables are pasted into a TXT file, with the extension renamed as *.nts. Thus, the first two lines of the NTS design matrix file will be like these for my data:

```
 1 356 11L 0
 sex sp1_sp6 sp2_sp6 sp3_sp6 sp4_sp6 sp5_sp6 sp1_sex sp2_sex sp3_sex sp4_sex sp5_sex
```

On the following lines, there will be numbers corresponding to the sex, species and interaction dummy variables, with the individuals ordered as in the NTS file with the raw coordinates.

With the two NTS files built as explained above, the MANOVA can be run in TPSRegr. Landmarks are loaded as *data* and the design matrix as *indep. var.* (independent variables). Clicking on *Consensus* does the Procrustes superimposition. Then, one has to click on *Partial warps* (see previous explanations on why this type of variables is used, and remember to skip any part of the output concerning partial warps), followed by another click on *Regression*. This is the first of three series of two regressions, with each pair of regressions testing one factor (i.e., one pair of regressions is needed to test the species by sex interaction; a second pair will be used to test species, and a third pair to test sex). All six regressions are necessary to obtain the same output as in a species by sex MANOVA performed in R or a commercial software. However, the most important pair of regressions is the one testing the species by sex interaction, which is the one I now describe. Thus, the very first regression represents the full model, as all independent variables are included. I call, for brevity, the full model 'species by sex', but I stress that it includes all blocks of dummy variables, i.e. species, sex and species by sex. For the second regression, testing the species by sex interaction, one employs a reduced model that I call 'species plus sex', because it only employs as predictors the species and sex blocks of dummy variables. Therefore, even if apparently counter-intuitive, in this second regression one must exclude precisely the factor (i.e., the block of dummy variables) a user wants to test. As I am testing the interaction, I leave out of the second regression all the variables in the third block (*Options, Select independent variables*, uncheck sp1_sex, sp2_sex etc.).

To explain the rationale of the procedure by which a block of predictors is excluded, I go back to the interpretation of Rsq. Rsq is, we said, the proportion (or percentage) of variance in the data accounted for by a factor. Put it another way, Rsq measures how well the data (shape, in this case) fit the model (predictors). Any additional predictor explains, even if merely by chance, an extra amount of variance in the data, which adds up to the variance already explained by the other predictors. This is why a full model (all predictors) always has a larger Rsq compared to a reduced model (some predictors excluded). However, if the predictors left out in the reduced model (i.e., the dummy variables coding the interaction in my first pair of regressions) have a negligible effect, the decrease in its Rsq compared to the full model will be very small and statistically negligible (i.e., in this case, a non-significant interaction).

After running a regression, the *File, View report* window of TPSRegr shows the numerical output of the analysis. TPSRegr, instead of the Rsq, shows one minus the Rsq, expressed as a percentage. From this percentage, corresponding to the variance unaccounted for by the regression, one easily computes the Rsq (100% minus the percentage of ‘unexplained’ variance). Finally, the difference in Rsq between the full (all dummy variables including the interaction) and reduced (species by sex dummy variables excluded) models represents the variance accounted for by the interaction. I emphasize again that the variance due to the interaction should be small, if SDM is similar across all species. TPSRegr computes the P value for the significance of the interaction using an approximated F ratio. The F ratio allows to estimate whether the full model (larger Rsq) is really ‘better’, in terms of variance explained, than the reduced model. More accurately, the P value for the interaction F ratio estimates the compatibility of the data with the null hypothesis that the full and reduced model are equally good. To obtain this test, users MUST remember to select *Options, Retain current resid. SS* before running the second regression. This is fundamental, otherwise the second regression is run correctly, but the software ‘does not know’ that it has to compare the fit of the first and second regression. It is this comparison, that provides the test for the effect of the factor (the interaction, in this case) that was excluded in the second analysis. For instance, after running the first regression of marmot mandibular shape onto all 11 dummy variables in my dataset, followed (having retained the residuals!) by a second regression using only the first six dummy variables (i.e., sex sp1_sp6 sp2_sp6 sp3_sp6 sp4_sp6 sp5_sp6), the relevant results in the *View report* window will be (with my comments in square brackets):

[first regression with all dummy variables, interaction included]

Percent unexplained = 75.4% [full model Rsq = 100% - 75.4% = 24.6%]

[second regression with only dummy variables for species and sex and the option *Retain current resid. SS* checked]

Percent unexplained = 76.2% [reduced model Rsq = 100% - 76.2% = 23.8%]

*** Testing difference between current residual SS matrix [the variance left unexplained by the reduced model] and the residual SS matrix retained from previous analysis. *** [Please, notice that this part of the output will not be there if one forgets to check the option *Retain current resid. SS*]

```
Multivariate tests of significance:
Statistic         Value   Fs      df1     df2      Prob     [Rsq]
Wilks' Lambda:    0.754   0.946   100     1590.2   0.6322   [24.6%-23.8% = 0.8%]
```

Thus, the species by sex interaction accounts for just 0.8% of total shape variance, which corresponds to a non significant (P = 0.6322 >> 0.005) F ratio of 0.946. This is a strong indication that SDM is statistically similar in all species.

Using the same logic, a researcher can test sex by including all predictors except the sex dummy variable to have a new reduced model to compare with the full model (all 11 dummy variables) in the first regression. Similarly, he/she can test species by, this time, excluding the species block (from the 2nd to the 6th dummy variable in my NTS file). These two main factors, however, are really important and worth being tested only when the interaction is not significant[14]. Then, with a non-significant interaction, to specifically test species and sex, it is better to re-run the two-way MANOVA without the species by sex variables. This is because, having demonstrated that the species by sex interaction is negligible for shape, the researcher no longer needs a full model that includes an interaction, which is too small to be important. The regression approach will compare again a set of reduced models with a full model. However, in this ‘species plus sex’ MANOVA, the full model becomes a regression of shape onto sex and the species block only: the ‘new’ full model is, thus, the one previously used as a reduced model, when the interaction was being tested.

[14] Overall results will be the equivalent of those obtained in the species by sex ANOVA of CS run in PAST, except for the different type of SS. However, as mentioned, PAST does not allow to re-run the ANOVA without interaction, if the latter is small and non-significant. In contrast, this can be done in TPSRegr, as described below.

For the 'species plus sex' MANOVA, even if the shape data are the same, it is better to shut down TPSRegr and reload both the raw coordinates and independent variables. This guarantees that no part of the output of the previous analyses is left, that might confound the results. Thus, after pressing *consensus* and *partial warps*, for a 'species plus sex' MANOVA, the first regression has both main factors as independent variables, which in my NTS file are coded by the first six dummy variables (*Options, Select Indep. variables* to include only: sex sp1_sp6 sp2_sp6 sp3_sp6 sp4_sp6 sp5_sp6). Then, to test sex, in the second regression, one has to exclude the sex dummy variable, check the *Retain current resid. SS* option, and run the regression using only the five dummy variables of the species block (sp1_ sp6 sp2_sp6 sp3_sp6 sp4_sp6 sp5_sp6). TPSRegr will compare the fit of the new full model ('species plus sex') with that of the new reduced model ('species' only as a predictor) to assess if omitting sex leads to a non-negligible loss of fit. The "Multivariate tests of significance" in the *View report* window ("*** Testing difference between current residual SS matrix and the residual SS matrix retained from previous analysis. *** ") is, now, testing the significance of shape SDM. To test species, a researcher will redo everything the other way round: shut down and restart TPSRegr, load the data and get the shape variables, regress them first on the first six dummy variables (sex sp1_sp6 sp2_sp6 sp3_sp6 sp4_sp6 sp5_sp6) and then, having retained the residuals, regress them again including only the sex dummy variable as predictor.

As in all analyses, also in the ANOVA, plotting the data is complimentary to testing. Variability in CS is easily visualized using box and jitter-plots. Profile plots can also be useful for univariate data. If the interaction is not significant, a profile plot of sex means (vertical axis) vs species (horizontal axis) should have lines connecting female means that are approximately parallel to those connecting male means. If they diverge strongly or cross, then there likely are important differences in magnitude and/or direction of SDM. A profile plot can be obtained in PAST after computing the CS means in a spreadsheet or in MorphoJ (after subdividing species, the command is *Preliminaries, Average Observations By* sex). Data in PAST will have to be organized in two columns, one for female and the other for male means, with species in rows. After selecting the two columns, the plot is requested with *Plots, Graph* using the *Line* option.

For multivariate data, ordinations and phenograms should show a within-species mix of females and males well separated from those of other species, if SDM is negligible (see also B1), and the opposite (i.e., a mix of sex and species or even same-sex clusters of different species), if SDM is large. However, there is no simple equivalent of profile plots for shape to graphically explore the species by sex interaction. Yet, having assessed that the direction of shape change between sexes is approximately similar across species using the visualization of female to male mean shape change (but see B1 for more on this) and focusing on the magnitude of SDM, a researcher could check whether the female to male mean shape distances are about the same in all species. This would be consistent with a negligible species by sex interaction in the MANOVA. To obtain the shape distances between sex means, users need a matrix with those mean shapes to load the data either in PAST (then, selecting the matrix and using the command *Statistics, Similarity and distance indices*, with *Euclidean* selected as a metric) or in TPSSmall (using an NTS file, clicking *data* to load it, then pressing co*mpute* and using *File, Save, Procrustes distances* to output the distance matrix). The result is a square symmetric matrix of, for the North American marmots, 66 pairwise distances. However, in this matrix, one will need to inspect and compare only the six within-species mean female to male distances.

A faster, graphical, approach to compare shape distances among means is to compute a phenogram (see A for more information on UPGMA phenograms using Euclidean distances and how to obtain them in PAST). If shape SDM is similarly small in all species, the phenogram of the 12 mean shapes should show the average female of each species paired with the average male of the same species. The shape distance, and thus the branch length of each pair of female and male mean shapes, should also

Table 3. Group differences in CS: species by sex ANOVA in PAST*. As in part A, in this and other tables, SS = sum of squared deviations from the factor group mean; df = degrees of freedom; MS = mean sum of squares, i.e. SS / df.

Samples	Factor	Ss	Df	Ms	F	P	Rsq
all species	species	14770.0	5	2954	127.7	*<0.000001*	63.9%
	sex	312.5	1	312.5	13.5	*0.0003*	1.3%
	interaction	87.0	5	17.4	0.8	0.5849	0.4%
	residual	7954.0	344	23.1			34.4%
	total	23123.5	355				
large samples only (i.e., cal, fla, mon)	species	11030.0	2	5514	223.7	*<0.000001*	58.3%
	sex	200.4	1	200.4	8.1	*0.0050*	1.1%
	interaction	72.4	2	36.2	1.5	0.2320	0.4%
	residual	7618.0	309	24.7			40.3%
	total	18920.8	314				

* As mentioned, PAST 2.17c does not provide details on the type of SS used in the ANOVA, but a comparison with results from R suggests that they are either type I or, more likely, type II.

be similar and, if SDM is small, short compared to interspecific mean species distances. In contrast, if the length of the branches varies (e.g., is short between female and male means of woodchucks and long between those of Olympic marmots) and/or sexes do not consistently form pairs within species, the phenogram suggests a variable amount of SDM and, therefore, a likely interaction between species and sex effects.

Results (B2)

Results of the species by sex ANOVAs are reported in Table 3 for CS and Table 4 for shape. In both tables, analyses include either all species or only the three species (woodchucks, hoary and yellow-bellied marmots) with the largest samples. For CS, I focus on results including all species, as they are in very good agreement with those of the largest samples. For shape, however, there are some small differences. They do not substantially change the conclusions, but results with or without small samples are reported and briefly compared.

The ANOVA confirms that SDM in size is very small (Rsq ≈ 1%), although, by pooling all species samples, the ANOVA achieves such a high power that SDM is significant ($P \leq 0.005$). Interspecific differences are also highly significant (P < *0.000001*), but, unlike the small effect of sex, they are very large (Rsq = 58-64%). This means that, for mandibular size, interspecific differences are ~ 60 times larger than SDM. The species by sex interaction is not significant and hardly accounts for any variance in CS (Rsq = 0.4%). That the pattern of SDM is similar across all species is supported by the almost parallel lines in the profile plot of female and male mean CS (Fig. 3a). The profile plot also suggests very small mean sex differences (with males a few mm larger than females, on average) and large interspecific variation (average CS from little less than 70 mm to almost 90 mm). Although PAST does not allow to repeat the ANOVA after excluding the interaction, its effect is so small that it cannot appreciably change results.

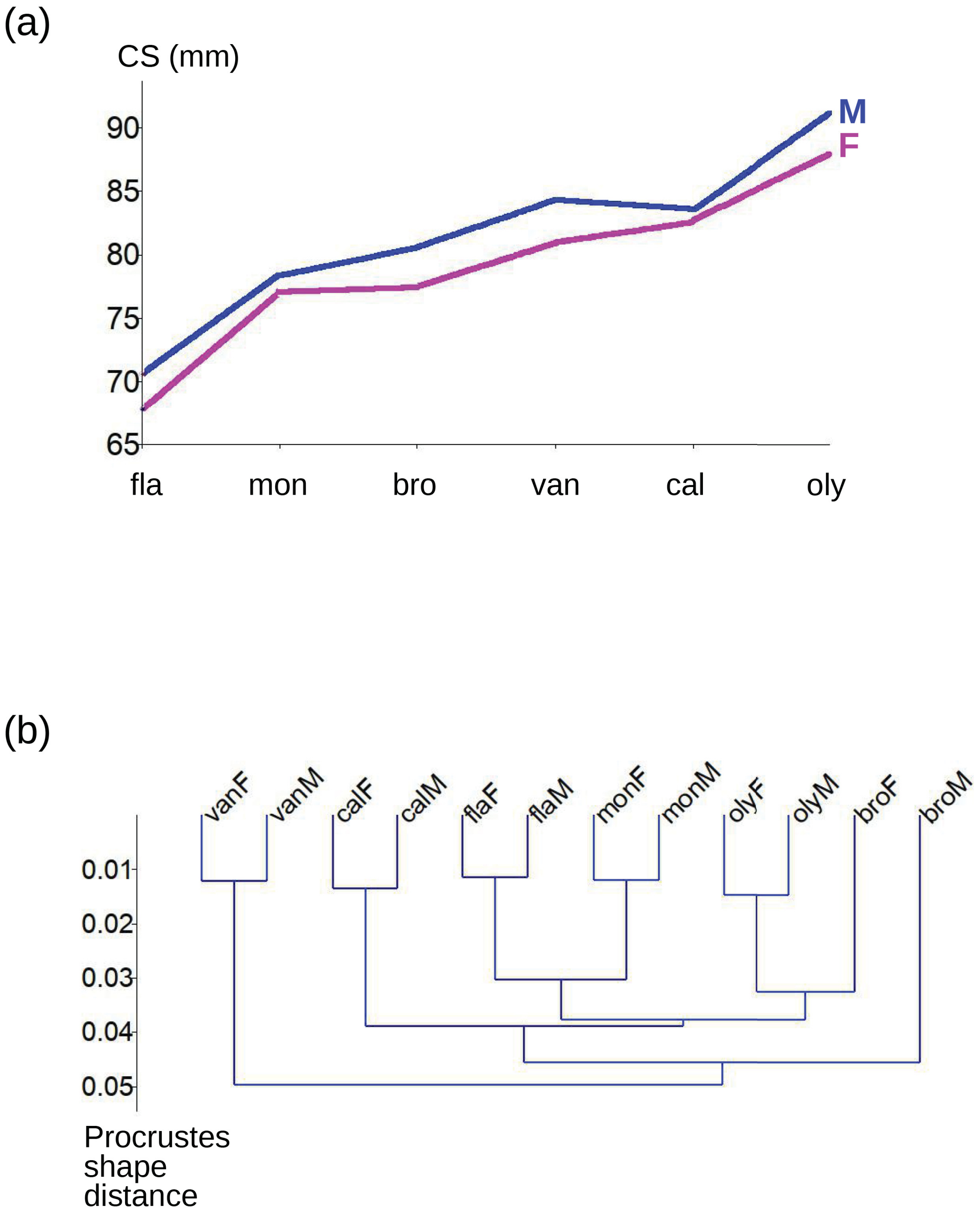

Fig. 3. Visualizations of species by sex interactions using group means.**a**. Mean CS profile plot. For size, males are on average slightly larger than females, but the difference is small and roughly similar in all species. Thus, lines are approximately parallel in the profile plot. **b**. Phenogram of mean shapes. In the phenogram, with the exception of the Alaskan marmot (bro) (whose sampling error is huge, having only eight individuals of known sex), female and male means are paired within each species with almost identical shape distances between sexes in each species. The similarity of SDM shape distances provides an information equivalent, in terms of the magnitude of the sex differences, to that of the parallel lines in the CS profile plot (Fig. 3a).

Table 4. Group differences in shape: species by sex MANOVA using dummy variables in TPSRegr[*].

Model & samples	Factor	Wilks' Lambda	Fs	Df1	Df2	P	Rsq
all species	species	0.026	17.6	100	1590	*<0.000001*	22.3%
	sex	0.919	1.4	20	325	0.1083	0.4%
	interaction	0.754	0.9	100	1590.2	0.6322	0.8%
	sp.+sex+interaction						24.6%
all species	species†	0.026	17.9	100	1614.6	*<0.000001*	22.5%
no interaction	sex	0.789	4.4	20	330	*<0.000001*	1.0%
	sp.+sex						23.8%
large samples only	species	0.088	34.3	40	580	*<0.000001*	17.0%
(i.e., cal, fla, mon)	sex	0.793	3.8	20	290	*<0.000001*	1.0%
	interaction	0.835	1.4	40	580	0.0695	0.6%
	sp.+sex+interaction						19.1%
large samples only	species†	0.088	34.6	40	584	*<0.000001*	17.3%
no interaction	sex	0.773	4.3	20	292	*<0.000001*	1.1%
	sp.+sex						18.5%

[*] TPSRegr uses type III SS both in MANOVAs and MANCOVAs (i.e., the tests for slope and intercept of allometric trajectories - see B6).

† If species are tested in a one-way MANOVA, regardless of sex, Rsq using all species or those with the largest samples are respectively 22.7% and 17.5%.

The species by sex shape MANOVA shows a small (Rsq = 0.6-0.8%) and non-significant (P = 0.6) interaction. After excluding the smallest species samples, despite the slightly smaller Rsq, however, the P value of the species by sex interaction decreases almost 10 times (P = 0.07). Thus, including only large samples, the interaction remains non-significant, but becomes closer to significance using a conventional alpha = 0.05 instead of the more conservative 0.005 threshold adopted in this study. This is a powerful demonstration of the potential influence of small samples on the results of statistical tests with a highly unbalanced design. Nonetheless, the interaction not only is always non-significant, but also, regardless of including the smallest samples or not, very small and, more precisely, 28 times smaller than the Rsq for species differences. Thus, the MANOVA can be repeated without the interaction to compare the magnitude of species and sex differences. In this second analysis (the 'species plus sex' MANOVA), SDM accounts for ~ 1% of shape variance, whereas species differences account for 22% and 17% of total variance, respectively including or excluding the smallest species samples. As with CS, both species and sex are significant. However, the effect size of species is ~ 20 times larger than that of SDM. That SDM is negligible compared to interspecific variation is supported also by ordinations, as shown in B1 (Fig. 2). Besides, the phenogram of species mean shapes, split by sex (Fig. 3b), confirms that the magnitude of SDM is generally small and similar in all species. With only one exception, female and male mean shapes cluster as 'sister' within species and mean female to male distances are almost identical in the phenogram. The exception is the Alaskan marmot whose female and male mean shapes do not cluster together, but are based on tiny samples (see Discussion). In conclusion, therefore, also for shape, SDM is similarly small in marmots, which justifies pooling sexes in all further taxonomic analyses of size and shape.

Discussion (B2)

The species by sex ANOVAs confirm the findings of the tests of SDM one species at a time. SDM in marmot manibular size and shape is negligible, especially if compared to the large interspecific differences. Results are robust to the exclusion of the smallest samples, although for shape the P value of the interaction term becomes almost 10 times smaller when they are excluded. Nonetheless, despite the increased statistical power using only the largest samples, the interaction remains non-significant (P > 0.05). More importantly, the Rsq of the interaction is approximately unchanged, with or without small samples, and remains consistently very small (< 1%).

The largely negligible interaction between species and sex is supported also by the graphical summaries for mean size (Fig. 3a) and shape (Fig. 3b). Male mandibles are consistently larger, on average, than female ones in all species, but the difference is always very small. In terms of magnitude, also shape SDM varies little among species and is smaller than interspecific differences, as evident in the phenogram. The Alaskan marmot is the only species, whose female and male mean shapes are not 'sisters' in the phenogram. In this species, with just three males and five males, sample means likely behave like 'outliers' (see Appendix A and Cardini *et al.* 2019). This is also the most plausible reason why the Alaskan marmot shape SDM Rsq, in the within-species tests (B1), was almost five times larger than the average in other marmot species.

Using the Rsq in the results of the species by sex ANOVA, a researcher can relate the magnitude of SDM to the magnitude of interspecific variation. Interspecific differences are, thus, found to be ~ 30 (shape) to 60 (CS) times larger than SDM. This observation, combined with the demonstration of similar SDM patterns in the mandibles of all species (the negligible species by sex interaction), provides a strong justification for pooling sexes in the taxonomic analyses. This conclusion is robust, even if SDM was statistically significant in the 'species plus sex' ANOVAs. As emphasized in part A, statistical significance should not be rigidly taken at face value and must be carefully interpreted. SDM is statistically significant, because the ANOVA (B2) is more powerful (N is larger) than tests of SDM one species at a time (B1). Statistical power becomes so large that allows to detect a small, but real effect (the tiny differences between female and male mandibles). However, this tiny effect, compared to a much larger interspecific variation, cannot appreciably bias the results of the taxonomic comparisons.

Once the interaction has been removed, results of the 'species plus sex' ANOVA (B2) should be very similar to those of the 'ME ANOVA' of part A. The level of individual variation is present only in the 'ME ANOVA', because of the replicates used to assess ME. In the main taxonomic analyses of part B, in contrast, replicates have been averaged and outliers removed. Averaging replicates somewhat reduces the total sample variance. Yet, if ME is small, the difference should be modest. With highly unbalanced samples, however, the congruence of the tests for species and sex in the two types of ANOVAs is less certain. This is because the 'ME ANOVA' in MorphoJ (Klingenberg *et al.* 2002) and the species by sex ANOVA, typical of GMM taxonomic studies (Rohlf *et al.* 1996; Corti & Rohlf 2001; Frost *et al.* 2003), employ models that take into account the effect of unbalanced sample size differently. This advanced statistical topic is covered in most statistical textbook, such as Howell (2013), where it is discussed in relation to the choice of type of SS that was briefly mentioned in the Methods. In general, when a researcher finds appreciable differences in the results of the 'ME ANOVA' in MorphoJ (which is hierarchical) and the 'species plus sex' ANOVA (based on a different, non-hierarchical, type of SS), that might be a warning that there are problems with the data: for instance, sample size is small and/or very heterogeneous or the assumption of homoscedasticity is not met[15]. For marmot mandibles, comparing Tables 4–5 of part A1 with Tables 3–4 of part B2, and focusing in A1 on results without outliers using the 12 landmarks configuration and in B2 on those including all species (because they are based on exactly the same data), one finds that: species explain 64% of CS variance and 22% of shape variance in both ANOVAs, and sex explains 1.5% and 1.0% of respectively CS and shape variance in the 'ME ANOVA' and, for the same variables, 1.3% and 1.0% in the species by sex ANOVA. Thus, with mandibular CS and shape in North American marmots, there is an almost perfect congruence between the analyses, regardless of the type of SS.

The 'ME' (A1) and 'species by sex' (B2) ANOVAs differ for the choice of type of SS, but also because interactions are not tested in the 'ME ANOVA'. With three factors (species, sex and individual), potentially there are three two-way interactions (the pairwise combinations of all three factors), as well as one three-way interaction (species by sex by individual). Why are interactions among factors left out of the 'ME ANOVA'? Mainly, to keep the analytical design simple in a preliminary analysis of ME. This simplification may be arguable in A1, but it is definitely inappropriate in B2, where the aim of the ANOVA is to decide whether or not we can pool sexes in the taxonomic comparisons. If the interaction is significant and has a large Rsq, the effect of sexual dimorphism varies depending on the species, which makes highly inaccurate pooling sexes. By pooling sexes in species with different patterns of SDM, I could end up comparing, for instance, the mean of one species with no or very small sex differences and the mean of a species with very large SDM. The first species would have a mean, which accurately

[15] If the ANOVA model was the same in both the 'ME ANOVA' in part A and the species by sex ANOVA in part B, results should be the same. Two-way (and, more generally, multi-way) ANOVAs have different options to partition the total variance of the data (more precisely their SS) among the factors being tested. When the ANOVA design is perfectly balanced (i.e., sample size is the same in all groups), all types of SS produce identical results and the distinction becomes irrelevant. However, as already mentioned, in taxonomy and descriptive evolutionary studies, one rarely has the luck of equal sample size across groups. The choice of the type of SS is, thus, an important but rather technical topic, with disagreements even among experienced statisticians about the 'best' option. Readers who want to learn more can look at the abundant literature on ANOVA types of SS. I suggest Howell (2013: 444–446, 587–590) as a concise but relatively simple introduction. A brief but interesting introduction to the meaning of interactions and different types of ANOVA is also found in the Appendix of Herler *et al.* (2010). Below, in this note, I provide an informal, concise, comment on the three main types of SS, so that beginners have at least an intuition of their meaning. However, I stress again that this is a really technical topic, which I am not an expert of and find difficult to fully understand (a very good reason for readers to look for more accurate explanations in the statistical literature!).

a. The 'ME ANOVA', as done in part A, is hierarchical (model I SS). I have explained (A1) why the hierarchical design is reasonable in that specific context. Here, I remark that the order in which factors are entered in a model I SS ANOVA is important; if altered, results will be different. In marmots, it is known from the literature that species tend to show large differences and have a small SDM. Thus, I first controlled for species differences (first factor in the model) and then for any potential SDM (sex as second factor), before testing individual variation. In the 'ME ANOVA', the main factors of species and sex are, therefore, included mainly to control for their effect, when the main interest is in individual differences compared to ME. In the species by sex ANOVA, in part B, in contrast, species, sex and their potential interaction are all a main focus.

b. The principal alternatives to model I SS are model II and III (Howell 2013). Type III, unlike model II where factors are weighted by the different groups' N, weighs all groups equally regardless of whether one is larger or smaller, as it happens when N varies across species and sex. This choice seems appropriate in taxonomy, as a variable N is not usually due to an underlying biological model that justifies different sample sizes across groups. Heterogeneous sample size in taxonomic studies is, generally, just a matter of what was available to measure. Type III SS are also called "unique SS" (Howell 2013), because SS which are accounted for by multiple factors (e.g., differences between species that are also part of within-species SDM) are left out of the model, so that each factor is assigned only the proportion of SS uniquely explained by that given factor. As a consequence, the sum of model III SS of all factors might not equal the total SS of the variables (e.g., $SS_{species} + SS_{sex} + SS_{species\ by\ sex} + SS_{unexplained} \leq SS_{total}$). This is why in Table 4 the sum of Rsq of each factor is slightly less than in a regression including all of them together.

c. Different types of SS have pros and cons and the choice might depend on the data and the specific aims of the analysis. Rohlf's TPSRegr employs model III SS and Howell (2013) also favours this model, which is the default option in some of the main commercial statistical programs. Others, especially in the R community, prefer model II SS. I do (and can) not argue about the 'best' choice. In this study, the decision was restricted by what is available in the free software I am using: MorphoJ uses model I, TPSRegr uses model III, PAST 2.17c does not specify the type of SS. In my personal experience, most of the time different ANOVA models lead to the same conclusions unless data are particularly problematic (e.g., very heterogeneous N, with several very small samples; strong deviations from the assumption of homoscedasticity; possibly, a large p / N ratio). With the marmot mandibles of this study, for instance, I repeated the ANOVAs in R using different types of SS, as well as including only the three largest samples or their random subsamples to have a perfectly balanced design. Results were highly congruent with those obtained in MorphoJ, PAST and TPSRegr.

summarizes morphological variation in that species. The second species, in contrast, would have a mean that is somewhat halfway between the female and male phenotype. In such a case, averaging sex might be biasing the results of interspecific tests, because one cannot accurately control for the effect of sex on species comparisons. The issue is well known and, yet, the importance of correctly taking sex into account in interspecific studies using GMM is often ignored. In macroevolutionary studies, the error introduced by comparing species of different genera or even families of a lineage regardless of sex, when SDM varies across group, might be less serious, but it is still there and certainly contributes to inaccuracies. For instance, Castiglione *et al.* (2019) do not mention sex in their broad study on mandibular shape convergence in ungulates, suggesting that SDM was not taken into account. This is very concerning in a group whose SDM varies so widely (depending on the ecology, mating system and type of social organization) that for decades they have been a model for studies on how secondary sex differences originate in mammals (Jarman 1974; Loison *et al.* 1999; Ruckstuhl & Neuhaus 2002; Pérez-Barbería *et al.* 2002).

The consequences of neglecting SDM in taxonomic and, more generally, evolutionary studies can be serious. I use also a made-up example to make the problem more tangible. I focus on size for simplicity, but the rationale will be analogous for shape. Let us say that I have two populations P1 and P2, whose mandible size is similar; there is also a degree of SDM, which is of similar magnitude and direction: in both, males are ~ 20% larger than females. However, I do not know in advance anything about these differences and, using samples, I want to discover whether mandibular size is the same or not in P1 and P2. Unfortunately, I have unbalanced samples with very few females in P1 and very few males in P2. To maximize N, I decide to pool sexes before comparing the two populations. Without a ‘sex-correction’ (which is rarely straightforward, as I mentioned in the ‘Material and methods’), the mean of P1 is male biased (‘looks larger’) and the mean of P2 is female biased (‘looks smaller’). Thus, because the impact of SDM was overlooked, a mere artefact of sampling error leads to a comparison that overestimates populations differences, which I may find to be large and significant, when there was none. Even if I used a weighted mean by computing the mean of the female and male means within each population, the mean of the sex with the smaller N would be less accurate (and maybe look like an outlier if N is really small) and, as a consequence, the mean population difference would also be inaccurate. If the samples were balanced (same or almost same N for females and males of both populations), pooling sexes would be less problematic. The mean of P1 would be halfway between the female and male mean and the same would happen for P2. The relative mean difference in size of P1 and P2 would, therefore, be accurate. However, sample variance is poorly estimated in both examples (either with unbalanced or balanced N), because sex differences and population variability within species are mixed up.

Besides, with a large SDM, even when N is the same for females and males, the total sample average is an abstraction, because it corresponds accurately neither to the female nor to the male mean of each population. Therefore, even with balanced samples, if pooling sex is necessary to increase statistical power, I prefer a ‘sex-correction’. For a ‘sex-correction’ to be accurate, however, SDM must be similar in all taxa, which is why it is crucial to test the species by sex interaction. For size, similarity of SDM means that males are larger than females by approximately the same amount in all populations or species (and vice versa, if females are larger). For shape, this is more complex, because a ‘sex-correction’ assumes not only that female and male means have similar Procrustes distances in all taxa (i.e., similar magnitude), but also that the variance-covariance structure of sex differences is approximately the same in all species. Testing the similarity of variance-covariance matrices requires large samples (Cardini *et al*., 2021, and references therein). With small samples, one might, at least, visually compare shape diagrams to verify that the mean difference between sexes suggests roughly similar shape changes across all species. If SDM is similar, for instance, displacement vectors should be similar in length and have approximately the same direction in all species. In contrast, the plot of the interaction between species and sex using a phenogram, which I used in Fig. 3b, is only based on Procrustes shape distances. Mean shape distances quantify the magnitude of SDM, but say nothing on direction. Thus, one might have a similar amount of SDM in two species, despite different or even opposite patterns of shape change. This is like, for a univariate trait, having the same absolute difference between sexes (say, a 10% average difference in size between females and males), but in opposite directions (i.e., females 10% bigger than males in one species and the opposite in the other species). As I discussed in B1, patterns of SDM can be visualized and compared qualitatively with shape diagrams (Klingenberg 2013), but also numerically assessed in terms of similarity of directional change by computing pairwise species angles between vectors of mean shape differences.

Carefully testing and controlling for the effect of sex on taxonomic comparisons is important. SDM, however, is not the only potential source of within-species variability that may confound taxonomic analyses. Age-related differences are another potential main factor. Usually, to work with homogeneous samples of comparable age, taxonomic studies focus on a specific ontogenetic stage. In mammals, and endotherms more generally, growth slows down after sexual maturity and, with rare exceptions such as elephants (Perry 1954), almost completely stops in fully adult individuals. Thus, adults are a convenient option for taxonomic studies. Besides, they have the advantage of being the most common age groups in museum collections. Using adults, therefore, helps to maximize sample size. However, within-species group variability (sex, age, different morphs etc.) can be a source of taxonomic information in itself and, whenever possible, it is interesting to have data covering the entire life cycle of an organism. For instance, as long as all main ontogenetic stages are adequately sampled, we might expect that developmental trajectories diverge because of evolutionary separation.

In ectothermic vertebrates and in many invertebrates, with the clear exception of holometabolous insects, growth may not stop after sexual maturity. Even in endotherms, in fact, ageing in adults can change morphology, something well known in humans (de Groot *et al.* 1996; Smith *et al.* 2021). Changes in adult mammals and birds, however, tend to be relatively small, whereas they can be large in ectothermic animals. This additional, and potentially large source of variability, makes comparisons potentially more challenging. For instance, if, in a species of lizard, one happens to sample young adults or underfed individuals and in another one adults that are older or simply better fed, despite including in both species only sexually mature specimens, size might differ because of adult age or food abundance. Age and condition-dependent size differences can introduce a bias, if they are not controlled for. Because size changes generally influence shape, a similar bias might also affect shape. ‘Size-corrections’ (see B6 in the Discussion) may help to mitigate the problem, but they are not as straightforward and almost automatic as sometimes believed. To avoid potentially misleading results, it is important to understand the methods, but an in-depth biological knowledge is the basis for a good analytical design, which must be tailored to the specific study organism.

To end this subsection, a brief comment on a methodological detail, which most of the time does not make any practical difference, but users should be aware of, and a second brief comment on a related issue, which also concerns parametric tests in TPSRegr but is potentially more important.

In the TPSRegr output for the species by sex interaction of the shape MANOVA, I selected, as a test statistics, the Wilks’ Lambda. Wilks’ Lambda is commonly reported for multivariate tests in commercial software, as well as in PAST. TPSRegr, however, provides four different multivariate tests in its report window. The Pillai’s trace, which is shown in the second line, after Wilks’ Lambda, is also a common multivariate test statistics. Pillai, for instance, is used in MorphoJ’s Procrustes ANOVA for the non-isotropic multivariate model (Klingenberg *et al.* 2002). Pillai is said to be more robust to violations of the MANOVA assumptions (Warne 2019), but that might in fact vary from case to case (Ateş *et al.* 2019). The help file of TPSRegr has a fairly detailed information on how the different test statistics are computed and multivariate parametric tests are also briefly discussed in Zelditch *et al.* (2004).

For ‘reasonable’ data, the different test statistics should produce similar approximated F ratios, and therefore lead to similar conclusions. The exception is often the Roy’s root, which is more liberal (rejects the null hypothesis more often) and is, thus, prone to a higher rate of type I errors (i.e., claiming differences when they are not there). This seems to be the case in my example, where all test statistics for the species by sex interaction had P = 0.63–64 in TPSRegr, except Roy’s root, whose P = 0.01. With the Rsq of the full model being only 3% larger than that of the reduced model (24.6% / 23.8% = 1.03), thus suggesting a tiny effect, Roy’s root might indeed be less reliable than the other test statistics in order to decide the significance of the species by sex interaction in marmot mandibular shape.

TPSRegr was written before methods for the analysis of semilandmarks were developed and became popular. If semilandmarks are used in this software, there are some caveats. On semilandmarks, I wrote more in Appendix A. These ‘special points’, used to measure curves and surfaces, are often treated using different algorithms to maximize their mathematical correspondence in a procedure called ‘sliding’ (Bookstein 1996). I stress that this manipulation is pure mathematics with no biological model behind and is not a compulsory step. In certain analyses, such as in studies of modularity/integration it may, in fact, sometimes do more harm than good (Cardini 2019). However, because semilandmarks are often slid during the superimposition, shape variance is reduced and this implies increased covariance and a further loss of information (i.e., degrees of freedom, in statistical jargon) beyond the usual loss of four (2D data) or seven (3D data) degrees of freedom in Procrustean GMM (Rohlf & Slice 1990) and Discussion in B4). For 2D data, for instance, a user may slide the semilandmarks (but see the Appendix A) in TPSRelw, restore scale[16] (read help on “Boas coordinates”), save the new coordinates as NTS and load the data in TPSRegr, or another program of the TPSSeries, for further analyses. In studies of shape, that means that the superimposition will be done again on the slid-rescaled data, an operation which typically does not change the shape coordinates. Yet, TPSRegr (or MorphoJ) does not ‘know’ that semilandmarks had been slid and, therefore, it might incorrectly compute degrees of freedom in parametric tests, leading to wrong P values. I emphasize that this only concerns parametric tests, as I discuss later with an example in relation to the DA/CVA (B4). With resampling tests using Euclidean distances in the tangent shape space, degrees of freedom are not computed for the dependent variables and, therefore, P values are correct. Because the TPSRegr species by sex MANOVA (and also the species by CS MANCOVA, in B6) use parametric tests for the factors in the analysis, these tests may incorrectly compute P values using slid semilandmarks. For this type of data, therefore, analyses may have to be run using the matrix of PCs of the Procrustes shape coordinates, with all those with zero variance excluded, in a commercial software or in R.

B3) Pairwise tests of species mean differences

Methods (B3)

From this subsection, the focus finally shifts on tests of size and shape differences between taxonomic groups, which are the main aim of a taxonomic study. If SDM is negligible (as it happens with marmot mandibles) or data can be ‘sex-corrected’ (see previous B2), these and all following analyses can be done by pooling sexes; otherwise, they should be run in parallel in females and males or, when limited by sample size, only in the sex with the largest sample.

[16] Bear in mind that this may not be necessary if further analyses are only done on shape only. However, if the software compels users to re-superimpose the slid data (as it happens in MorphoJ or the TPS Series), the Procrustes shape coordinates will be virtually unchanged but CS extracted from the TPSRelw shape coordinates (*Procrustes aligned specimens*) will be in all specimens equal to one (plus a tiny amount of error due to the approximation of decimals in the computations). It is one because data are already in the shape space with a standardized CS = 1. Yet, it is easy to forget this, so that a user ends up picking up the variable called CS in, say, MorphoJ and using it as if it was the true CS. This would lead to mistakes in all analyses of CS and allometry. For this reason, I suggest to always restore scale in data slid in TPSRelw, before they are re-used in another GMM program of the same series or in MorphoJ.

Table 5. Pairwise permutation tests of mean CS or mean shape interspecific differences, performed in MorphoJ. Rsq above the main diagonal, P values below.

Data	Species	bro	cal	fla	mon	oly	van	Species mean rsq
CS	bro	–	11.7%	20.4%	0.5%	72.4%	25.6%	26.1%
	cal	*<0.0001*	–	64.8%	26.8%	22.8%	0.4%	25.3%
	fla	*<0.0001*	*<0.0001*	–	38.7%	51.5%	58.5%	46.8%
	mon	0.4405	*<0.0001*	*<0.0001*	–	42.3%	28.2%	27.3%
	oly	*<0.0001*	*<0.0001*	*<0.0001*	*<0.0001*	–	37.0%	45.2%
	van	*<0.0001*	0.4416	*<0.0001*	*<0.0001*	*<0.0001*	–	29.9%
shape	bro	–	10.9%	9.0%	8.4%	15.0%	22.4%	13.2%
	cal	*<0.0001*	–	13.2%	20.1%	6.1%	22.9%	14.6%
	fla	*<0.0001*	*<0.0001*	–	10.0%	6.0%	17.0%	11.1%
	mon	*<0.0001*	*<0.0001*	*<0.0001*	–	6.5%	21.5%	13.3%
	oly	*<0.0001*	*<0.0001*	*<0.0001*	*<0.0001*	–	13.3%	9.4%
	van	*<0.0001*	*<0.0001*	*<0.0001*	*<0.0001*	*<0.0001*	–	19.4%

Pairwise interspecific tests of mean differences are done on both size and shape. It is the same test used for SDM within species (B1; for an alternative test, see the Discussion in this chapter). Thus, CS or shape is regressed onto a binary dummy variable coding a pair of species (e.g., coding hoary marmots -1 and VAN 1, in a covariate), with the significance of the corresponding Rsq tested using 10 000 (or more) permutations in MorphoJ. Now, however, the comparison involves pairs of species and has to be repeated for all possible species pairs. With g = number of groups, this means g * (g -1) / 2 pairwise tests, which, with six marmot species, becomes 15 tests for size and another 15 for shape. The usual considerations about the potential inflation of type I error rates apply.

For mandibular size, species differences can be visualized using box and jitter-plots. For shape, both ordinations and phenograms, which I defer to B5, more specifically dedicated to this topic, provide effective summaries of patterns of interspecific variability and similarity relationships. These summary plots are, as usual in GMM, interpreted in combination with shape diagrams (Klingenberg 2013).

Results (B3)

The species by sex ANOVAs demonstrated that SDM in mandibular morphology is negligible, but also that interspecific differences are large. Pairwise tests of mean species differences in size and shape (Table 5) support this observation and provide finer details on the magnitude of interspecific variation.

For CS, all tests are significant except two. The Alaskan marmot and the woodchuck, both member of the subgenus *Marmota*, have similar CS. In these two species, mandibular size is intermediate between those of the species belonging to the subgenus *Petromarmota*, the small yellow-bellied marmot and the other, larger, members of the hoary marmot species complex (i.e.,VAN, Olympic and hoary marmots). Within *Petromarmota*, hoary marmots and VAN do not differ for average size, although the range of CS is slightly larger in hoary marmots (Fig. 1b). Of all study species, the Olympic marmot is the largest on average, but its largest representatives are mostly within the range of variation of hoary marmots (Fig. 1b). Overall, the average Rsq of the pairwise interspecific mean differences in CS is 33%. However, there is a wide range, with Rsq close to 0% in the two pairs of species with non-significant differences, I mentioned above, and an average Rsq ≥ 45% in the pairwise comparisons of the yellow-bellied marmot or the Olympic marmot.

For shape, all pairs of species show significant mean differences. The average pairwise Rsq varies depending on the species. It ranges from 9% (Olympic marmot) to 19% (VAN), with a total average Rsq of 13% considering all 15 comparisons. Patterns of interspecific shape variation are, as anticipated, graphically summarized in B5.

Discussion (B3)

Pairwise tests for mean differences confirm the significant interspecific variation shown in the ANOVAs. Yellow-bellied and Olympic marmots are, respectively, the smallest and largest of all marmots (Armitage 1999) and it is, thus, unsurprising that they have on average, respectively, the smallest and largest mandibles among North American species. Alaskan marmots and woodchucks have intermediate size, whereas both VAN and hoary marmots tend to be similarly large. In terms of body mass, in fact, VAN is somewhat smaller than hoary marmots (Armitage 1999), but estimates of body mass are likely based on smaller samples and are less straightforward than measuring the size of bones. Body mass varies in marmots within the active season and, depending on the length and harshness of winter or the abundance of food in the summer, there might be some inter-annual variability even at the time of emergence and immergence, when wild individuals are weighted (Armitage 2014).

In general, the range of mandibular size variation often largely overlaps and there is more variation within species with larger populations and distribution ranges, as expected because of both adaptive and plastic changes in relation to the broader variability of local conditions. This means, for instance, that the largest yellow-bellied marmots are larger than the smallest hoary marmots, whereas the largest hoary marmots and VAN are almost as large as the largest Olympic marmots (Fig. 1b). It is worth remarking how the comparisons of mandibular CS show the usefulness of post-hoc pairwise tests. The ANOVA (Table 3) of CS had a very large Rsq for species (accounting for almost ⅔ of size variation). Yet, this cannot be taken as evidence that all species show differences, as demonstrated by the two pairs of species with similar average mandibular size (Table 5). On the other hand, with shape, the ANOVA Rsq (Table 4) was much lower (accounting for less than ¼ of variance), but all pairwise comparisons (Table 5) were highly significant and accounted for ~ 10% of variance or more. Indeed, important differences can be found among species means despite overlaps which are evident for both mandibular CS (Fig. 1b) and shape (Fig. 2, but also Fig. 4 in B5).

In terms of software and the approach used for comparing species pairwise, I did the tests using regressions on dummy variables in MorphoJ. The reasons are the same as in B1, where the test was used to assess SDM. There are disadvantages, however. Subsets are relatively fast to obtain in MorphoJ by using *Preliminaries, Include or Exclude Observations* and the species classifier to select the appropriate pairs, but running a series of pairwise tests in MorphoJ is tedious, as one has to build 15 subsets (with the corresponding covariates) and run 15 analyses for size and another 15 for shape. If I was only interested in P values, a much faster options is MorphoJ's *Comparison, Canonical Variate Analysis* using species as the grouping variable. The CVA in MorphoJ computes pairwise permutation tests of mean shape differences for all pairs of species. In these tests, the test statistics is the Procrustes shape distance between two means, as well as their Mahalanobis distance, which uses standardized shape variables to express differences in units of standard deviations (SD) (Albrecht 1992). The two types of distances are generally in good agreement for large samples. I usually focus on Procrustes, which measures distances in the untransformed shape space produced by the superimposition and provides results largely equivalent to those of the multivariate regressions. However, the drawback of the CVA tests in MorphoJ is that they are specific to shape (thus, they cannot be run for CS) and also do not compute the Rsq, which is necessary to estimate how much sample variance is accounted by the mean difference between two species. In PAST, users can find parametric pairwise post-hoc tests using the one-way ANOVA for CS (*Statistics, One-way ANOVA*) and the MANOVA/CVA for shape (*Multivar, MANOVA/CVA*), but none of them also computes the Rsq between species pairs. The extra effort to perform the pairwise tests using a series of regressions on species dummy variables is, therefore, worth, because the same identical permutation test is done on size and shape and Rsq are obtained together with the corresponding P values.

B4) Species discriminant analysis (DA)

Methods (B4)

The DA has been briefly mentioned in the Introduction, as well as in part A. A detailed, but clear description of the method is in Albrecht (1992) and Neff & Marcus (1980), and an extensive discussion of its application to Procrustes shape data is in Klingenberg & Monteiro (2005). A DA is also called CVA, when there are more than two groups and the focus is on group separation in scatterplots. However, the two terms are used interchangeably in the literature. Thus, unless I refer to the name adopted by a specific software for this type of analysis, I will be also using DA and CVA interchangeably as synonyms. In fact, a more accurate name would be linear discriminant analysis (LDA), because there is a curvilinear variant of the DA, the quadratic discriminant analysis. The quadratic DA has been developed for non-homoscedastic data (see Appendix A on assumptions), but is less common, requires larger samples for accuracy and cannot be used for ordinations (Neff & Marcus 1980). I am not covering this method, which is shortly described by Neff and Marcus (1980). For simplicity, as in most software, I will use the name DA instead of LDA.

In this study, the DA is mainly employed for group classification. This was the main aim why it was originally developed by Fisher, who "asked what linear combination has the greatest difference of sample means relative to its sample standard deviation" (Anderson 1996: 30). In a DA, a set of multivariate descriptors (shape, for instance) is used to predict groups (species, in my case), after the original variables are standardized and linearly combined to maximize the between group variance (i.e., the mean differences). A new data space is produced, whose dimensionality is g – 1[17]. For instance, using the mandibular shape data with six marmot species, there will be five uncorrelated discriminant axes (or CVs), with the first explaining more between group variance than the second, which in turn accounts for more between group variance than the third, and so on. Unlike in a PCA, the percentages of multivariate variance accounted for by the DA/CVA axes refer only to the fraction of total variance that captures differences between the group means. Thus, if I had just two species to classify, there would be only one DA axis that accounts for 100% of the differences between the two species. However, the group differences captured by this single discriminant axis could be just a small fraction (say, as an example, 10% or just 5% or even less) of the total multivariate variance in the original data space.

The reduced-dimensionality space of a DA/CVA is appropriate to increase the chance of correctly affiliating each individual to one or the other group based on the original predictors. For instance, if interspecific differences are large compared to the variability within species, as it happens with size when the small yellow-bellied marmots are compared to the very large Olympic marmots (see Results), one expects to be able to correctly predict species using mandibular size. This is a trivial case, because the ranges of mandibular size variation do not overlap in these two species. However, when, for example, Olympic marmots are compared to hoary marmots, that are also big, a classification of individuals based on CS might be less simple and a method, such as the DA/CVA, that maximizes between to within group differences, becomes useful to increase classification accuracy.

In the DA/CVA data space, an individual will be classified in the group whose mean is closest to that individual. To be rigorous, this is true if the probability of belonging to any a priori group is considered equal (the default option in PAST – see Albrecht (1992) for an in-depth discussion of different options to perform a DA/CVA, including prior probabilities and weighted vs unweighted analyses). The proximity

[17] This is true unless p, the number of variables in a dataset, is $< g - 1$ and, thus, p sets the limit to the DA/CVA space dimensionality. This is, for instance, the case when the DA/CVA is done using CS and the perimeter of the landmark configuration to estimate size variation, because $p = 2 < g - 1 = 5$. This is a special case, in which the between group DA/CVA space captures the entire variance in the two-variables dataset, even if data are rotated to maximize group mean differences.

of each individual to the group means is calculated using Mahalanobis distances, which are Euclidean distances (i.e., straight lines connecting two observations) in the transformed standardized space of the g -1 CVs. The use of distances measured in SDs seems appealing, as it is easier to interpret. However, as explained in the next paragraph, Mahalanobis distances are accurate only if the assumptions of the DA are not violated. Also, they tend to become larger in relation to the number of predictors (Rohlf 2021) and, therefore, they cannot be easily compared across studies, when the anatomical structures or landmark configurations are different[18].

The DA/CVA, like a parametric MANOVA, assumes multivariate normality and homoscedasticity. DA/CVAs are, by design, prone to overfitting the data (Kovarovic *et al.* 2011; Rohlf 2021). To put it simply, the problem with overfitting is that the analysis is using a data space built by knowing the a priori groups of a sample of individuals to classify those same individuals in those groups. This is a type of circular reasoning that almost always overestimates classification accuracy. To mitigate against it, the classification must be cross-validated. The most common cross-validation is a leave-out jack-knife approach. With this method, the DA/CVA functions are built using N – 1 individuals and, then, used to classify the individual which has been left out. The same procedure is repeated for all N individuals. Because the specimen to be classified is not employed to derive the functions used for its own classification, the 'circular reasoning' is avoided. Therefore, unlike a standard DA/CVA, a cross-validated DA/CVA produces accurate results that do not spuriously inflate group separation. This is true, in general, as long as there are no strong violations of the assumptions of this method and the sample size is large in relation to p, the number of predictors. As in the ANOVA, balanced (or almost balanced) samples, also help to make the analysis more accurate.

A DA/CVA becomes computationally impossible if total N – g < p. For instance, if I had to compare two species samples, each with 10 individuals, I could not do a DA using the 20 PCs of mandibular shape, because 20 – 2 = 18, which is < 20 PCs. Sometimes, to overcome this limitation, DA/CVAs are computed after first reducing dimensionality using a PCA. As usual with dimensionality reduction, this requires great caution and a convincing demonstration that most variance is preserved in the selected PCs. When the DA/CVA is performed after dimensionality reduction, it is also desirable to explore its sensitivity to the inclusion of more or less PCs (Kovarovic *et al.* 2011; Evin *et al.* 2013). Besides, N – g < p is a minimum mathematical requirement, but accuracy usually requires N >> p (Rohlf 2021). As p becomes closer to total N – g , results may be inaccurate even when cross-validated (Evin *et al.* 2013).

For the DA/CVA, PAST offers more flexibility and a more comprehensive output compared to MorphoJ. MorphoJ can only perform DA/CVAs using Procrustes shape coordinates and only predicts group affiliation for two groups at a time. Unfortunately, not even PAST provides options for setting unequal prior probabilities and has no information on posterior and typicality probabilities, on which I provide a basic introduction in the Discussion. Also, if used for scatterplots, PAST 2.17c seems to inaccurately report the percentages of between group variance accounted for by the CVs. These (using weighted means, as explained in the help) are correctly computed in MorphoJ, that can, thus, be used for the scatterplots or just to get the percentages of variance. However, as I suggest below and later discuss, I favour a bgPCA instead of a DA/CVA for group ordinations.

The analysis in PAST is simple to obtain from the menu *Multivar, MANOVA/CVA*. Because at least two variables are needed to predict groups, I used CS together with the perimeter of the landmark configuration to assess the cross-validated classification accuracy of mandibular size. The perimeter is easy to calculate in TPSUtil with the *Compute area & perimeter* option. From the perimeter, I excluded landmark 11, on the mental foramen, because it lies within the mandibular outline in side view, but this

[18] This holds 'by definition' also for Procrustes distances, because different landmark configurations imply incommensurable (and therefore incomparable) shape spaces.

is optional and makes a negligible difference. CS and the perimeter are very highly correlated (within-species $r > 0.97$). Highly correlated (collinear) variables are generally undesirable in multivariate analysis (Hair *et al.* 2013), as the information is redundant and there might be computational problems. However, it is not inappropriate in a special case where the aim is exclusively to estimate the cross-validated classification accuracy for what is, basically, a univariate size measurement based on landmarks. Thus, the classification table is obtained for mandibular size by 'tricking' the DA with two highly correlated size variables.

For shape, the DA/CVA in PAST is done using again the menu *Multivar, MANOVA/CVA*, after, now, selecting all shape PCs with non-zero variance. Non-zero variance PCs are automatically saved in a PCA in MorphoJ[19] and can be easily exported as TXT and imported in PAST[20]. The reason why it is advisable to do the DA/CVA in PAST using the PCs of the Procrustes shape coordinates, instead of the shape coordinates themselves, is clarified later in the Discussion. However, in this study, to assess the sensitivity of the classification accuracy to small differences in the information captured by the shape variables, the DA/CVA on shape is replicated using: (a) all 24 Procrustes shape coordinates, as well as (b) the corresponding 20 PCs; (c) only the first 10 PCs; (d) all PCs of Procrustes shape, recomputed after excluding the two smallest species samples (the Alaskan and Olympic marmots); (e) the first 10 PCs of size-corrected Procrustes shape data (see B6). The shape information in the first two DAs is identical, and, thus, unless there are computational errors, classification accuracy also should be identical. All other DAs of shape, in contrast, use less shape information because of the exclusion of higher order PCs (in c and e, leaving out PCs from the 11th to the 20th), smallest samples (d) or static allometric shape (e).

The output of a DA/CVA in PAST is extensive and includes: a test for the overall significance of group differences (Wilks' lambda or Pillai's trace); the pairwise parametric tests for all pairs of groups (post-hoc multivariate tests); the ordination (CVA) scatterplot; and the *Confusion matrix*, which is the classification table. The test of significance for the overall multivariate difference among groups is identical to a parametric one-way MANOVA, which is why PAST calls the analysis MANOVA/CVA. I skip this part of the output, because a P value for the Rsq of species differences has already been computed in B2–B3. Likewise, pairwise tests for species differences have also already been done in B3. I usually avoid also the CVA scatterplot, as already mentioned. For ordinations of group differences, I opt for a simpler, related method (see next subsection), that tends to be less prone to overfitting, unless one has a very large p/N ratio (Cardini *et al.* 2019; Cardini & Polly 2020; Rohlf 2021). Finally, the fourth part of the output of the DA/CVA in PAST, the classification table, is the most important one, for our aims, because it is the analysis that estimates the taxonomic accuracy of a classification based on size or shape. In order to have the classification table cross-validated, however, users must first click on *Confusion matrix* and then, always!, check the box *Jackknified*. If this is not done, the group prediction is not based on the 'leave-one-out' method and will almost always be biased towards overestimates of the true classification accuracy.

In the 'confusion matrix', the number of individuals correctly classified in their a priori group is on the main diagonal; the misclassified individuals are off the main diagonal. The counts are easily converted into percentages ('hit-rates') by dividing in a spreadsheet by the total sample size of each species, which is reported in the last column of the table. The expectation for large group differences is that cross-validated classification accuracy will be high (close to 100%, if there are very large differences and little overlap among species) and clearly larger than random chance. Random chance, with two groups of

[19] TPSRelw relative warps are identical to MorphoJ's PCs, but only computable for 2D data.

[20] Users should also export the group (species, in my case) classifier and must change ID in the first column of the TXT file with a dot or any label other than ID, which is not accepted by PAST. The group classifier, once the file is opened in PAST, can be used to colour-code groups, as customary in PAST.

Table 6. Cross-validated DA hit-rates, computed in PAST.

Species	CS and perimeter	24 shape coord.	All PCs	First 10 PCs	All pcs without bro, oly	10 PCs size-corr.*
bro	13%	88%	94%	69%	–	56%
cal	61%	81%	87%	81%	92%	84%
fla	85%	71%	90%	85%	90%	85%
mon	38%	79%	91%	81%	92%	77%
oly	79%	79%	71%	86%	–	79%
van	36%	84%	96%	92%	98%	90%
total	60%	78%	90%	84%	92%	82%

* This is exploratory, because slopes are significantly different and, thus,the main assumption of the size-correction is not met.

equal sample size, is 50%; with three balanced samples, it is 33%; with four, 25% etc. The computation is less straightforward if groups have heterogeneous sample sizes, which is the case with my marmot data. Kovarovic *et al.* (2011) suggest a formula and, more importantly, a randomization experiment to estimate random chance expectations empirically. However, this method, originally developed by Solow (1990), is not available in any software I know, but it is not hard to program in R (Evin *et al*. 2013).

Results (B4)

In the cross-validated DA, mandibular size moderately discriminates species (Table 6, first column). The overall average hit-rate is 60%. The lowest classification accuracy (13%) is found in the Alaskan marmot and the highest in the yellow-bellied and Olympic marmots (≥ 79%).

Shape has a predictive accuracy higher than CS, with overall hit-rates ranging from 78% to 92%, depending on the set of shape variables selected for the analysis (Table 6). In terms of species-specific hit-rates, VAN has the highest classification accuracy using shape (92% on average), whereas Alaskan and Olympic marmots have the lowest (on average, respectively 77% and 79%). With shape, however, there is some variability in hit-rates depending on the specific set of shape variables used in the DA (Table 6). Surprisingly, using the 24 Procrustes shape coordinates, the total hit-rate is appreciably lower (78%) than using all PCs of those same coordinates (90%). They should be identical, because the overall information is the same. In the Discussion, I will explain why this is likely to be a computational error in PAST using Procrustes shape coordinates and, thus, another valid reason to employ PCs in the DA/CVA and, in general, in most analyses of shape in PAST and other programs, which are not specific to Procrustean GMM.

In the DA of shape, the sensitivity of results to p/N and small samples is briefly explored by either reducing the number of group predictors, including only the first 10 PCs, or by leaving out the species with the smallest N, which are the Alaskan and Olympic marmots, with respectively 16 and 14 specimens. In these two samples, the number of individuals is smaller than the shape dimensionality (i.e., $N < p = 20$). Summarizing shape with only the first 10 PCs (90% of total shape variance), all species samples have more specimens than variables, but the DA produces slightly lower hit-rates (84% accuracy, overall, compared to 90% using all PCs). In contrast, if Alaskan and Olympic marmots are excluded and all PCs (recomputed in the reduced sample) are analysed, the resulting cross-validated hit-rates are slightly higher (92% overall accuracy) than including all species and PCs.

Finally, in the last column of Table 6, purely for didactic aims, I show the hit-rates using the first 10 PCs of size-corrected shape data. They are slightly lower (82% overall accuracy) than using the first 10 PCs of total shape (84%). In fact, however, this analysis should not be done, because allometric trajectories are not statistically parallel (see Results and Discussion in B6).

Discussion (B4)

The estimates of cross-validated classification accuracy in the DAs are consistent with the conclusions of the pairwise tests. Despite large size differences, shape is better at discriminating species and results of shape DAs in marmots are not strongly affected by p/N. Thus, I mainly focus the Discussion on the DA of shape using all species and all PCs, which correctly classified on average 90% of individuals. In this analysis, only the Olympic marmot had a somewhat lower hit-rate (71%), but this is the smallest sample (N = 14) and three of the four misclassified individuals were affiliated to either hoary marmots or VAN, which is not unreasonable since all three species are part of the hoary marmot superspecies complex (Kerhoulas *et al.* 2015). In contrast, VAN had the highest hit-rate (96%) of all North American marmots, which is congruent with this species highest average Rsq in pairwise tests. Thus, as in previous studies (Cardini 2003; Nagorsen & Cardini 2009), VAN is confirmed as a highly distinctive species for mandibular shape. As with its almost uniformly dark fur (Armitage 2009) and unique kee-aw alarm call (Blumstein 1999), phenotypic change seems to have been faster in this isolated insular endemism.

From a methodological perspective, it is reassuring that replicating the shape DA/CVA, so that p/N < 1 in all species, had a small, mostly negligible, impact on total hit-rates (ranging from 84% to 92%). In contrast, it is worrying that a DA/CVA using all shape coordinates or their PCs produced different estimates of classification accuracy (Table 6). With PCs the total hit-rate is 90%, but using the Procrustes shape coordinates it drops to 78%. This difference is remarkable and should not be there at all, because the information being used is 100% identical. As explained in the methods, a PCA using the variance covariance matrix of the Procrustes shape coordinates is simply a rigid rotation of the axes that leaves all pairwise multivariate distances in the sample unaltered. This is easily verified by unfolding the two pairwise distance matrices (i.e., stacking data below the main diagonal – or above, if one prefers – so that, for each matrix, they are in the same, single column) to plot them one against the other: the distances should lie on a line with a slope of one and the matrix correlation between Procrustes shape distances and Euclidean distances in the PCA space should also be one (or virtually one, since PCs are computed on data projected in the tangent space – see Introduction in A and explanations in V&C). Thus, using a metaphor, a rigid rotation is just like looking at the tips of needles, stuck in a transparent plastic box, under different angles: the relative positions of the tips (like the specimens in a scatterplot) will look different, depending on the view angle, but they are in fact identical.

It seems likely that the inaccurate DA hit-rate, based on Procrustes shape coordinates in PAST, is a computational inaccuracy that happens because of the strong collinearity (‘redundancy’) in the coordinates. Similar inaccuracies can happen in multivariate parametric tests in PAST and other statistical programs, which are not specific to Procrustean GMM. Depending on how computations are done, statistical programs may not ‘recognize’ that the Procrustes shape coordinates have an additional amount of covariation (beyond any real pattern of covariance among landmarks) as a consequence of the superimposition. Thus, for example, if a DA is done on the Procrustes shape coordinates of the marmot mandibles, the degrees of freedom for the predictors are computed taking into account 24 variables (twice the number of 2D landmarks). But this is incorrect, because four dimensions are lost (i.e., the corresponding information is removed) during the superimposition: one is removed by standardizing CS to one in all individuals; another two degrees of freedom are lost by centroid-centering all individuals along the horizontal and vertical axes; and a fourth degree of freedom is lost by minimizing rotational differences. With 3D landmarks, the reasoning is the same, but the loss of information is seven degrees of freedom, because there is a third axis of translation and two more rotational planes[21]. With slid semilandmarks, if present, there will be, very approximately, another degree of freedom lost for each semilandmark. Likewise, as generally appropriate in taxonomic applications, using only the symmetric component of Procrustes shape in structures with object symmetry, which may be appropriate in taxonomic applications, there will be further redundancy in the Procrustes shape coordinates because asymmetric variation has been removed (Klingenberg *et al.* 2002).

It is easy to check that P values of multivariate parametric tests in PAST change depending on whether they are done on the 24 Procrustes shape coordinates or the corresponding 20 shape PCs with non zero variance. A specific test, for instance a MANOVA or a multivariate regression, is done first using the 24 shape coordinates and then repeated with the 20 PCs of the same shape data: the degrees of freedom reported by PAST will be different, as well as, likely, the value of the test statistic (Wilks' Lambda or its approximated F ratio, for instance) and the corresponding P value. Checking if computational inaccuracies in a DA/CVA are causing inaccuracies in cross-validated hit-rates is less immediate, but still relatively simple. One can create an N by p matrix of random normally distributed numbers with the same mean and variance as in the Procrustes shape coordinates using a spreadsheet (the function usually is '=randnorm(mean, SD)', without the single quotes). Data are then imported in PAST for a PCA, whose scores are saved. It does not matter, for this aim, that there are no group differences. Differences could also be simulated, but now I am only interested in comparing hit-rates between the original variables and their PCs before and after a Procrustes superimposition. Thus, a DA/CVA is done on either the original random coordinates, as they were in the spreadsheet, or their PCs and everything is repeated after superimposing the random coordinates, which introduces covariance and, thus, redundancy as in the real mandible dataset. The DA/CVAs on random coordinates or their PCs produce identical cross-validated hit-rates in PAST. In contrast, after the random coordinates are superimposed, hit-rates are different if the CVA is done on the superimposed coordinates or their PCs.

To avoid mistakes, unless users are very certain that the software can correctly deal with redundant variable in all multivariate computations and also to adjust degrees of freedom in parametric test, I strongly suggest to always analyse PCs with non-zero variance from a PCA of the data projected in the tangent space, as those computed by MorphoJ or TPSRelw. A small practical disadvantage of analysing PCs in PAST is that some of its multivariate analyses have options for visualizing Procrustes shape variation. If one wishes to draw shape diagrams in PAST (e.g., the expansion factors, which are not available in MorphoJ or the TPS Series), he/she must do the statistics (scores, hit-rates, tests etc.) using the PCs of the Procrustes shape coordinates and, then, redo the same analysis (a PCA or DA/CVA, for instance) using the Procrustes shape coordinates ONLY for visualizing shape changes. Nonetheless, even when the Procrustes shape coordinates are employed in PAST exclusively for visualizing shape variation, I suggest to quickly compare the deformation grids with those done in MorphoJ to be sure that they look the same.

As I am discussing some specific aspects of Procrustes shape data in the context of multivariate analyses, I take the chance for stressing another potential misuse of this type of data, as well as a limitation of the user-friendly software used in this study. Although PAST has no option for stepwise DAs, commercial statistical programs and also R might allow to subset the variables, so that group separation is maximized while parsimony is simultaneously achieved by reducing the number of group predictors. For instance, a stepwise DA might suggest that hit-rates are larger when only certain PCs are included (say, as a made-up example, PC1, PC2, PC4, PC7 and PC8). In previous work, we have already warned

[21] If $N \leq p - 4$, for 2D landmarks ($p - 7$ for 3D), however, the non-zero variance PCs will be $N - 1$, because sample size limits the 'real' dimensionality of the data (Zelditch *et al.* 2004; Rohlf 2021). For instance, with 20 marmots and 12 mandibular landmarks, a PCA would result in 19 PCs with non-zero variance, instead of 20. A number of PCs smaller than the original number of variables (minus the degrees of freedom lost in the superimposition, for Procrustes data) should, however, act as a first warning of a potentially problematic p/N ratio.

that stepwise approaches are sensitive to small variation in sample composition and may, therefore, produce results which are less easy to generalize (Kovarovic *et al.* 2011). Also, even when the sample is the same, a stepwise procedure might select different variables depending on the analysis, so that, for instance, different classification methods employ different information to predict the same groups. Thus, if dimensionality reduction is really necessary, it might be better to explain why; then, demonstrate that a subset, including an entire block of the first PCs (e.g., in my analysis, the first 10 PCs) likely preserves the most important information in relation to the main study question; and, finally, consistently use those same PCs in all analyses that cannot be performed in the full Procrustes shape space. Zelditch *et al.* (2004) provide a brief overview on dimensionality reduction using PCs, but there are many alternatives such as, for instance, Horn's method (Glorfeld 1995) or the modified scree-plot using correlations of shape distances (first PCs vs total Procrustes shape) explained in Cardini *et al.* (2010). Regardless of the approach, taxonomists should consider that the subset of PCs selected for the analysis should be adequate to accurately summarize Procrustes shape distances not only in the total sample (e.g., all six marmot species with 445 individuals in total), but also within subsamples or groups, so that estimates of parameters within subsamples (means, variances, covariances etc.) are also accurate.

PAST provides more flexibility for the DA/CVA compared to MorphoJ, but it does not have the larger set of options, and results, available in most commercial programs or R. I have already mentioned that, for instance, PAST users only have the option of equal prior probabilities for any group. The output of PAST also lacks an important part, that is particularly useful in taxonomic and forensic applications. A DA/CVA not only decides the affiliation of each individual to the a priori groups, but also estimates its probability to be a member of one or the other group based on its distance to the means. This probability is called posterior probability (PP), because it is obtained after the discriminant functions have been calculated. PPs across all groups sum up to one. With two species (S1 and S2), for instance, I may find that PP of individual X is 0.03 for species S1 and, thus, 0.97 for species S2, which is clearly the most likely group for X. Another individual, XX, may also be classified in S2 as the most likely group, but its PP could be much lower; for instance, PP_{XX} could be 0.51, which means that, despite ending up in S2, XX has a 49% probability of being S1. Knowing PPs, therefore, allows to accurately assess the confidence with which each individual is affiliated to one or the other group.

Unfortunately, PAST does not have a table with PPs. Besides, PPs only represent a relative probability that refers exclusively to the available groups. It could be that, even if X is indeed much closer to S2 (PP = 0.97) than to S1 (PP = 0.03), within S2 X is an outlier. To make this more intuitive, I go on with the simple example where there are only two species, but same applies for any number of a priori groups. With only two species, there is a single DA/CVA axis, that can be visualized as a line passing through the means of S1 and S2. It might happen that X is on the side of the CV1 line opposite to the mean of S1 (i.e., it is not in the space in between the two means). Thus, X is very far from the mean of S1, but this does not exclude that, despite being relatively closer to S2 than S1, it is also far from the mean of S2. Therefore, X belongs to S2, because it is comparatively much closer to its mean, but it could be, nevertheless, an outlier for this species. The absolute probability of an individual being at a given distance from the mean of the group, where it is classified, is called typicality probability (Albrecht 1992). Typicality probabilities, like PPs, are estimated using a multivariate normal distribution. Thus, X could have a PP of 0.97 of being in S2 compared to S1, despite a typicality probability of, say, < 0.01. Such a low typicality probability suggests that X might be better classified as unlikely to belong to either S1 or S2. Not only PAST but, in fact, most statistical software does not provide estimates of typicality probabilities. Those who need to compute typicality probabilities might want to explore the *typprobClass()* function of Morpho (Schlager 2017) in R, whose use is exemplified in the help of the *CVA()* function. If the outlier detection (A2) has been careful, however, it is unlikely that any specimen will have a very low typicality probability in a DA. This is another reason to be cautious when potential outliers are found in the preliminary screening of the data.

B5) Summary and visualization of species shape differences

Methods (B5)

Morphometricians use ordinations to explore and summarize patterns of shape variation. To this aim, cluster analyses are another possibility. The complementarity, as well as some of the different pros and cons of these methods, are outlined in part A. Ordinations and phenograms can also be used to summarize differences among samples and/or the similarity relationships of group mean shapes. I first introduce ordination methods to explore patterns of variability in shape by plotting individuals and groups. Later, I explain how to obtain group mean shapes, visualize them and summarize their similarity relationships. With Procrustes shape data, there are also some special aspects, related to the biological arbitrariness of the superimposition that need to be considered. Basically, any interpretation of variables or landmarks one at a time (including loadings in ordinations, coefficients in regressions etc.) must be avoided and, thus, patterns of shape change should be described using only shape diagrams (Klingenberg 2013). In the Discussion, I will provide some more detail on this point.

In taxonomic research, groups can be plotted using different colours and symbols in a PCA (as anticipated in Figs 2 and 4a - see also below). For the computation, it is important to use the variance covariance matrix, instead of the correlation matrix, because Procrustes shape coordinates are in the same unit of measure and the covariance matrix faithfully preserves the similarity relationships in the Procrustes shape space. When software, such as MorphoJ or TPSRelw, are used, one can be confident that the default computations are correct, because these programs are specific to Procrustes shape data. In PAST, users have to select the correct options (*Multivar, Principal components*, checking the *Var-covar* and *Disregard Groups* boxes). Also, because PAST is not specific to GMM, higher order PCs with virtually zero variance will be shown (*View scatter, View numbers*) and have to be manually discarded. For instance, for the Procrustes shape coordinates of marmot mandibles, imported in PAST from MorphoJ, the last four PCs have eigenvalues (i.e., variances) in the order of 10^{-17}, which practically means zero, as expected in relation to the loss of degrees of freedom in the superimposition (see B4). If a user is uncertain that PCs are correctly computed in a non-GMM software, he/she can compare scores with those of MorphoJ or check that pairwise Euclidean distances computed using the PCs are virtually identical to the Procrustes shape distances obtained in TPSSmall (see B2 for instructions on how to obtain them).

PAST and MorphoJ have options to add group-specific confidence ellipses based on the PC scores visualized in a scatterplot (e.g., PC1 vs PC2 or PC3 vs PC4). In PAST, as usual, groups are specified using colours (*Edit, Row colour/symbol*). PAST can also show groups using convex hulls which, unlike ellipses, do not require normally distributed data. The ellipses, however, have the advantage that they take into account sampling error. This means that they will be much larger for small samples. In MorphoJ, using a classifier like species, the ellipses can be drawn for the sample (*Equal frequency ellipse*) or the group mean (*Confidence ellipse for the mean*). The former are the same as in PAST and provide an estimate of variability in a sample, like confidence intervals based on the SD of univariate data. The latter are related to the standard error of the mean ($SD/\sqrt{N}$), which is smaller than the SD and useful to interpret statistical tests of group mean differences (see Howell 2013 or Moore & McCabe 2005 for an introduction). In this respect, however, users must bear in mind that the ellipses drawn in the scatterplots are based only on the PCs being shown, whereas group mean differences are tested in the full multivariate shape space which, for marmot mandibles, is made of all 20 PCs.

A PCA maximizes total variance regardless of groups. For this reason, PCA summary scatterplots are suboptimal (or ‘conservative’ (Rohlf 2021)) for exploring patterns of group differences. If groups are well separated in a PCA, a researcher can be fairly confident that there are differences; if they are not,

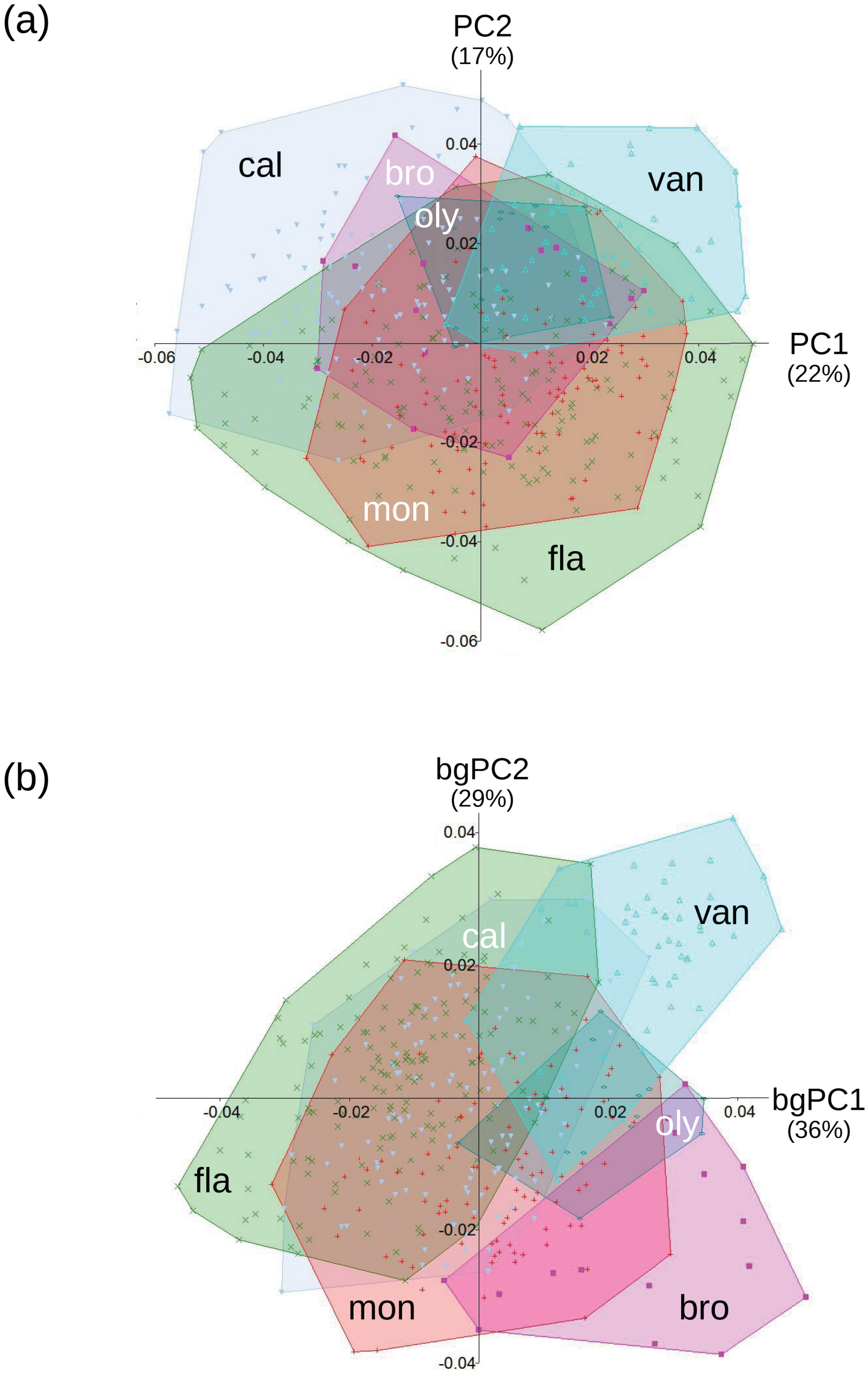

Fig. 4. Ordinations summarizing species variation in shape using the first two axes of (a) a conventional PCA (total variance in parentheses) or (b) those of a bgPCA (between group variance in parentheses).

however, it could be that none of the first PCs aligns well with the direction of group differences, which may still be there but subtle and masked by the pattern of overall variation in the data. In this respect, DA/CVA scatterplots are better to summarize multivariate group structure, but tend to inflate separation (Kovarovic *et al.* 2011; Rohlf 2021). Another possibility is a bgPCA. Using a simplified description, a bgPCA is 'as if' a DA is performed without standardizing the data (Cardini *et al.* 2019). In a bgPCA, individual shape coordinates are projected on the vectors one would obtain by doing a PCA on group means. Thus, differences among groups are maximized, but, unlike a DA/CVA, the scores of a bgPCA remain in the Procrustes shape space. The bgPCs simply carve out a subspace of the total shape space according to group separation, instead of total sample variance as in a conventional PCA.

As in a DA/CVA, the number of bgPCs is g -1, each with an associated proportion of between group variance: bgPC1 explains the most, followed by bgPC2 etc. Also, and again as in DA/CVA, the bgPCA subspace only captures between group variation. However, in a bgPCA, one can in theory compute also PCs of the residual non-between group variance. This option is not available in PAST, but it is present in the R package Morpho using the *groupPCA()* function (Schlager 2017). In Morpho, users can also decide whether or not to take into account group differences in sample size in the computation of between group vectors. However, if samples are fairly balanced, results will be similar regardless of weighting by N (default in Morpho) or not (the only option in PAST). Importantly, unlike a DA/CVA, a bgPCA can be computed even when p > N. Yet, the larger the p/N ratio, the more likely it is that group separation is inflated in the between group subspace (Cardini *et al.* 2019; Rohlf 2021). To avoid this type of spurious result, one could plot cross-validated bgPC scores, which are calculated using the same rationale as in a leave-one-out DA/CVA (Cardini & Polly 2020). A cross-validated bgPCA is, for now, only available in R (Schlager 2017; Thioulouse *et al.* 2021). However, when N is adequate (> or, better, >> p), the difference between cross-validated and non-validated bgPCs is usually minimal in biological data, where there is real covariation among the landmarks and their Procrustes shape coordinates (Cardini & Polly 2020). A bgPCA can also be used to predict group affiliation, but this is another option not yet available in user-friendly programs such as PAST. PAST, however, is one of the few user-friendly programs that computes a bgPCA. The analysis is obtained from the same menu as for the PCA (*Multivar, Principal components* using the *Var-covar* matrix), but requires the user to first specify groups using colours and then check the *Between Group* box.

With the marmot mandible shape data, I exemplify summary scatterplots for groups using the PCA and bgPCA in PAST. When ordinations are done on Procrustes shape coordinates, PAST has also an option (*Shape deform. (2D)*) for 2D shape diagrams using wireframes, thin plate spline grids (Klingenberg 2013) and even 2D expansion factors. The expansion factors, which are not available in MorphoJ or the TPS Series, employ colour-coding based on the thin plate spline deformation grids to emphasize regions where there is a local expansion (shown in yellow-orange-red, going from modest to larger changes) or a local shrinking (green-cyan-blue, from moderate to larger). The shape diagrams in PAST correspond to the changes occurring along a PC (or CV or bgPC). To draw them, once the *deformation from mean shape* window has been opened, the user can select one of the first six axis (*Component*) and writes the score (*Score*) he/she wants to visualize on that axis. Showing shape diagrams corresponding to differences between the total sample mean shape (the origin of the axes) and both of the opposite extremes of an axis is common to aid the interpretation of variability in ordinations of Procrustes shape data. Often, differences are magnified to make them more evident. For example, if the highest positive and negative scores on PC1 were, respectively, 0.05 and -0.035, a two-fold magnification would visualize the shape diagrams of an hypothetical individual with a PC1 score of 0.1 or, for the opposite extreme of the PC, -0.07. If differences are magnified, the magnification factor must be stated in a figure used in a publication or presentation. I stress again that, as already discussed, Procrustes shape coordinates in PAST can be used for ordinations and the visualization, but are better being avoided for parametric tests and group prediction (including in the *Multivar, MANOVA/CVA* of PAST), because there can be issues with computations and degrees of freedom in matrices with redundant information (i.e., very high collinearity, as inevitable after the superimposition).

Methods described in the previous paragraphs allow to summarize, explore and visualize the variability within and among groups. If mean groups differences are found, it is also useful to describe the pattern of similarity of the mean shapes. The methods are basically the same, with the exception of the CVA/DA and bgPCA, which are specific for group separation and not applicable to sample means, unless one wants to group them in supraspecific clusters (e.g., subgenera, genera etc.). Thus, the variability in mean shapes can be typically summarized using a PCA and/or a phenogram. Both methods have already been explained (mainly in part A). However, when applied to group mean shapes, there are some specific aspects to consider. The main one is that the analysis of mean shapes, whose computation in based on the specimens available in a specific sample, typically does not take into account the uncertainties around their estimates. I will briefly discuss this problem and suggest a simple approach to start exploring the issue in user-friendly programs. First, however, I explain how to compute shape diagrams for group mean shapes. I use SDM in hoary marmots, as a simple example that only involves two means, but it works similarly for all pairwise comparisons of group means (including species, as in B3). With more than two groups, to visualize mean differences in the full shape space, a researcher can do the relevant pairwise comparisons or compute the grand mean (the mean of the group means) and compare each species to the grand mean. With the North American marmots, a researcher would, thus, be using the same approach as with SDM (see below) but compare, for example, the mean hoary marmot with the grand mean of all six marmot species; then, he/she could do the same with the mean yellow-bellied marmot etc.

Mean shapes are easy to compute in MorphoJ using *Preliminaries, Average Observations By* and an appropriate classifier. The classifier could be species or, within a species, sex. Mean shape SDM, that I am using as an example of shape diagrams, was not visualized in the results of B1–B2, because it is largely negligible, but it comes handy as a simple case of shape differences with just two shapes, the mean of females and the mean of males. Thus, for interpreting mean differences, one can visualize the shape changes when the female species mean is compared to the male mean of the same species. In MorphoJ, this requires splitting the data by species (which had already been done in B1) and, then, within each species, averaging individuals using the classifier for sex. If there are unsexed individuals, those will be excluded before computing the female and male means (*Preliminaries, Exclude or Include Observations*). Then, a PCA is done on the dataset with only the female and male means of a species, which produces a single PC with just two points. There is only one PC, because using two mean shapes, N = 2 and, thus, regardless of the number of landmarks and shape coordinates, there is only N -1 = 1 PC axis. However, the shape differences between the two group means are in the total shape space of the 24 Procrustes shape coordinates. The two points on PC1, shown with vertical bars in MorphoJ, are the female and male means. To know which is which, if in doubt, the user has to colour-code the groups (right clicking on the plot in the *Graphics, Shape changes* window and selecting *Color the data points* using the sex classifier). The differences in shape on this single PC are the mean shape differences between sexes in that species. As usual, if necessary, differences can be magnified to aid the interpretation. This is particularly useful with mean shapes, as their differences are smaller than those between, say, the opposite extremes of bgPC1 or CV1. For instance, if the PC1 score for the male mean of hoary marmots is ~ 0.01, right clicking on the *Shape changes* window of a PCA in MorphoJ and setting the *scale factor* (the name MorphoJ uses for ordination scores) to 0.05 would correspond to a 10 fold[22] magnification of the difference with the female mean (whose score for the visualization is -0.05, using the same magnification as for the male).

MorphoJ shows shape differences with the target (e.g., the male mean) superimposed on the start shape, which, by default in this software, is the average of a sample and, therefore, in this case, the grand mean of the female and male mean shapes. As we suggested in V&C (Viscosi & Cardini 2011), however, it is generally better to avoid superimposed shapes and displacement vectors, as they easily lead to misleading interpretations such as describing differences in terms of landmark movements (e.g., landmark 1 moving forward, landmark 2 backward etc.). In contrast, by separately showing a start and a target shape one next to the other, a researcher is compelled to visually integrate variation simultaneously across all landmarks. Thus, he/she will more naturally describe it in terms of change in the space in between the landmarks, instead of interpreting shape differences one landmark at a time, which is wrong (Cardini & Verderame 2022). For instance, in the visualization of SDM in hoary marmots, using superimposed diagrams, it might look as if if the tip of the coronoid process moves backward in the male mean (Fig. 5a). However, this is misleading, because Procrustes shape changes must be described regardless of the superimposition, in terms of what happens in the entire space spanned by a landmark configuration. Thus, one might more correctly say that the whole region of the coronoid process expands and becomes longer in males (Fig. 5b) compared to females (Fig. 5c), as suggested by plotting one shape next to the other, instead of on top of it.

To obtain separate shape diagrams in MorphoJ, one must first change the visualization options using *Preliminaries, Set Options for Shape Graphs.* At the bottom of the window opened by this command, the user unchecks the box *Show starting shape*. Then, he/she first shows one extreme of PC1 (the male mean, say), and saves the diagram as SVG (to be edited in a vector graphics software such as the free open-source Inkscape - https://inkscape.org/) or simply as JPG or PNG. Later, the researcher does the same for the opposite extreme of PC1 (the female mean), using the same magnification. Finally, the two separate diagrams are pasted one next to the other in a graphic editor or in a slide for a presentation. Right clicking the *Shape changes* window in MorphoJ, it is also possible to select the type of diagram (displacement vectors, which this software calls 'lollipops'; thin plate spline deformation grids; wireframe or outline). The help file has a good amount of information on the different options, discussed in detail by Klingenberg (2013) and exemplified by V&C, and also explains how to build a wireframe (*Preliminaries, Create or Edit Wireframe*) or draw, format and import an outline TXT file.

Phenograms, when applied to group means, are a particularly effective approach to graphically summarize similarity relationships. This is because the number of observations is much smaller than using individuals in samples, which makes the tree easier to interpret. However, phenograms have

[22] The magnification would be fivefold, if done relative to the mean of the two means. This type of visualization, relative to the mean of the observations, regardless of using individuals or group mean shapes, is the default option in MorphoJ. MorphoJ also adopts a default 0.1 score for the visualization of shape change along any PC. This default option must be borne in mind, because it means that almost all the time users are not seeing a specific score in their observed data space. They visualize the shape that a specimen might have if there was an individual on PC1 = 0.1 (or PC2 = 0.1, PC3 = 0.1 etc.). But 0.1 may be larger or smaller than the observed highest PC score on that axis. Often, it tends to be larger, unless variation is big in a study, and that means that the shape visualized by default in MorphoJ is frequently outside the observed range of PC scores. By setting a new '*Scale Factor*' on a specific PC to the value of the individual with the largest score on that PC, users will be able to visualize the observed shape with no magnification. For instance, with the means of females and males of hoary marmots, PC1 ranges between -0.01 and 0.01, which is 10 times less than visualized using the default 0.1 (or -0.1, for the negative extreme). That the difference is magnified is not an issue, as long as the user is aware of this and states the magnification in his/her publication. The term '*Scale Factor*' is, however, an unfortunate choice for the observed score in a MorphoJ plot, because it is already used in TPSDig and the TPS format to specify the value to convert pixels in units such as mm or cm, and also because 'scaling' in GMM is generally employed in relation to size. For instance, we typically say that specimens are 'scaled' to unit CS in the Procrustes superimposition. Users will, therefore, have to be careful and remember the different meaning of 'scale' in MorphoJ compared to the TPS Series and its common use in GMM.

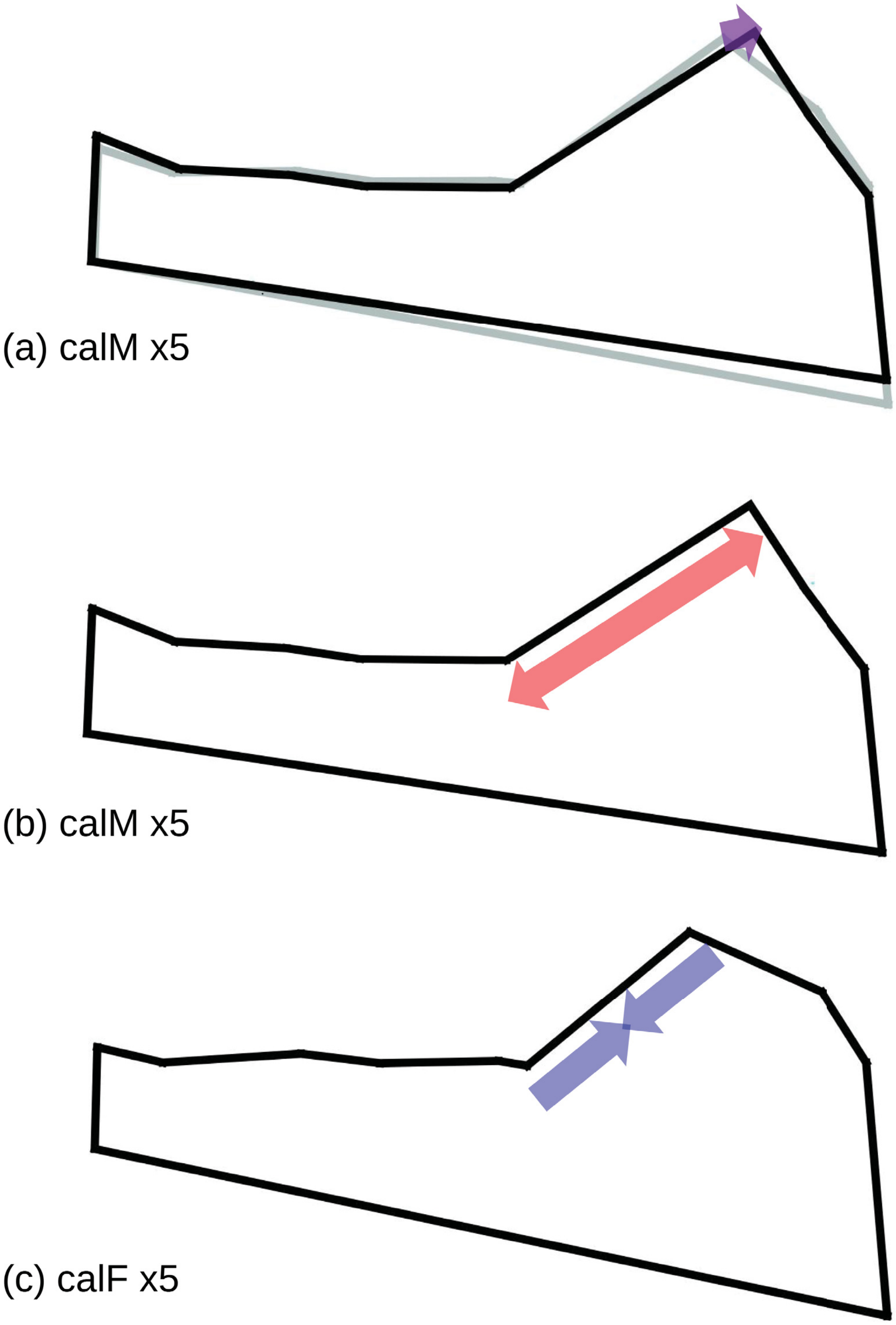

Fig. 5. Example of visualization of shape change: hoary marmot SDM illustrated using (a) superimposed shapes (male mean, in black, and grand mean of female and male means, in grey) or separate diagrams for male (b) and female (c) mean shapes. Focusing on the coronoid region, the violet arrow shows the potentially misleading effect of the superimposition, suggesting a backward 'movement' of the tip of the coronoid in males. Separate diagrams (b–c), in contrast, correctly suggest that change happens in the region whose boundary are marked by the landmarks, with the rostral margin of the coronoid becoming longer (red arrow) in males and shorter (blue arrows) in females.

several limitations: a) they are only about phenetic relationships, and generally cannot be interpreted as phylogenetic hypotheses (Felsenstein 2004); b) all shape information is used to build a phenogram, but the resulting tree distorts the relationships captured by the original Procrustes shape distances (de Queiroz & Good 1997); c) uncertainties in the reconstruction of the tree topology are, usually, not taken into account (Pearson *et al.* 2015). In a study on differences in relation to taxonomy, a) is not an issue, but b) and c) are, as better explained in the B5 Discussion. In fact, the same limitations concern ordinations of mean shapes, that are purely phenetic in nature and do not take into account sampling error.

To mitigate against c) and, to a smaller degree, b), I suggest a subsampling experiment that helps exploring uncertainties in the estimates of group means. It is not equivalent to a confidence interval and has its own limitations. Besides, it may not be doable if a sample is small, and results are potentially biased when N is highly heterogeneous across species. The main advantage of this approach is, however, its simplicity. The idea is to use random subsamples of the species with the largest samples to re-estimate their means in smaller samples. The size of the subsamples can be the number of individuals found in the species with the smallest sample. The variability among random subsamples means should, therefore, be a crude proxy for the amount of error expected in samples as small as in the species with the smallest N. To summarize this variation in relation to interspecific differences, the means of the random subsamples can be analysed with a PCA and/or a cluster analysis.

The size of the smallest species sample in my dataset is 14, for the Olympic marmot. In the Alaskan marmot, which has almost the same N (16), I just randomly excluded two individuals. The difference is minimal and one could have included all 16 individuals. In all other species, in contrast, N was much larger and, thus, I randomized the order of the specimens, before selecting mutually exclusive subsamples of 14 individuals. The randomization of the specimens order within each species can be done in TPSUtil (see 'Digital images and landmark configuration' in the Material and methods of part A) with the data later reloaded in MorphoJ together with a classifier for each of the N = 14 subsamples. Using the classifier, it will be easy to compute in MorphoJ the means of the randomized subsamples (*Preliminaries, Average Observations By ...*), before, finally, combine all species subsample means in a single dataset (*Preliminaries, Combine datasets ...*). When the total sample size was not a multiple integer of 14, however, a few specimens had to excluded. Overall, therefore, in relation to the avalaible number of mutually exclusive random subsamples in each species, I analysed one mean shape for both Olympic and Alaskan marmots, three for VAN, seven for hoary marmots[23] and woodchucks, and eleven for yellow-bellied marmots, for a total of 30 mean shapes. With these 30 mean shapes, I run a PCA and a UPGMA cluster analysis using tangent space Euclidean shape distances, which are virtually identical to the corresponding Procrustes shape distances (see part A and Marcus *et al.* 2000). Finally, I checked if multiple estimates of mean shapes of the same species clustered together 'within' that species, as expected if N = 14 is precise enough to estimate species mean shapes when the aim is to summarize average similarity relationships in marmots. Alternatively, if N = 14 is too small for precision, I would expect mean shapes of different species to be mixed in the scatterplot and the phenogram, with no clear prevalence of species-specific clusters.

Results (B5)

Because all figures summarizing the patterns of mandibular shape differences in North American marmots use the abbreviations of the species names, I remind readers that they can use Table 1 to check scientific and common names, as well as the abbreviations. Thus, it is easy to observe that, despite significantly large differences in mandibular shape and high cross-validated hit-rates in the DAs (Tables 4–6), ordinations of individuals including all six species (Fig. 4) show a large overlap in

[23] In fact, as I realized while correcting the proofs of the accepted version of this paper, by mistake, one of the seven mutually exclusive subsamples of hoary marmots had N = 12. However, its mean shape is virtually identical to that obtained by adding another two random specimens and, thus, the results of the analysis are unchanged.

shape variation among species. The PCA suggests a fairly circular scatter within each species (Fig. 4a). The area of the convex hulls seems to indicate that variance is somewhat proportional to sample size, with yellow-bellied marmots varying the most on PC1–2 and Olympic marmots the least. The bgPCA scatterplot (Fig. 4b) captures most of between group variance (65% using bgPC1–2) and shows a pattern of variation which is mostly congruent with the PCA. There are, however, some differences. Compared to PC1–2, along bgPC1–2, the within-species scatter is more elliptical and the apparent differences in variance among the three largest samples are less pronounced. In fact, on bgPC1–2, woodchucks, instead of yellow-bellied marmots, seem to vary the most. Also, in the bgPCA scatterplot, VAN is better-separated from other North American marmots than in the PCA.

Shape similarity relationships of North American marmots are summarized also using ordinations and phenograms of means (Figs 6–7). For these summaries, I am using the mean shapes of the random, mutually exclusive, balanced subsamples. This is optional, as generally morphometricians use the total sample mean shapes. However, as outlined in the methods, using the means of the subsamples can be a first step to explore the uncertainties in relation to sampling error. A researcher might replicate the summaries using the total samples mean shapes, but this is redundant when, as in my case, the between species separation is so evident using the balanced subsample means.

Figure 6 is a PC1–PC2 scatterplot, which accounts for a total of 62% of variance in the balanced subsamples mean shapes. There is a complete interspecific separation in the space of PC1–2: each species with multiple mean shapes forms tight within-species clusters, totally separated from those of other species. However, the means of the two smallest species samples, the Alaskan and Olympic marmot, represented each by a single mean, are only slightly separated from the cluster of VAN means at the negative extreme of PC2.

The shape change at the opposite extremes of bgPC1 and bgPC2 is shown using deformation grids and expansion factors, magnified five times relative to the gran mean, which is in the origin of the axes and has zero scores (Fig. 6). The woodchuck and the hoary marmot, lying on opposite sides of PC1, are characterized by differences in the extension of the mandibular angle (longer in hoary marmots), masticatory tooth-row (wider in woodchucks) and incisor alveolus (showing a horizontal expansion between the mental foramen and the incisor in hoary marmots and a contraction in woodchucks). The other species have intermediate PC1 scores, with partially overlapping ranges, but are well separated on PC2. The yellow-bellied marmots, at the positive extreme of PC2, are completely separated from the group formed by VAN, Alaskan and Olympic marmots at the opposite, negative, extreme of the same PC. PC2 scores in yellow-bellied marmots are associated with a relatively slender mandible, with a sharp contraction of the dorsal margin of the mandibular symphysis and a pronounced expansion of the vertical ramus between the coronoid and condylar processes. Negative PC2 scores, typical of VAN, Alaskan and Olympic marmots, suggest a recurved coronoid process, forward bent apex of the condyle, dorsoventrally compressed vertical ramus and an elongated dorsal margin of the symphysis.

If PC3 (18% of variance) and PC4 (6% of variance) were also explored (not shown), the Alaskan marmot would be found isolated at the positive extreme of PC4. The other five species, in contrast, mostly overlap on PC4, but, with the exception of the Olympic marmot (partly separated from hoary marmots on PC4, but overlapping on PC3), they are very well separated on PC3: the woodchuck is at the negative extreme and VAN at the positive one; hoary and yellow-bellied marmots are in the middle, with the former closer to woodchucks and the latter to VAN.

The UPGMA phenogram of the balanced subsamples mean shapes (Fig. 7) is complimentary to the PCA of Fig. 6. All species represented by multiple mean shapes using balanced random subsamples form within-species clusters isolated from those of other species. In terms of relative positions in the tree, VAN is basal, which indicates its distinctive mandibular shape, consistently with the on average larger Rsq of this species in the pairwise tests of mean shape differences (Table 5) and a good degree of separation from other marmots in Fig. 4b. The Olympic and Alaskan marmots are 'phenetic sisters' in the phenogram, but they are separated by long branches, indicative of large differences. They are apparently slightly closer to the hoary marmot than to the yellow-bellied marmot or woodchuck. However, the

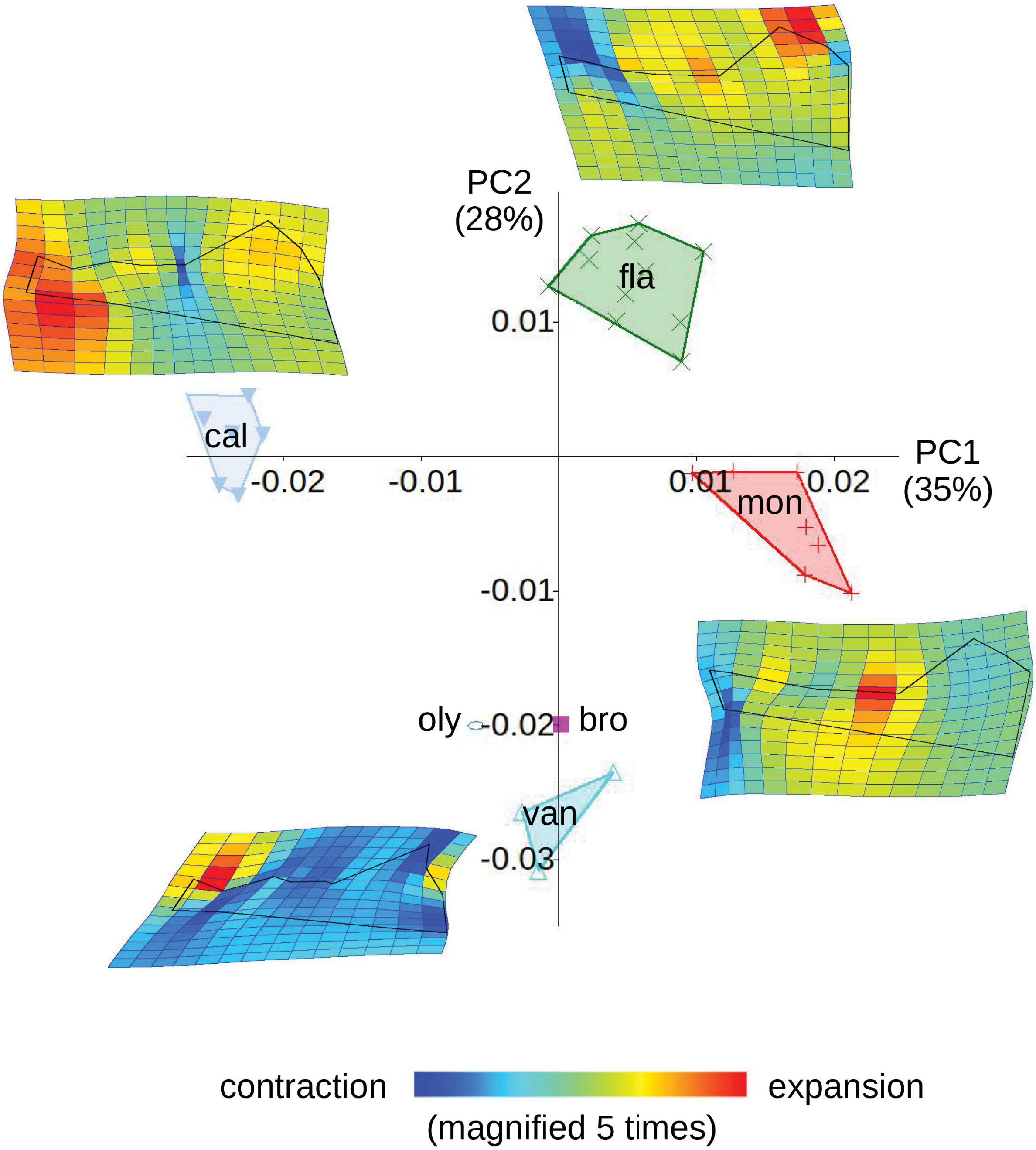

Fig. 6. PC1–PC2 of mean shapes for the random, mutually exclusive, species subsamples. Shape variation (magnified five times) at the opposite extremes of each PC is shown using wireframes, as well as deformation grids and expansion factors computed in PAST using the thin plate spline interpolation. (In these wireframes, unlike those in MorphoJ, the mental foramen is also connected by a line to its neighbouring landmarks, as PAST constrains users to link all landmarks: the difference is, however, minimal and purely visual).

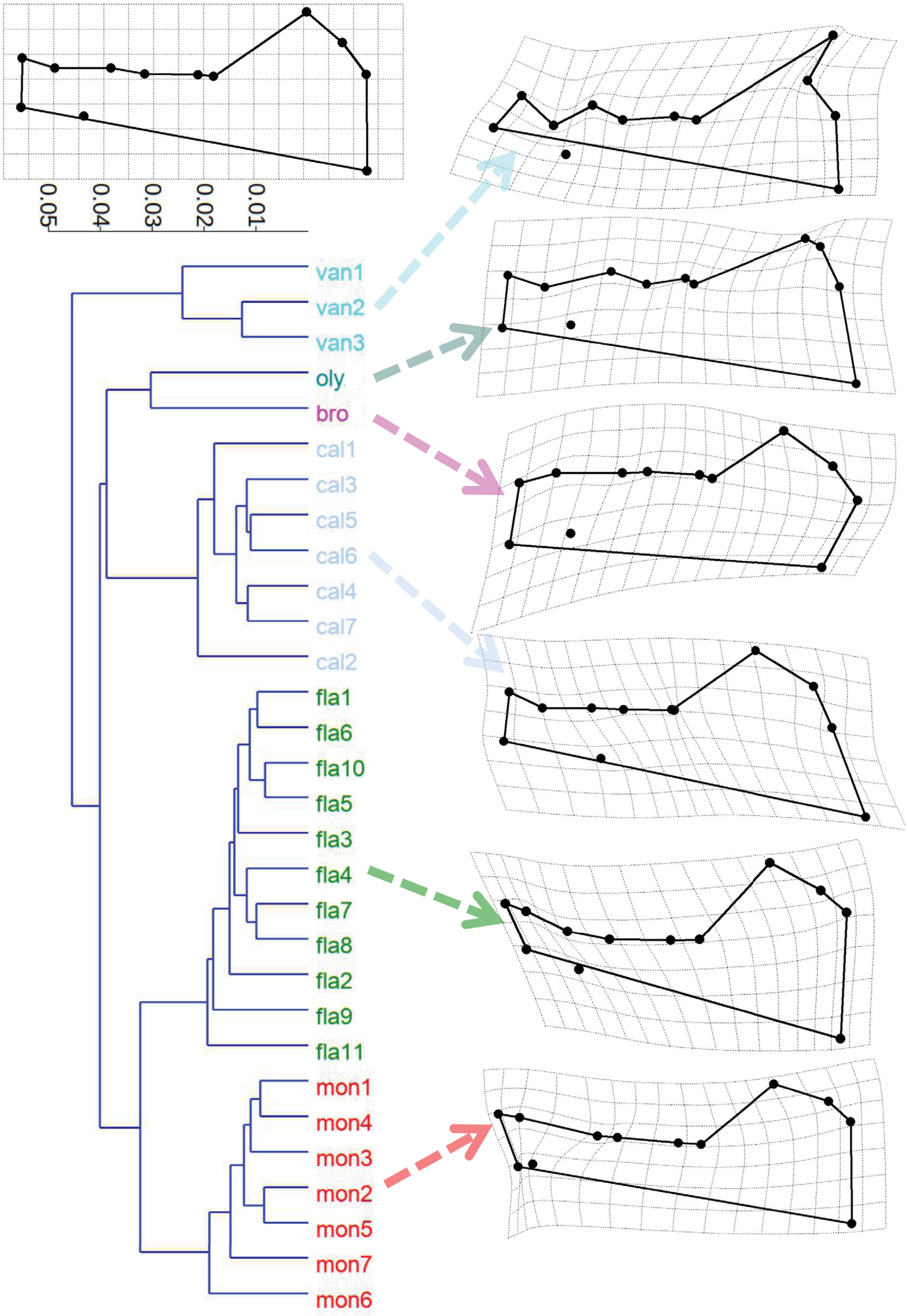

Fig. 7. UPGMA phenogram of Procrustes mean shape distances for the random, mutually exclusive, species subsamples. Shape variation (magnified five times, relative to the grand mean of all species) is illustrated using the six species mean shapes (all specimens included) with wireframes and thin-plate spline deformation grids (drawn in Morpheus et al. - Slice 1999) - but equivalent to those made using MorphoJ or the TPS Series).

branch length of the cluster formed by hoary, Olympic and Alaskan marmots is very short, so that it is probably better to interpret this node as a 'phenetic polytomy'. The yellow-bellied marmots and woodchucks are 'sister' clusters, but the corresponding branch length is also fairly short. In contrast, branches separating the yellow-bellied marmot cluster from the woodchuck cluster are almost twice longer than that joining these two species. In short, the phenogram strongly supports the within-species similarities of mean shapes estimated using small samples, as well as the distinctive mandibular shape of VAN; the clusters between different species, however, are more dubious.

Figure 7 also illustrates, next to the tree with the means of the species random subsamples, the differences between the grand mean shape (mean of the means of all North American marmots) and the mean shape of each species, estimated, for the visualization, using the complete species samples. As in Fig. 6, VAN has a relatively longer superior margin of the symphysis, with a narrowing of the incisor alveolus approximately between its caudal end and the mental foramen. The vertical ramus is dorsoventrally compressed, with relatively short angular and condylar processes and an elongated coronoid. Both the Alaskan and Olympic marmots have robust incisor alveoli, with grids suggesting a dorsoventral expansion in the horizontal compared to the vertical ramus. In the vertical ramus, there are clear differences, with the Alaskan marmot having relatively short coronoid and angular processes, which make the condyle look more prominent, whereas the Olympic marmot has a longer coronoid and more pronounced angle of the mandible. The depth of the vertical ramus, with an enlarged, prominent angle, and the expansion of the incisor alveolus are the most evident aspects of shape change in the hoary marmot. Finally, both the yellow-bellied marmot and woodchuck have relatively slender mandibular shapes, but this relative horizontal expansion and dorsoventral compression occurs almost uniformly in the woodchuck, with the exception of a sharp contraction between the mental foramen and the inferior margin of the incisor alveolus. In contrast, in the yellow-bellied marmot, only the anterior part of the mandible is comparatively thin and elongated, whereas the vertical ramus is comparatively deep and with a relatively long condyle.

Discussion (B5)

It is almost a truism that summarizing differences in multivariate shape data is not as simple as using univariate plots for CS. Moreover, the interpretation of multivariate results of analyses of Procrustes shape variables cannot be based on coefficients, such as PCA or DA/CVA loadings, as in traditional morphometrics. Using coefficients to argue whether one or the other shape variable is more important for the computation of a certain PC or DA/CVA axis is akin to doing inaccurate and misleading per-landmark analyses or visualizations. The mistake of per-landmark interpretation is, as explained in part A, a consequence of having specimens superimposed using a convenient, but biologically arbitrary mathematical superimposition (Moyers & Bookstein 1979; Richtsmeier *et al.* 2002; Cardini & Verderame 2022). In contrast, all interpretations, and analyses, must be done in the total multivariate shape space. This means carefully avoiding parts of the output such as PCA or DA/CVA loadings in PAST or other multi-purpose statistical programs. Thus, shape diagrams (Klingenberg 2013) are the correct approach to describe the changes captured by a multivariate analysis of Procrustes shape data. For instance, in an ordination, one might show the shape changes happening along a specific PC (Fig. 6); for tests of group means, the average shape of each group is a simple and effective summary (Figs 5, 7); in a multivariate regression, one can visualize the shapes corresponding to the regression predictions for the smallest and largest values of the independent variable (Fig. 8 in B6).

In a taxonomic comparison, ordinations of individuals, with groups marked using different colour or symbols, are typically used to summarize shape variation before visualizing and interpreting mean shape changes. The three main ordination methods in GMM are a simple PCA, a CVA and a bgPCA. The advantages and disadvantages of these methods are summarized in Rohlf (2021). As I mentioned in the Methods, in taxonomic studies using shape, I generally opt for the bgPCA. A bgPCA better separates groups compared to a simple PCA but, using morphometric datasets with covarying anatomical landmarks and a reasonable p/N ratio (Cardini & Polly 2020), is less prone than a CVA to inflating group differences. With marmot mandibles, the bgPCA seems to perform well, as it captures most between group differences in just two axes (bgPC1–2, Fig. 4b). It also suggests a degree of separation of VAN, whereas other species mostly overlap. That VAN is more separated in the scatterplot is a result in good agreement with the pairwise tests of species mean shape differences (Table 5), in which VAN has on average the largest Rsq (almost 20% vs ~ 10–15% in other marmots). Yet, it is apparently puzzling that all pairwise tests of mandibular shape are highly significant when there is so much overlap in the ordinations (Fig. 4). In fact, there is no disagreement, if P values are discussed in relation to the corresponding Rsq. Then, it becomes clear that on average interspecific differences are not negligible, but, even if tests are significant and Rsq moderately large, ~ 80% or more of total shape variance remains unaccounted for by species mean shape differences. Similarly, when all species were simultaneously tested in the MANOVA, the Rsq of interspecific variation was about 20% (Table 4). A 20% multivariate Rsq is not small, but, with some 80% of variation unrelated to species mean differences, a large overlap among species in the ordinations is unsurprising. Besides, when shape variance is summarized including all species, axes in ordinations represent a compromise that tries to captures all main group differences at the same time. In contrast, when just three species are included, as in Fig. 2, bgPCA axes capture all interspecific differences and more effectively align with the main direction of between species variation. Figure 2 likely better separates the three species with the largest samples also because the pattern of shape variation is less strongly dominated by VAN, whose clear distinctiveness on bgPC1–2 contributes to push other species in a smaller region of the scatterplot, where they largely overlap.

In terms of species-specific shape variation, results are broadly congruent with those of previous studies on marmot mandibles, despite using fewer landmarks and often smaller samples (Cardini 2003; Nagorsen & Cardini 2009). The posteriorly curved, long coronoid, for instance, is confirmed as a highly distinctive feature of VAN that is uncommon in other species and almost diagnostic. The relatively long mandibular angle in hoary and Olympic marmots, the largest species, and the slender horizontal ramus of the woodchuck and yellow-bellied marmot are reminiscent of similar shape features observed within species in, respectively, older and larger individuals compared to younger and smaller ones (Cardini & Tongiorgi 2003; Cardini & O'Higgins 2005). Thus, they might be part of a common allometric trend that occurs during ontogeny and is, partly, observed also in relation to interspecific size divergence in adults. A longer angular process is also likely to help providing a larger surface for the insertion of the main masticatory muscles (mainly, the masseter) and, possibly, a more effective lever arm in bigger animals.

The PCA scatterplot of Fig. 6 and the phenogram of Fig. 7, which summarize the balanced random subsample means and support within-species similarities and between species average differences, suggest that most of the variation that causes overlap in the interspecific ordinations of individuals (Fig. 4) are the small differences that make each individual unique. Some of these may be due to within-species genetic variation and others represent plastic responses to the variability of environmental conditions. More likely, they are a mix of the two. When individuals are averaged, small inconsistent differences are 'smoothed out'. In these samples, just 14 individuals seem to be enough for a fairly accurate estimate of a species mean shape, at least in the four species with samples large enough for random mutually exclusive subsamples to be drawn (i.e., all except the Alaskan and Olympic marmot). With means based on 14 individuals, species form tight clusters well separated from those of other species. N = 14 might be approximately appropriate also for the Olympic marmot, with its small population living in a restricted geographical area, but probably not for the Alaskan marmot, that occupies the much wider region of the Brooks Range, stretching over more than 1000 km and with peaks reaching almost 3000 m a.s.l.

The preliminary clues on an adequate sample size for estimating species mean shapes in marmot mandibles do not differ much from the lower boundary for a minimum N suggested in some previous

GMM studies on mammals. For instance, Cardini *et al.* (2015) found that 10–20 individuals sharply reduced the scatter of mean premolar shapes in Icelandic ponies. However, that was an intraspecific study and only one group had a sample large enough for random subsampling experiments. In a broader study on sampling error in mammal craniofacial shape, Cardini *et al.* (2021) showed that ~ 15 individuals may be enough for producing precise estimates of means in relation to interspecific mean differences, but demonstrated that no less than 20 (and often many more) specimens are necessary for repeatable reconstructions of interspecific relationships based on mean shapes. Also, for precision in estimates of within-species shape variance (which I did not assess in this study), Cardini *et al*. (2021), as well as other researchers (as summarized in their Discussion), suggested a minimum requirement of 20–40 individuals. The majority of studies of sampling error in morphometric studies of within and between species variation, among closely related species (see references in Cardini *et al.* 2021), seem to agree that, below N = 10, estimates are highly inaccurate and precision drops. Therefore, when such small samples cannot be improved, the corresponding groups might have to be excluded or results flagged and re-verified, as I did, after leaving out the smallest samples.

The phenogram of random subsample means (Fig. 7) hints at a modest degree of mandibular shape similarity between yellow-bellied marmots and woodchucks, which for North America are, respectively, the smallest and second smallest species. Likely, the source of this moderate similarity is convergence driven by size, since the two species are only distantly related (Steppan *et al.* 1999, 2011). Mean shapes, however, are sample estimates and, as explained in the methods, neither a cluster analysis nor a PCA of mean shapes take into account the uncertainty around these estimates. Resampling methods are a promising approach to assess confidence around mean shapes. PAST, for instance, has an option to bootstrap variables and infer how strongly the data support different branches of a phenogram. However, this method is inappropriate for Procrustes shape data and does not address the issue of the effect of sampling error on the topology of a tree using mean shapes (Cardini & Elton 2008). An alternative type of bootstrap should be used to this aim (Caumul & Polly 2005; Cardini & Elton 2008; Nagorsen & Cardini 2009; Pearson *et al.* 2015), but this is not available in any user-friendly software, although it is fairly easy to program in R. In a PCA on individuals (as the one of Fig. 4), the option for drawing confidence ellipses around means, available in MorphoJ, is an interesting one, but the PCA plot is not the same as with group means and uncertainties are estimated only in relation to the pair of PCs used in the scatterplot.

In the methods, I anticipated that not only summaries using mean shapes but also and, in particular, phenograms have limitations: a) they are not phylogenetic hypotheses (Felsenstein 2004); b) they distort phenetic relationships (de Queiroz & Good 1997); c) uncertainties in the tree topology are not taken into account (see above). As I have already discussed the third problem in the previous paragraph, I add here a few comments on the other two.

In general, Procrustes shape data are not very suitable for phylogenetic inference (Varón-González *et al.* 2020) and are probably best used as a source of preliminary evidence on taxonomic distinctiveness (Cardini *et al.* 2022) or interpreted (see below) in a ‘post-cladistic’ context (Smith 1990). However, that a phenogram is not a cladogram, except under very restrictive assumptions (Felsenstein 2004), is not a problem when the aim is that of summarizing phenetic relationships. The tree simply provides a different type of information that is ideally interpreted in relation to a well-supported molecular phylogeny (post-cladistic approach). With VAN, for example, I showed that, as in previous work (Cardini 2003; Nagorsen & Cardini 2009), its mean shape lies on a relatively isolated basal branch of the mean shape phenogram. Genetic data, however, show that VAN is nested within the radiation of the hoary marmot (Kerhoulas *et al.* 2015; Mills *et al.* 2023). The discrepancy supports the hypothesis of accelerated morphological evolution of this population on the island, as common in insular mammals (Millien 2006). This type of post-cladistic interpretation can be facilitated by graphical comparisons.

An effective way to complement phenetic information from shape data and phylogenetic hypotheses is to plot phenograms and cladograms one next to the other (e.g., Cardini 2003: fig. 6) or to project a cladogram onto a PCA scatterplot (see *Map Onto Phylogeny* in the help of MorphoJ). Alternatively, one could also map shape changes onto a phylogenetic tree, which for 2D data is doable in TPSTree (Rohlf 2015).

Distortions in phenograms are more problematic (de Queiroz & Good 1997), but can be assessed using cophenetic correlations (automatically shown in PAST and mentioned in part A). Integrating the information of a phenogram with that of a PCA scatterplot might also help to detect large distortions. Phenograms usually preserve shorter distances, i.e. those among the 'leaves' in a terminal branch, better than larger ones, which are more accurately captured in a scatterplot of the first PCs. For cluster analyses, there is a variety of algorithms. However, in taxonomy and, in general, in biology, UPGMA trees are most common option and usually perform relatively well (Rohlf 1970). In conclusion, therefore, as long as one is clear about the aim of a phenogram, and the uncertainties and limitations in the reconstruction of a tree as a summary of phenetic variation, cluster analyses do not pose any particular problem and are another helpful technique in the analyses of highly dimensional multivariate data, as those typical of Procrustean GMM.

B6) Relationship between shape and size within and across species

Methods (B6)

Allometry is rarely the main subject of a taxonomic investigation. However, as I discuss later, taxonomists might be interested to assess the effect of size on shape differences and, if there is an effect, they might want to investigate whether allometries are similar in different taxa and account for a large proportion of taxonomic differences. The conventional approach to answer these questions uses a multivariate regression of shape on size to test the significance and magnitude of allometry (Zelditch *et al.* 2004). In this analytical framework, to take group differences into account, a MANCOVA is used to test the similarity in allometric trajectories (i.e., the linear multivariate regression lines specific of each taxon). Potentially, the MANCOVA also allows to 'size-correct' shapes, before re-testing 'allometry-free' species differences (Zelditch *et al.* 2004). There are, however, alternative approaches, as discussed by Klingenberg (2016, 2022), which are not considered here, and a modified version of the conventional MANCOVA is presented in Elton *et al.* (Elton *et al.* 2010).

Before comparing allometric trajectories among species, it may be interesting to first explore, one species at a time, whether there is any appreciable allometric variation. This step is optional, because overall allometry (i.e., simultaneously for all species) is tested in the MANCOVA. Alternatively, one could first do the MANCOVA and later use within-species allometric analyses to discover how much allometric variation, if any, is present in each of the study groups, which is the equivalent of pairwise post-hoc comparisons in an ANOVA.

For testing within-species static allometry, I used linear multivariate regressions of the Procrustes shape coordinates on CS in MorphoJ. The test statistic, as in all regressions in this software, is the Rsq, whose significance is assessed using permutations (10 000 in my case). The type of analysis is the same used to test SDM or pairwise species differences. The only difference is that, now, the regression is testing the association between shape and a continuous predictor, instead of a binary grouping dummy variable, as in the tests for mean differences of B1 and B3. The Rsq in the regression represents the amount of shape variation accounted for by the covariation with the CS of the individuals in a sample. The analysis is specified in MorphoJ as already explained in the subsections, B1 and B3, on group differences. The only change is that the predictor (a *covariate* in MorphoJ's jargon) is CS. The same test can be done in TPSRegr (for 2D data only, using both permutations and parametric tests) and PAST (*Model,* *linear ... multivariate*). In PAST, the multivariate regression is only parametric and, therefore, assumes normality. Also, in PAST, the multivariate regression should be done using the PCs of the Procrustes shape coordinates to avoid miscalculations of the degrees of freedom (see Discussion in this chapter). If none of the species shows significant allometric variation (non-significant P with a small Rsq), the MANCOVA is unnecessary. In contrast, if allometry is significant in some species, the MANCOVA might be performed.

The species by CS MANCOVA is well explained in Zelditch *et al.* (2004), but also in the help file of TPSRegr, where it is called "test for common slopes". The rationale is very similar to the two-way MANOVA. The difference is that, instead of two grouping variables (such as species and sex), there is now one grouping factor (species) and a continuous covariate (CS). In a species by CS MANCOVA (using type III SS, as in the two-way MANOVA), first one assesses the interaction term. The interaction tests if the allometric trajectories of the different groups have the same slope. The species-specific regression lines in this MANCOVA, with the interaction included, have slopes and intercepts identical to those obtained in the regressions done one species at a time (previous step). If slopes are statistically similar, and thus allometries are approximately parallel, it means that changes in mandibular proportions, correlated to size, are similar in all species. For instance, in marmots, larger individuals generally tend to develop deeper mandibles to increase robustness (Cardini & Tongiorgi 2003; Cardini & O'Higgins 2005).

As anticipated in the Introduction, the degree of similarity or divergence in allometries is, in itself, an interesting information for a taxonomist, because one expects smaller differences in allometric patterns when there is less evolutionary divergence. A non-significant interaction might, therefore, suggests closer phylogenetic relationships. As in other tests, however, non-significance can happen even if allometric trajectories form large angles, because of low power and/or inaccuracies in small samples (Cardini & Elton 2007). When samples are large, because multivariate tests tend to be powerful, the opposite may also occur, with allometries similar in direction (i.e., approximately parallel), despite a significant interaction suggesting differences in slopes (Klingenberg 2016). As usual, the P value for the interaction term should be shown together with, and interpreted in relation to, its Rsq. I will go back on the issue of Rsq and the angles between trajectories when I discuss the statistical models to compare allometric vectors.

If the effect of the interaction is negligible, one can 'size-correct' shapes before testing again species differences. That the interaction is negligible is a fundamental assumption for going on with the size-correction; in contrast, if the interaction is significant and has a large Rsq, one cannot size-correct shapes. Size-correcting shapes means that the static allometric variation in all species is 'statistically removed'. This is done by repeating the MANCOVA after excluding the interaction term. In this second MANCOVA, only species and CS are tested. The MANCOVA compares groups (species, in my case) using the residuals of the allometric trajectories, which are the component of shape variation unrelated to size differences. The comparison, however, is meaningful only as long as the allometric trajectories are parallel, and this is why the species by CS interaction must be tested first. With a size-correction, it is as if shape variation (unrelated to CS) is 'squeezed' in each group around the species-specific shape predicted for a given size. Crucially, the size chosen for the predictions is identical in all species (e.g., the average CS of all species). Thus, the predicted means at that specific size become the new species mean shapes to which the regression residuals (i.e., non-allometric variation) are added. The overall procedure, therefore, 'removes' the effect of size on shape and produces size-corrected shape data. Because allometries are parallel, the choice of the 'common' size used to control allometric variation is irrelevant: whether it is the smallest, largest, the total sample mean or the grand mean size of all species (or even an abstraction such as CS = 0), results are the same because, with parallel allometries, the relative distances among the predictions used to compute the 'new' mean shapes are constant across the entire range of CS values. In contrast, if trajectories were not parallel and, as an example, diverged as size increases, results of comparisons of size-corrected shapes would change depending on the size chosen to control for allometry: in this case, size-corrected shape differences between species are smaller, and maybe negligible, at the lower extreme of size variation (where allometric lines are closer), but bigger at the larger extreme, where allometric lines diverge more.

Assuming one has demonstrated similar allometric patterns in the species by CS MANCOVA, the second 'species plus CS' MANCOVA without interaction allows to focus on size-corrected species shape differences. In practice, in the first MANCOVA, a researcher usually only looks at the interaction term (species by CS); in the second, he/she focuses on the species factor. This is analogous to what we saw in the species by sex MANOVA, where first one focuses on the interaction and, later, if that is negligible, looks at SDM. Thus, if species was significant without size-correction (tests in B2–B3), but it is not in the 'species plus CS' MANCOVA, the conclusion is that species shape differences are purely allometric. Of course, this conclusion is peculiar to the evidence provided by the morphological structure and the specific configuration of landmarks chosen for the taxonomic assessment. With a different structure, or possibly even with a fairly different landmark configuration, results may change. In contrast, if controlling for allometry in the MANCOVA does not remove species differences (i.e., species is significant in the 'species plus CS' MANCOVA), that indicates that allometry alone does not account for all interspecific variation. Assuming size is more plastic and evolutionary labile, shape differences, when not purely related to size, might hint at a likely deeper evolutionary separation.

In terms of user-friendly software, the MANCOVA has the same limitations as the two-way MANOVA. Neither PAST nor MorphoJ can do this type of analysis, although MorphoJ offers some alternatives, that can produce equivalent results, as I explain in the Discussion. A taxon (species, in my case) by size (CS) MANCOVA is available in commercial statistical software, as well as in R, with different options for the SS. The distinctions in terms of model SS are the same I mentioned for the ANOVA/MANOVA. With balanced samples, the type of the SS does not matter. However, at least in my experience, with 'reasonable' data with large and fairly homogeneous N, results in moderately unbalanced samples tend to be similar using type I, II or III SS.

For 2D shape data, the species by CS MANCOVA can be run in TPSRegr using dummy variables. I provide some guidelines on how to do it without a detailed explanation. The MANCOVA in TPSRegr is clearly exemplified in its help file where it is called "*Example of test for common slopes*". As in the MANOVA, the file format is NTS. At least for the dependent variables (the shape data), one could in theory reuse the file employed for the species by sex MANOVA. However, I need new NTS files for both dependent and independent variables in the MANCOVA because, with a negligible SDM (B1–B2), the specimens of unknown sex can now be included. Thus, I have one NTS file with the landmark coordinates (averaged between the two digitizations) of 445 individuals, and a second NTS file with the design matrix for the independent variables for the same individuals. Specimens must be in the same order in both files. As in the MANOVA, it is convenient to have first all individuals of one species, followed by those of the second species etc.

The design matrix is made of three blocks of variables, two for the main effects (species and CS) and one for their interaction, for a total of 12 variables in my dataset. CS is simply a single column with the CS of all individuals. The second block, which is species dummy variables, can be built as shown in the help and seen in the two-way MANOVA. Thus, there will be five variables, each coding a species as 1 (or -1) and all others as zero except one species (always the same!), which is coded -1 (or 1) in all five dummy variables. This time I coded the Alaskan marmot -1 in all five species dummy variables; in these same variables, other species were coded 1 or zero, depending on the dummy variable (e.g., hoary marmot = 1, others = 0; VAN = 1, others = 0 etc.). However, this choice is arbitrary.

I could have, as in the MANOVA, coded -1 VAN in all variables and 1 or zero the other species. Finally, there is the third block, which consists of six dummy variables for the interaction. These variables have each the CS of one species and zeros for all the others.

In this paragraph, I describe in detail how to test the interaction, whereas in the next I explain how to test species holding the effect of size constant (i.e., using size-corrected shapes). The test of species size-corrected mean shape differences is conditional on the negative outcome of the test of the interaction. Thus, with the variables I made as described in the paragraph above, one runs a first regression of shape onto the species and interaction blocks of the independent variables (*Options, Select indep. Variables*), which means that only the CS column of the design matrix is excluded. This is the full model, where each species has its own group-specific slope in the allometric regressions. The slopes are identical to those obtained one species at a time (first part of B6). However, because they are all regressed in the same analysis simultaneously, the Rsq corresponds to shape variance accounted for by static allometries in the total sample (N = 445). For instance, in my dataset, the MANCOVA fits six separate species-specific regression lines. As usual, in the *View report* window, TPSregr reports the percentage of variance unexplained, which is subtracted from 100% to obtain the full model Rsq. Now, one needs a second regression which least square fits lines (one for each species) so that they are constrained to be parallel. This is the reduced model where all species are forced to have the same slope. The reduced model is specified in TPSRegr, after checking the *Retain current resid. SS* option, by selecting the CS column and the species block as independent variables (thus, excluding the six variables in the interaction block). The difference between the Rsq of the full and reduced model is the Rsq of the species by CS interaction, which is the improvement in the goodness of fit of the regression when slopes are separate compared to when they are forced to be parallel (and, thus, with regression lines which are suboptimal to a smaller or larger degree). The interaction Rsq should be very small if slopes are really similar and, thus, almost parallel.

In the report window, the multivariate tests after the label *** *Testing difference between current residual SS matrix and the residual SS matrix retained from previous analysis.* *** are those testing the species by CS interaction. As in the two-way MANOVA, I report the test using the Wilks' lambda, but one can choose another test statistics with the caveat I have already made about Roy's root being probably too liberal. By comparing the full and reduced model, TPSRegr estimates if the Rsq of the former is significantly larger than the Rsq of the latter: if it is, slopes (first regression) are not homogeneous, because the deviation from a model with parallel trajectories (second regression) is not negligible. When that happens, the analysis stops there. As explained before, one cannot size-correct the data with divergent allometric trajectories. If, however, lines with separate slopes do not improve appreciably the Rsq compared to parallel ones, the analysis produces a non-significant species by CS interaction that allows the researcher to go on testing species differences using size-corrected shapes (next paragraph).

If the interaction is negligible, a researcher runs a second pair of regressions. For this, I suggest to shut down TPSRegr, reload the data and, then, re-run the same regression with parallel lines I explained above (i.e., shape onto the CS column plus the species block). Residuals are retained and another regression is run, this time including only CS as predictor. In this second pair of regressions, one is testing if allometries are just parallel or, in fact, they produces overlapping lines. The analysis is also called test for the homogeneity of intercepts because, if regression lines overlap, their intercepts (i.e., the values corresponding to CS = 0) must be almost identical. The regression with parallel lines has, now, become the full model. The reduced model, instead, is the one where a single regression line fits all 445 individuals regardless of species. TPSRegr is, thus, testing the species factor in a 'species plus CS' MANCOVA (interaction excluded). As usual, a researcher should pay attention not only at the P value of the multivariate test, but also at the magnitude of the difference in Rsq between the full and reduced model and report it together with the P value. If the Rsq of the test for homogeneity of intercepts is very small, after having shown in the first MANCOVA that the allometric model is statistically similar in all species, one has demonstrated that species differences become totally negligible once the effect of allometry is statistically removed from the data. To put it the other way round, if species is not significant and its Rsq small in the ‘species plus CS’ MANCOVA, mean shape species differences, if present, are purely allometric in nature.

The visualization of allometric change is typically done using scatterplots and shape diagrams for the predictions at opposite extremes of an allometric line (i.e., for the smallest and largest individual). The shape diagrams are the usual ones (Klingenberg 2013) that are provided as part of the output of multivariate regressions in both MorphoJ and, for 2D data only, TPSRegr.

I focus mainly on one species at a time regressions in MorphoJ, which is both 2D and 3D and also offers a summary scatterplot reminiscent of (but not identical to!) a bivariate scatterplot in a univariate regression. MorphoJ, in fact, can perform regressions with multiple groups, in a way that takes group structure into account. I will briefly mention this option in the Discussion, where I suggest a variant of the MANCOVA to test species differences using size-corrected data. For the visualization of ‘multi-group’ allometric regressions, however, I avoid MorphoJ’s scatterplots, because allometries in this software are always forced to be parallel (as in the ‘species plus CS’ MANCOVA). This is because MorphoJ assumes that users have already demonstrated, in a different statistical software, that group-specific allometries are statistically parallel (the non-significant species by CS interaction of the MANCOVA). Thus, to summarize variation in allometric predictions among groups I opt for an alternative method, which I explain below, after first describing MorphoJ’s graphical output for within-species allometries.

Once the regression of the Procrustes shape coordinates on CS has been done in MorphoJ, one finds a scatterplot of shape vs CS in the *Graphics, Scores* window. Shape (on the vertical axis) is summarized using regression scores. Regression scores are the projection of the observed shape coordinates onto the vector of the regression coefficients. Thus, they correspond to the shape information that has the highest covariation with CS[24]. Regression scores help to produce a scatterplot that, as said, looks like the conventional visualization of a univariate regression, where the dependent variable is plotted on the vertical axis and the independent predictor on the horizontal axis. However, the two types of scatterplots are not equivalent. Unlike in a univariate regression scatterplot, the regression scores do not show all the residual unexplained variance. They only display the component of total shape variance that covaries the most with the predictor, but this component could account for a very small proportion of total shape variance. Therefore, regression scores are useful, but require caution, as they can mislead users in their perception of how strong an allometric relationship is. To estimate the strength of a relationship, one has to check the Rsq in the *Results* window. In contrast, regression scores help to spot influential observations or outliers and, to some extent, also potential deviations from linearity or homoscedasticity.

Having summarized the linear relationship between shape and its predictor (CS, in this case), one has to interpret the specific shape changes along the regression line. In MorphoJ, this part of the output is found in the *Graphics, Shape changes* window. In this window, MorphoJ compares the sample mean shape to the shapes predicted by the regression, which in my case is the allometric trajectory. One can,

[24] This is analogous to using DA or bgPCA scores to summarize the shape information that ‘covaries’ with, in those cases, group differences. With both, regression scores or DA/bgPCA scores, one should resist the temptation to recycle these statistical summaries for other analyses. For instance, testing group differences using regression scores rarely makes sense. If the regression was a test for allometry, one might argue that comparing groups using regression scores is appropriate to find allometric differences. It is not, because a univariate score cannot capture the full pattern of allometries in the multivariate shape space and the correct approach is to test allometric differences directly in that space (e.g., using the MANCOVA model or testing angles between group-specific allometric trajectories). For similar reasons, it is incorrect to test allometric variation using bgPCA or DA scores which maximize group differences but were not optimized to capture multivariate allometry and, therefore, can only produce misrepresentations of allometries.

as usual, select wireframes, lollipops etc. Rick-clicking on the plot and selecting *Set scale*, users can specify the value of CS for which the predicted allometric shape is visualized. However, this option is poorly explained in the help file. Apparently, CS (or any other predictor) is specified relative to the mean CS, so that positive and negative 'scale' values correspond respectively to individuals larger or smaller than average. For instance, the average CS in VAN is 83 mm and the range of CS goes from 77 to 91 mm. Thus, if a user specifies *Set scale* = -6, the allometric prediction for the smallest specimen in the sample is shown; if he/she specifies *Set scale* = 8, the allometric prediction corresponds to the largest individual in the sample. With a two-fold magnification for the extremes of the allometric trajectory, *Set scale* becomes -12 and 16; with a threefold magnitication, *Set scale* is -18 and 24 etc.

The graphical options I have considered until now mostly concern results of multivariate allometric regressions one species at a time. In a taxonomic study, where multiple groups are present, it is also interesting to summarize interspecific differences in allometries. A simple option consists in saving the species-specific allometric shapes, putting them together in the same dataset and performing a PCA on these shape data (Adams & Nistri 2010). The result, for marmots, is a summary scatterplot with six series of points, each on a straight line. These 'lines' are the species-specific allometries, with PC1–PC2 representing most of the interspecific variation in allometric trajectories; PC3–PC4 the second highest variation in allometries etc. In MorphoJ, the plot can be obtained in two different, but equivalent, ways. The first is precisely what I suggested above. One performs the allometric regressions one species at a time. Then, he/she selects the branch of the project tree with the output of the regression, right-click on it and saves the *Regression prediction*. This is repeated for each species. Finally, the six set of data are combined in a single TXT file (in doing this, carefully leave the column names only in the first dataset, below which data for other species are pasted excluding their own first row with the column names). This TXT file is loaded as a new dataset in the MorphoJ project, species classifiers are re-imported and the data are superimposed and subjected to a PCA. The PCA will provide the summary of the allometric trajectories. Re-superimposing the data does not typically alter the predictions, because they were already superimposed, even if one species at a time. Because to draw a straight line (which corresponds to the regression prediction for a species) one needs two points, it is as if each species is represented by two individuals (the opposite end points of the allometric trajectory). Thus, with six species, the allometric prediction dataset produces 11 PCs with nonzero variance, as it would happen in a PCA with N = 6 * 2 = 12.

The second way to produce the summary PCA scatterplots for the allometric predictions is faster and uses all data after a single common superimposition. Using a TXT file with the same ID (identifier) employed for importing the raw coordinates in MorphoJ, a user can import as covariates the species and interaction blocks of the design matrix used in the MANCOVA. If data in the TPSRegr MANCOVA are in the same order as the raw coordinates in MorphoJ, one only has to add, in a spreadsheet, the ID column before pasting the 11 (for the marmot data) dummy variables (all those used for the MANCOVA except the one with CS for all individuals). The data are then saved in a TXT file, imported in MorphoJ and used to regress shape (all 445 individuals of all marmot species) onto the set of dummy variables. This means performing, with a single command, a multivariate regression of Procrustes shape coordinates on CS with species-specific independent slopes. As explained before, the slopes and intercepts are identical to those obtained in the regressions one species at a time. Finally, a user selects the regression output and performs (with the usual MorphoJ's commands – see footnote in A2) a PCA on the *Regression prediction.* The outcome is, for marmots, an 11 dimensional allometric shape space, as in the 'one species at a time approach', and, therefore, the same type of graphical summary of allometries using predictions' PCs. Results are, generally, virtually identical, but the second approach has the advantage of employing a single common superimposition; it is also faster, as one generates few intermediate files.

Results (B6)

Static allometries (Table 7) are significant in all species except the two with the smallest samples (the Alaskan and Olympic marmots). The strength of the relationship, however, is modest, with an average Rsq of 8%. Only in the yellow-bellied marmot, the species with the largest range of variation in CS (Fig. 1b), static allometry is stronger, so that the Rsq goes up to 16%.

The test for common slope (interaction between species and CS in the MANCOVA) is significant regardless of including or not small samples (Table 8). The full model with species-specific slopes fits the data only 2% (32% vs 30% including all species, and 31% vs 29% excluding smallest samples) better than forcing species allometries to be parallel. This might suggest that, despite significance, the divergence of allometric lines is negligible and shape data may be size-corrected (i.e., doing a second MANCOVA testing species, without the species by CS interaction, or doing pairwise tests of shape differences using MorphoJ's size-correction, as explained in the Discussion). However, the summary of allometric trajectories in Fig. 8 strongly suggests divergence. The trajectory of the woodchuck looks almost orthogonal to those of the species of *Petromarmota*. In the woodchuck, allometric shape makes smaller individuals (Fig. 8) somewhat resemble adults of yellow-bellied marmots (Fig. 7), whereas the larger individuals look fairly similar to the species mean shape (Fig. 7). In *Petromarmota*, in contrast, the allometric change (exemplified in Fig. 8 using the yellow-bellied marmot) is reminiscent of the differences found in studies of ontogenetic allometry, with marmots having a slender rostral portion of the horizontal ramus, if small, and an expanded angular process, if large (Cardini & Tongiorgi 2003; Cardini & O'Higgins 2005).

The large divergence of the woodchuck allometric trajectory is confirmed by repeating the test for common slopes after excluding this species: the difference in Rsq between full and reduced models becomes even smaller (1.5%–1.2% respectively including or excluding the two smallest species samples) and that is enough to make the interaction no longer significant using a 0.005 threshold. The Alaskan marmot belongs to the same subgenus (*Marmota*) as the woodchuck. This species also has a somewhat divergent trajectory in the PC1–PC2 space of allometric predictions, with a direction apparently intermediate between woodchucks and Olympic marmots. However, the Alaskan marmot's trajectory is very short and in this species, as well as in the Olympic marmot, estimates are less reliable, because their samples are small (Cardini & Elton 2007).

Table 7. Within-species 10 000 permutation tests for multivariate allometric regressions testing the null hypothesis that shape is independent from CS (performed in MorphoJ).

Species	P	Rsq
bro	0.6361	5.1%
cal*	*<0.0001*	7.7%
fla*	*<0.0001*	15.9%
mon*	*<0.0001*	4.9%
oly	0.5840	6.9%
van*	*0.0017*	6.4%

* If angles of allometric trajectories are compared pairwise to test if they are significantly smaller than expected by chance, because of sampling error, all pairwise tests have P < 0.005, with angles between 35° and 51°, except tests with *M. monax*, which produce angles between allometric regression vectors ranging from 62° and 72°.

Overall, the effect of static allometry is modest, with the exception of yellow-bellied marmots, and the main differences in allometric patterns seems consistent with the subgeneric separation of North American marmots. However, at least within *Petromarmota*, where samples are large, allometries are broadly collinear.

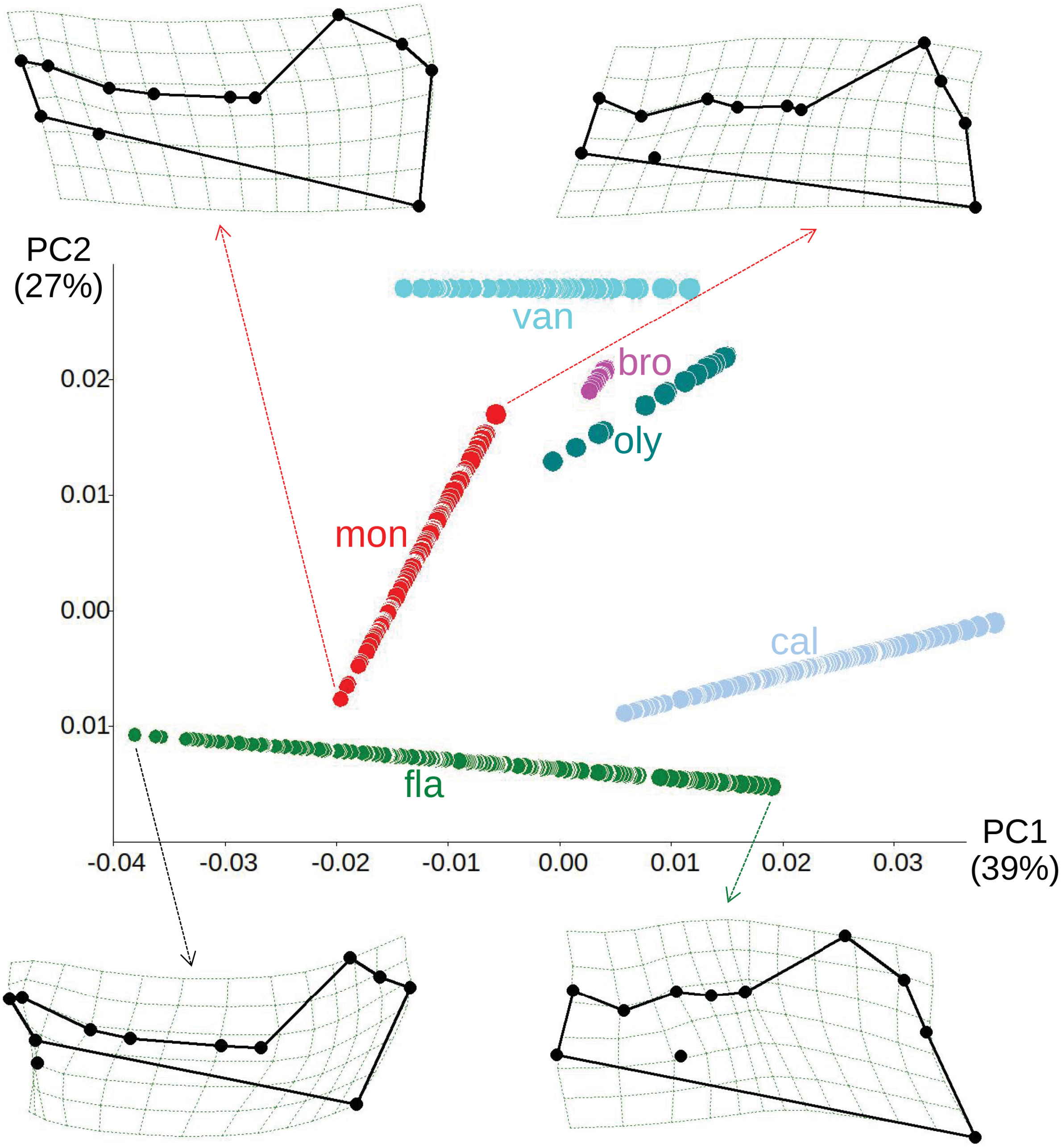

Fig. 8. PC1–PC2 summary scatterplot of species-specific allometric predictions (66% of total allometric shape). As an example, the opposite extremes of the allometric trajectories of yellow-bellied marmots (fla) and woodchucks (mon) are shown (magnified five times) using wireframes and thin plate spline deformation grids, drawn in TPSRelw by 'warping' variation along the regression lines in the PC1–PC2 subspace.

Table 8. Interspecific static allometry: test for common slope and test for homogeneity of intercept, performed using dummy variables in TPSRegr.

Model & samples	Null hypothesis	Wilks' Lambda	Fs	Df1	Df2	P	Rsq of model
all species	common slope	0.613	2.1	100	2024.4	*<0.000001*	32.0%
	homogeneous intercept†	0.033	20.6	100	2048.8	*<0.000001*	29.8%
large samples only	common slope	0.651	3.0	60	1158.4	*<0.000001*	30.8%
(i.e., cal, fla, mon, van*)	homogeneous intercept§	0.059	30.8	60	1167.4	*<0.000001*	28.8%

* VAN is included because, unlike the species by sex MANOVA, where its sample is small using individuals of known sex, pooling sexes and including individuals of unknown sex make its sample size large.

† The significant interaction is mainly due to *M. monax*: if this species is removed, the interaction has a P = 0.04492 with all species (P = 0.01481, using only species with large samples); also, the difference in Rsq between the full (including the species by CS interaction) and reduced (no interaction) models becomes even smaller (from ca. 2% tp ca. 1.5-1.2%). These observations are congruent with the tests using angles, which are significant and large mainly when comparisons involve *M. monax*.

Discussion (B6)

I focus this part of the discussion on the species with the largest samples because static allometric trajectories tend to be short in adult mammals and, thus, require large samples for accurate estimates of a small effect size (Cardini & Elton 2007). This means that there are big uncertainties for Alaskan and Olympic marmots whose allometric analyses have to be confirmed using larger samples.

Within-species, static allometric variation in marmot mandibles is significant, when samples are large and power adequate. However, allometry is generally modest in magnitude. Yellow-bellied marmots are unusual, among North American species, because adults show an amount of allometric change which accounts for almost three times more variance than in other marmots (16% vs an average of 6% in other species). This strong effect of allometry is likely related to the larger range of size differences within this species (Fig. 1b). Compared to other North American species, yellow-bellied marmots seem to have, across their geographic range, more variability in habitat conditions (from arid to mesic) and colony elevation (from hills less than 1000 m high to mountain prairies up to 3500 m) (Armitage 2005, 2013). Among other factors, because the genus *Marmota* conforms to the predictions of the Bergmann's rule (Armitage 2005), marmot body mass is expected to vary in relation to temperature and, therefore, size differences in populations of *M. flaviventris* may be partly related to altitude and latitude. Yellow-bellied marmots also a have larger genetic variance in mitochondrial DNA, compared to other *Petromarmota* species (Rankin *et al.* 2019). Probably, unlike hoary marmots and VAN (Nagorsen & Cardini 2009; Polly *et al.* 2015; Kerhoulas *et al.* 2015; Rankin *et al.* 2019), this species did not undergo strong genetic bottlenecks in its recent history. Yet, even if the yellow-bellied marmot really has more genetic variation, which may contribute to variability in size, we do not know to what extent within-species population differences in body mass of marmots are genetic or plastic in nature. It seems likely that both types of responses to environmental effects are present, but studies are needed to support this assumption and assess the relative magnitude and potential interaction of genetic and environmental effects. Regardless of the explanation, the evidence for a larger variability in size of yellow-bellied marmot mandibles looks robust and, as size changes, a degree of allometric adjustment in proportions is expected in relation to possible developmental constraints and/or to preserve function (Emerson & Bramble 1993; Voje *et al.* 2014).

Unlike in yellow-bellied marmots, the magnitude of static allometry in woodchucks is small (Rsq = 5%), which, as mentioned, seems typical of most marmot species. However, the woodchuck allometric trajectory diverges from those of the members of *Petromarmota*. The distinctiveness of the woodchuck static allometry is evident in the summary plot of allometric predictions (Fig. 8), but it is also supported by comparisons of pairwise angles between regression vectors. After excluding the smallest species samples, angles are on average 66°, when woodchucks are compared to *Petromarmota* species, but just 44° within *Petromarmota*. This difference represents an increase in allometric divergence of ~ 50%. Thus, it is not surprising that woodchucks, a large sample with a distinctive allometric pattern, are the main responsible for the significant species by CS interaction in the MANCOVA. Indeed, if woodchucks are excluded from this analysis, the interaction is no longer significant.

We know, however, that significance must be interpreted together with the estimates of Rsq. In this respect, the reduction in the variance accounted for by the MANCOVA seems negligible when the full model, with separate species-specific regression slopes, is compared to the reduced 'species plus CS' MANCOVA with parallel regression lines: including woodchucks, Rsq is 32–31% with separate slopes and 30-29% with parallel regressions (Table 8), which represents a reduction of just 2%; without woodchucks, the reduction becomes slightly smaller (~1.5%), but this is enough to make it statistically negligible. One might therefore argue that the decrease in Rsq is always so small that, regardless of statistical significance, parallel regressions fit the data almost as well as regressions with separate slopes and, therefore, allometries are close to parallel and shape data can be size-corrected. But is this a sound conclusion?

The evidence from the data is not always unambiguous. With the North American marmot mandibles, results of the allometric analyses are somewhat contradictory. The difference in Rsq using independent or parallel regressions is very small. However, not only the species by CS interaction is highly significant both including all species and using only the largest samples. We have also the graphical summary of species-specific allometric predictions that strongly suggests the divergence of the allometric trajectory of the woodchuck. Besides, there are large and likely uncertainties in the estimates of allometry for the small samples of Alaskan and Olympic marmots (Cardini & Elton 2007) and the statistical power for detecting their potential divergence is also very low. Overall, considering the modest effect size of static allometries in marmot mandibles (Table 7), it seems more cautious to avoid a potentially inaccurate analysis of size-corrected data. If done, as in the last column of Table 6 using the DA/CVA cross-validated hit-rates for didactic aims, the species factor would have remained highly significant (not shown) and the results of interspecific comparisons almost unchanged. That these results are similar, however, could be accidental, as I argue in the next paragraphs.

Alaskan marmots, with the woodchucks, are the only North American members of the subgenus *Marmota*. This species, interestingly, consistently forms large angles in all pairwise comparisons of allometric trajectories. The angles range from 78° (Alaskan marmots vs woodchucks) to 86–87° (Alaskan marmots compared to hoary and Olympic marmots), with an average of 83°. Because the Alaskan marmot sample is very small, however, results cannot be trusted and likely represent an overestimate of allometric divergence. Yet, I take advantage of this species, with its apparently very distinctive static allometry, to exemplify what might happen if data are size-corrected in spite of divergence.

Figure 9a plots PLS1 of a partial least square analysis (Rohlf & Corti 2000) of allometric predictions using separate slopes *versus* CS (used as a covariate in the PLS). PLS1 is used as a summary to maximize the covariation of the allometric predictions with CS, but the plot would be almost identical using PC1 of the predictions (i.e., the PC shown on the horizontal axis of Fig. 8). A size-correction must be independent of the choice of the specific 'common' size used to calculate the size-corrected species shapes, which become the new mean shapes to which the regression residuals (non-allometric shape) are added back, as explained in the methods (B6). However, it is evident from the plot that using, for instance, the mean size of all 445 individuals (CS = 77 mm, emphasized with a vertical light grey line in Fig. 9a), size-corrected mean shapes will be much closer (i.e., more similar) than using the CS of the smallest individual (CS = 56 mm, emphasized with a vertical yellow line in Fig. 9a).

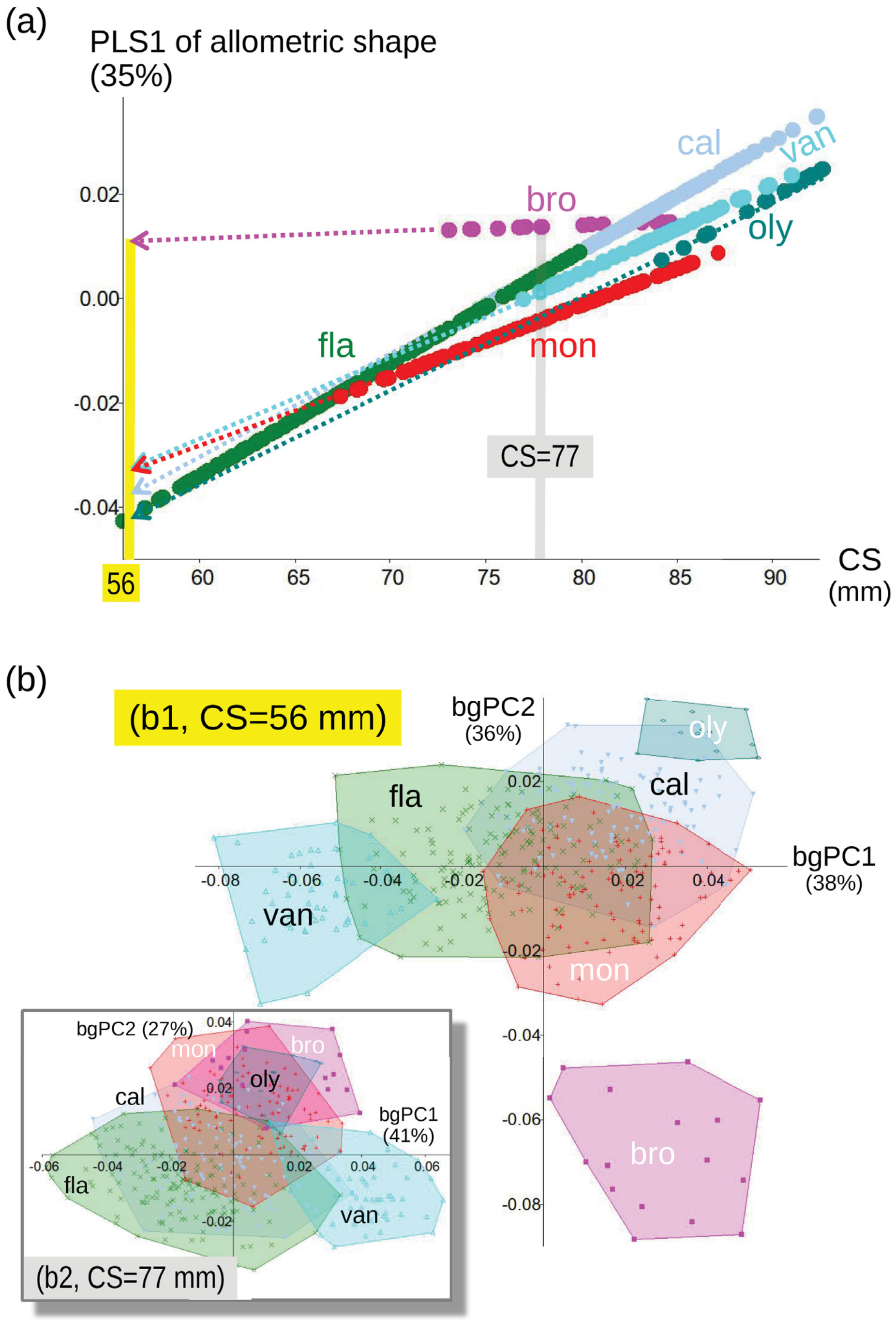

Fig. 9. Divergent allometries and their effect on size-corrected shape. (a) PLS1 summarizing allometries (35% of variance in allometric predictions) vs CS. The vertical lines emphasize the scores of species-specific predicted allometric shapes for either the smallest mandible of all North American marmots (CS = 56 mm, emphasized with a vertical yellow line and arrows to show the extrapolations of the allometric trajectories to CS = 56 mm) or the mean CS of all species (CS = 77 mm, emphasized with a light grey vertical line). (b1) Scatterplot of bgPC1–2 (percentages of between group shape variance in parentheses) for the size-corrected shapes predicted using species-specific allometries (i.e., separate slopes) and CS = 56 mm as 'common' size. (b2, inset) Scatterplot of bgPC1–2 of size-corrected shapes using independent trajectories (as in b1) and CS = 77 mm: if differences in slopes were negligible, b1 and b2 should be almost identical.

The analytical impact of the strong interspecific divergence in the allometry of the Alaskan marmots becomes even clearer if size corrected-shapes are indeed computed using the two different CS and, then, analysed. Thus, in a bgPCA of size-corrected shapes using separate slopes and CS = 56 mm, the Alaskan marmot is totally isolated from all other species on bgPC2 (Fig. 9b1). In contrast, using CS = 77 mm for the size-correction (Fig. 9b2), species show much less separation and the Alaskan marmot largely overlaps with the woodchuck and Olympic marmot. This second bgPC1–2 scatterplot of size-corrected shapes (using CS = 77 mm and independent slopes – Fig. 9b2) is, in fact, very similar to what one would obtain using parallel lines (not shown), as one would do with truly negligible species differences in the direction of allometric trajectories. All the scatterplots (using separate or common slopes and regardless of ‘common’ CS) should, however, be identical if the assumptions for a size-corrections were met. This is because, if the interspecific divergence in allometries was small, the difference in height of the species-specific regression line would be approximately constant for all values of CS (from the smallest to the largest). As this is not the case, the bgPCA scatterplots of size-corrected shapes become strongly influenced by the specific choice of ‘common’ size. But there is no biological justification to opt for one or the other CS in the size-correction and, therefore, the size-correction is inappropriate for the mandibular shape of the North American marmots.

That the divergence of static allometry is generally larger between than within the two subgenera of marmots, however, is interesting and supports the potential usefulness of the comparison of allometric patterns to assess taxonomic differences. For now, with only two species in the subgenus *Marmota*, of which one has a small sample, it is premature to make any strong claim. Nonetheless, it might be interesting in the future to further explore allometric differences by also including large samples of Eurasian species of the subgenus *Marmota*. Previous studies that measured angles of allometric vectors in marmots are scant and in those studies, as in the present one, relatively few species were included. For instance, Cardini & O’Higgins (2005) compared mandibular and cranial ontogenetic allometries among marmots, using a total of six species, two belonging to *Petromarmota* and four to *Marmota.* They found that results were largely congruent between the two structures, but the separation of the subgenera in terms of allometric patterns was incomplete. Their samples, however, not only were relatively modest in overall size, but also biased in age composition, as specimens were mostly adults, whereas younger age classes were poorly represented. In spite of this, also in Cardini & O’Higgins (2005), the mandibles of the woodchuck showed a distinctive allometric trajectory with angles on average larger than between any other pair of marmot species. One might, thus, speculate that the long evolutionary separation of the woodchuck (Steppan *et al.* 1999, 2011; Mills *et al.* 2023), as well as its behavioural (being the only solitary marmot) and ecological (living at low altitude at the boundary between forest and meadows) peculiarities (Armitage 2000), might have contributed to the change in how mandibular shape covaries with size in this species.

Besides the specific findings, the allometric analysis is useful to exemplify the caveats of a size-correction and the main steps of this procedure. Since a correct application of the method requires caution and some effort, however, one might wonder whether size-correcting shape data is really worth being considered in a taxonomic context.

The potential usefulness of exploring size-corrected data might be better appreciated using a couple of hypothetical examples. First of all, in general, with large differences in the size of a structure across individuals and populations or species, a large proportion of taxonomic variation in shape could be due to allometry (Emerson & Bramble 1993; Klingenberg 1996, 2016; Voje *et al.* 2014). But this does not answer the question on whether it is really interesting to check that shape variation is not exclusively allometric in nature when taxa are compared. Let us say, however, that, using mandibles, I am studying two poorly known parapatric marmot populations, p1 and p2. p2 has smaller size. Between p1 and p2, there are also significant differences in shape. After controlling for allometry, however, shape differences become tiny and no longer significant. Later, it is discovered that p2 lives in a region with less favourable conditions, so that the limitation in food quantity and quality not only reduces individual fitness, but also limits growth. The smaller size might, thus, mainly be a plastic response, with shape differences related to an earlier truncation of the same ontogenetic model of allometric change. One does not normally expect such a simple explanation for population differences, but this simplistic, made-up, scenario shows why a taxonomist may be interested in allometry. Using a less abstract case, dwarf populations of large mammals are a common occurrence on islands (Foster 1964; Damuth 1993; Lomolino *et al.* 2013). Especially in young insular populations, if shape changes are found, a taxonomist may want to know if they are independent of, rather than merely driven by, size reduction.

The allometric variation I have investigated in North American marmots is both static (within species) and evolutionary (interspecific similarities and differences in static allometries). To estimate size, I employed CS. The use of CS in allometric studies using Procrustean GMM is almost a default option. The Procrustes superimposition standardizes CS differences to calculate shape. The shape variation, which is left, is, by definition, independent of CS unless there is allometry (Klingenberg 1996). Thus, because CS is based on the same landmark coordinates used to calculate shape, it seems appropriate to test allometry in Procrustean GMM using CS. Sometimes CS is transformed using the natural logarithm to better approximate the multiplicative nature of growth, but interpretations are less straightforward and, most of the time, logging makes no difference to the main conclusions (Klingenberg 2022). However, the specific measurement of size used for exploring its effect on shape really depends on the study question (Hallgrímsson *et al.* 2019). In taxonomy, body mass could be as interesting as CS. In practice, CS is often seen as a proxy for body mass, a type of information that is rarely available for all individuals in a sample. In fact, whether CS of a specific anatomical structure, measured using a certain landmark configuration, is indeed a good approximation of body mass cannot be taken for granted. However, using CS to measure size in studies of hard tissues has another advantage. Unlike body mass, the CS of a bone might be less dependent, at least in adults, on condition, seasonality and other sources of environmental variation. This is particularly true in marmots whose body mass varies geographically and almost double during the active season, before hibernation (Armitage 2000, 2014).

The multivariate regression of Procrustes shape variables on CS is not the only approach for studying allometries in GMM. Klingenberg (2022) provides a detailed overview of some of the main alternatives, including form spaces, where size is analysed together with shape. For instance, one can append the natural logarithm of CS to the Procrustes shape coordinates or, alternatively, restore CS after the superimposition or simply omit the CS standardization in the superimposition (Mitteroecker *et al.* 2013; Klingenberg 2016). Klingenberg (2022) shows that, if there is no appreciable allometry or there is clear allometry, results of different approaches are consistent. The analysis of form can be interesting in itself, regardless of focusing on allometry. As I have anticipated, most of the analyses in this paper can be done using Procrustes form variables instead of shape. However, form spaces tend to be dominated by size differences. If that happens, taxonomic comparisons too may become largely equivalent to analysing a set of linear measurements, with some results mirroring those of a simple analysis of CS. This seems undesirable in taxonomic applications. Then, because a taxonomist is interested in analysing shape separately from size (with the exception of allometry), Procrustes shape becomes the set of variables of choice and, in this case, a multivariate regression of shape onto CS might be the most appropriate method (Klingenberg 2022). Because allometric variation may not be linear, however, besides exploring log-CS, a researcher can try a curvilinear model by doing the multivariate regression onto a polynomial expansion of CS (e.g., Larson *et al.* 2018). The model inevitably becomes more complicated and, to be worth, the improvement in the fit of the regression (i.e., the Rsq) should be substantial.

The approach using multivariate regressions, plus, when there are groups, the group by size MANCOVA, has become almost a standard procedure in Procrustean GMM (Zelditch *et al.* 2004; Klingenberg 2011;

Rohlf 2015). Its implementation in user-friendly programs, unfortunately, is simple for the one-species at a time multivariate regression, but not straightforward for the MANCOVA. I used a series of regressions on dummy variables in TPSRegr, but this method has at least two main limitations: it takes some effort to understand how it works to carefully build the dummy variables and run the series of regressions; even more limiting is the fact that it cannot be used on 3D data, because TPSRegr is specific to 2D Procrustean GMM analyses. For 3D data, for now, there might be no other option than to export the PCs of shape from MorphoJ and do the MANCOVA in R or a commercial software.

However, there is an alternative, using a different method, which is applicable to both 2D and 3D shapes in user-friendly free software. Instead of using the species by CS MANCOVA, one can pairwise test in MorphoJ whether the angles between the allometric vectors of species specific regressions (those done in the first part of B6) are significantly smaller than expected by chance. More exactly, the software provides "*P*-values of the test against the null hypothesis that the vectors have random directions in the shape tangent space", as explained in the help file. Thus, with small and significant angles[25], one might infer that the allometric trajectories are less divergent than random expectations. This is not exactly the same as testing whether they are parallel, but provides at least some evidence that allometries might be approximately parallel. The conclusion should, therefore, generally be similar to that obtained with the test of the interaction term in the species by CS MANCOVA.

The test for the angle of allometric vectors must be done for each pair of species or, at least, for those with the largest samples. To obtain the test in MorphoJ, users have to select a regression (e.g., the regression of shape on CS in the hoary marmot) and, then, use the command in the menu *Comparison, Compare Vector Directions*. The software opens a window, where one can specify the second regression (e.g., shape on CS in woodchucks), with which to compute and test the vector angle. As usual with pairwise analyses, however, running the tests may be tedious, if there are many groups. For instance, with just six species of North American marmots, one has to run 15 tests. If all 15 living marmot species had been analysed, there would have been 105 tests. Because of multiple testing, there is also the need to be cautious and consider that some results may be significant simply because the rate of type I errors is inflated (see part A).

MorphoJ can also perform an analysis which is equivalent to testing the species factor in a 'species plus CS' MANCOVA. As with the MANCOVA, first one has to demonstrate that slopes of different species are not significantly different, so that allometries can be considered approximately parallel. This can be explored using angles, as explained above, if data are 3D and/or the user has no software for the MANCOVA. Once this is done, he/she can hold the effect of CS on shape constant in MorphoJ and pairwise test species differences in size-corrected shape data. To do it, for the marmot dataset, a user selects the full 445 specimens dataset and does a regression of the Procrustes shape coordinates on CS. However, this is done now, for this specific aim, without any test, but after having checked the box *Pooled regression within subgroups*, with species selected as a subgroups. With this option, the regression is equivalent to that of a 'species plus CS' MANCOVA which forces the allometric lines to be parallel. Finally, by selecting the output of the regression in the project tree, one can run a CVA on the regression residuals (*Comparison, Canonical Variate Analysis*, after carefully verifying that the type of data is *Residuals* and checking the box for the permutation test). The software will test pairwise size-corrected mean shape differences among all species. In fact, it takes slightly longer, but it might be even better to first perform a PCA on the regression residuals (selecting the regression output and then using the usual commands in MorphoJ). The resulting size-corrected PCs of shape can be exported to PAST where, by running the CVA, users obtain the test of significance of overall differences (a species

[25] I stress that to suggest approximately parallel allometries angles must be not only significantly smaller than random chance, but also fairly small. Sheets & Zelditch (2013) discuss other types of tests for vectors (e.g., testing if angles are significantly different from zero), which may be more specific and accurate to assess parallel or overlapping allometric trajectories.

MANOVA on size-corrected shape), the pairwise post-hoc multivariate parametric tests (equivalent to the permutation tests in MorphoJ's CVA), and also the cross-validated classification table (*confusion matrix*). If controlling for allometry has an impact on taxonomic differences, results on size-corrected shape data should be substantially different from those (B3, B4, B5) using the full shape information.

Even if allometry and a potential size-correction are exclusively explored in MorphoJ without running the MANCOVA, a researcher should, nonetheless, at least compare the Rsq of separate and parallel regressions. This can be done using the same dummy variables, as in TPSRegr, by running first a regression of the species and interaction block and, later, a second regression on CS and the species block. Also, allometries should be graphically summarized, but, in this respect, the options are the same as those I used for the MANOVA (i.e., a PCA of the predictions using separate regressions and, possibly, a plot of PLS1 of the regression predictions vs CS as in Fig. 9a).

At the end of this subsection of the Discussion, I would like to stress again that non-significant results should be interpreted with caution. For instance, if samples are small and the species by CS interaction in TPSRegr and/or the pairwise tests of allometric vectors in MorphoJ are not significant, that is more likely to hint at low statistical power rather than demonstrate parallel allometries. When N is small in all or most samples, even simple pairwise tests for group mean differences in CS or shape are at best preliminary and more sophisticated analyses, such as the MANCOVA, are probably not worth the effort, since tests are not powerful and results can be inaccurate. With small samples, thus, simply but cautiously exploring differences in plots is possibly the best choice. In contrast, with better data (i.e., large samples and reasonable p/N ratios), there are more options and analyses can become much more sophisticated. In that case, for those interested in learning more on comparisons of shape trajectories, including tests which provides finer resolution of the differences in allometric trajectories, I suggest the review by Sheets & Zelditch (2013).

B7) Species comparisons of the magnitude of size and shape variance

Methods (B7)

Homogeneity of variance is an assumption of all tests of group mean differences I have described and, therefore, large deviations from homoscedasticity should have already been spotted when assumptions of previous analyses were explored. However, here I focus specifically on a biological aspect, which is whether there are differences in variance magnitude between VAN, a small insular population, and the hoary marmot, its sister species, with a vast distributional range on the continent. For both CS and shape, the following comparisons will be done:
a) modern VAN vs the total sample of hoary marmots;
b) all VAN, both modern and subfossil, vs the total sample of hoary marmots;
c) all VAN vs two mutually exclusive random subsamples of hoary marmots of approximately the same size ($N \approx 50$) as the total VAN sample;
d) the total sample of hoary marmots vs mutually exclusive subsamples of either yellow-bellied marmots or of woodchucks, whose size is approximately the same ($N \approx 50$) as in the total VAN sample.

a–b) are used to assess if there are indications of reduced variance in the recent past of VAN, as one might expect because of population bottlenecks in a peripheral isolate[26]. Even if, given this expectation,

[26] Demographic bottlenecks are the most likely explanation for reduced variance in VAN (Cardini *et al.* 2007; Nagorsen & Cardini 2009; Kruckenhauser *et al.* 2009). However, at least for the present-day population, one cannot exclude that morphology is somewhat less variable also because individuals live in a very restricted geographic range: almost like in a 'common garden' experiment, in a relatively uniform environment, changes due to plasticity are unlikely to substantially increase phenotypic variation. In contrast, in the large variety of environmental conditions (latitude, altitude, humidity and rainfall, types of vegetation etc.) found across the huge range of the hoary marmots, both genetic and plastic responses are likely to be important in producing the observed phenotypic differences.

a one-tailed test is more appropriate, I will use two-tailed tests, which are simpler to implement and more conservative (i.e., less powerful, but also less prone to type I errors). c–d) help to explore whether differences in variance might be simply due to the ~ 50% smaller size of the sample of VAN ($N \leq 50$) compared to hoary marmots ($N = 108$). If that is the case, one would expect to find no differences between VAN and subsamples of hoary marmots with $N \approx N_{VAN} = 50$ (c), but significant differences between the total sample of hoary marmots ($N = 108$) and the randomized subsamples of other marmot species (d) with $N \approx N_{VAN} = 50$.

For both CS and shape, differences in the magnitude of variance are tested using a permutational version of the Levene's test, the most common univariate test of homogeneity of variance (Gastwirth *et al.* 2009), which is easily extended to multivariate data (Willmore *et al.* 2006). For univariate data like CS, the variance is reported in PAST when descriptive statistics are computed (*Statistics, Univariate statistics*). With multivariate data, there are different, complimentary, ways to calculate the magnitude of variance in a multidimensional space (Foote 1997; Drake & Klingenberg 2010; Fontaneto *et al.* 2017). None of them can capture the full complexity of a multivariate dataset. The multivariate extension of the Levene's test employs the simplest of these statistics, which is obtained by summing up the variances of each Procrustes shape coordinate or, which is the same in terms of result, the variances of their PCs. The sum of the variances of the PCs is sometimes called the 'trace' of the variance covariance matrix. These computations are easily done in a spreadsheet, but one can find the *Total variance* in the *Results* window of MorphoJ, immediately below the eigenvalues of a PCA. Thus, the total shape variance can be computed in this software by doing separate PCAs for each of the species of which one is testing differences in variances (VAN, hoary marmots, subsamples of these or other species etc.).

The rationale of the test is that, when variance is similar, the absolute deviations of individual measurements to their mean in a sample should, on average, be statistically the same as the deviations from the mean in a second sample. For size, the absolute deviations can be computed in a spreadsheet by taking the absolute value of the difference of individual CS from the mean of its species sample. This computation is done first for one sample (e.g., VAN) and then for the second sample (e.g., hoary marmots). Deviations are then pasted one next to the other in two adjacent columns of PAST.

For shape, the absolute deviations are the Procrustes distances of each individual to the mean of its sample. These are readily obtained, one sample at a time, by loading the raw coordinates[27] in TPSSmall, superimposing the data (click on *compute*) and saving the Procrustes distances from the mean shape (*File, Save, Procrustes d to reference* – see Discussion on the small, potential, inaccuracy introduced by this procedure). As for the absolute deviations of CS, the resulting distances are pasted one next to the other in two different columns in PAST.

Finally, the similarity in mean absolute deviations (be it size or shape) is tested using the permutation test of the t-test (select the two columns and click on *Statistics, F and T tests (two samples)*). By default, 9999 permutations (that become 10 000 including the observed difference) are used, but one can increase this number, if necessary. The P value is the last one, at the bottom of the output window, right above the box where the number of permutations is specified. Everything else in this window can be omitted and users should not be mislead by the P value called *p(same variance)*: this is the parametric P of the Levene's test performed on the observed data (the observed CS, for instance), but it is incorrect if applied to the deviations of sample measurements from the mean. The reason not to do the Levene's test directly on the observed data is, mainly, that this would work in PAST only for size, which is univariate. Another reason is that I tend to prefer resampling statistics (permutations) to avoid the assumption of

[27] In this instance, one could also use the Procrustes shape coordinates. Size is not used in TPSSmall. However, users must remember that the data are those with averaged replicates, i.e. the same data used in all comparisons of groups in part B.

Table 9. Comparisons of the magnitude of variance (var) between the first (species 1) and second (species 2) species*.

Data	Species 1	N	Var 1	Species 2	N	Var 2	Var 1/var 2	P
CS	van modern	22	13.0	cal	108	16.0	0.8	0.7877
	van	50	11.8	cal	108	16.0	0.7	0.3634
				cal1	54	17.0	0.7	0.3531
				cal2	54	15.4	0.8	0.4455
	fla1	52	30.8	cal	108	16.0	1.9	*0.0043*
	fla2	52	30.4				1.9	*0.0014*
	fla3	52	34.9				2.2	*0.0009*
	mon1	50	23.7				1.5	0.0287
	mon2	51	21.3				1.3	0.2126
shape	van modern	22	0.00098	cal	108	0.00173	0.6	*0.0001*
	van	50	0.00119	cal	108	0.00173	0.7	*0.0001*
				cal1	54	0.00173	0.7	*0.0007*
				cal2	54	0.00168	0.7	*0.0002*
	fla1	52	0.00199	cal	108	0.00173	1.1	0.0584
	fla2	52	0.00189				1.1	0.2796
	fla3	52	0.00182				1.1	0.5231
	mon1	50	0.00172				1.0	0.9978
	mon2	51	0.00169				1.0	0.8641

* In the main tests, species 1 is VAN using only contemporary individuals or adding the subfossils, and species 2 is the hoary marmot. However, to explore the sensitivity of results to the smaller N of VAN, tests are repeated also: a) comparing VAN (all individuals) to mutually exclusive random subsamples of hoary marmot, so that VAN and hoary marmots have similar N (balanced design); b) using mutually exclusive random subsamples of yellow-bellied marmots and woodchucks, with approximately the same size as VAN (all individuals), compared to the total sample of hoary marmots.

normality of the parametric version of the Levene's test. This is, usually, a minor issue, as the test is generally robust to non-normality (Gastwirth *et al.* 2009).

There is no specific visualization for this type of comparison of variance magnitude. For size, it is appropriate to use the same box and jitter-plots already drawn for the comparison of group mean differences (Fig. 1b). For shape, the area occupied by one or the other group in the ordination scatterplots (Fig. 4) might help to spot large differences in variance, even if the interpretation must be cautious, because one is inspecting a subspace of the multivariate dataset and this may not accurately reflect the structure of the observations in the full shape space.

Results (B7)

The results of the comparisons of the magnitude of size and shape variance are in Table 9. First, I discuss size, for which both modern VAN and the total VAN sample have CS variances ~ 20–30% smaller than found in hoary marmots. However, none of the comparisons is statistically significant. To explore the impact of sample size on results, interspecific comparisons of CS variance are repeated using random subsamples. If the hoary marmot is split into two mutually exclusive subsamples, whose size (N = 54) is approximately the same (N ≈ 50) as the total sample of VAN, results do not change: VAN still has smaller variance, but none of the tests is significant. This first randomization experiment preliminarily suggests that the larger variance of the hoary marmot is not simply due to its larger sample; nevertheless, the difference is small and statistically negligible.

The sensitivity of results to N is explored also by taking the opposite approach, which means comparing the total sample of hoary marmots (N = 108) with subsamples of other marmot species of approximately the same size (N ≈ 50) as the total sample of VAN. This sensitivity analysis is run only using woodchucks and yellow-bellied marmots, whose large N allows to randomly extract respectively two and three mutually exclusive subsamples of ~ 50 specimens. Now, it is the hoary marmot that, despite a larger sample size, display less CS variance in all tests. Differences are large and significant when hoary marmots are compared to yellow-bellied marmots, whose CS variance is about twice as big as in hoary marmots. When hoary marmots are compared to woodchuck subsamples, tests do not reach the 0.005 significance threshold, but variance is consistently ~ 40% larger in woodchucks. Combining the results of these and the previous tests, the conclusion is that hoary marmots have a relatively modest variance in CS. Despite the modest mandibular size variation, hoary marmots have still larger CS variance compared to VAN, although the data do not allow to confidently claim that the difference is statistically sound.

The comparisons of shape variance produce a rather different outcome. For shape, both the total sample and the subsamples of hoary marmots show significantly larger variance compared to VAN, whereas hoary marmots shape variance is similar to those of random subsamples of other marmots. Thus, the data suggest that shape variance is similar in all large samples of marmots except in VAN, whose variance is 30–40% smaller.

Discussion (B7)

The comparisons of the magnitude of CS and shape variances focused on VAN, an endemic species with a small and critically endangered population (https://www.iucnredlist.org/species/12828/22259184). VAN has been isolated at least since the end of the last glaciation, as sea level rose at the boundary between the Pleistocene and the Holocene, forming the narrow strait that now separates the Vancouver Island from the mainland. Genetic data, however, suggest that VAN has a longer history of incomplete isolation from the parental lineage on the continent, the coastal clade of the hoary marmot (Kerhoulas *et al.* 2015). The main split may have occurred between 0.4 and 1.2 million years ago (if not earlier (Rankin *et al.* 2019)), but hybridization events may have happened even after the separation (Kerhoulas *et al.* 2015). Despite hybridization and an isolation longer than generally assumed (Steppan *et al.* 2011), VAN has gone through one or several bottlenecks, which probably occurred in refugia on or nearby the island, where the ancestors of modern VAN survived during glaciations (Kruckenhauser *et al.* 2009; Brashares *et al.* 2010; Jackson *et al*. 2015). Thus, molecular studies show that a reduction in genetic variance of VAN started much earlier than the dramatic demographic decline recorded in the last few decades. As in previous research (Nagorsen & Cardini 2009), my study supports this conclusion. VAN shows less morphological variance than hoary marmots not only when the modern sample is analysed, but also when I pooled both modern and subfossil marmots, some dating back to several thousands of years ago (Nagorsen & Cardini 2009). Therefore, despite small differences between modern and subfossil specimens of VAN, the sample shows a relatively homogeneous mandibular morphology. Because this finding has already been discussed before (Nagorsen & Cardini 2009), I focus here first on a novel observation and later on methods.

In general, there is a good correspondence, in term of relative differences in intraspecific variability, between the genetics and the morphology of marmot mandibles. DNA analyses show that, at least using specific markers, VAN varies less than hoary marmots (Kruckenhauser *et al.* 2009; Brashares *et al.* 2010; Jackson *et al*. 2015), which, in turn, are much less variable than yellow-bellied marmots (Rankin *et al.* 2019). Likewise, CS varies in VAN less than in hoary marmots, although the difference is not significant, with hoary marmots significantly less variable than yellow-bellied marmots. When we look at shape, the pattern is similar, with VAN varying less than hoary marmots, which are slightly less variable than yellow-bellied marmots, but now it is VAN to show significance, compared to hoary marmots, whereas the comparison of hoary and yellow-bellied marmot variance is not significant. The sensitivity analyses using random subsamples suggests that differences between VAN and hoary marmots are not simply due to the smaller sample size of VAN.

Thus, mandibular data support genetic results in terms of variability increasing from VAN to hoary marmots and from the latter to yellow-bellied marmots. To my knowledge, a smaller phenotypic variance in hoary marmots has not been reported before, but it is in agreement with the predictions of the palaeoecological model of Polly *et al.* (2015). These authors suggested that, during the last glacial maximum, the hoary marmot "was extirpated from almost all of its modern range during glacial phases [and] impacted more than any other marmot species". The correspondence between genetic and phenotypic data goes even further for shape. Hoary marmots vary less than yellow-bellied marmots, but, when they are considered as a superspecies complex, together with Olympic marmots and VAN, the amount of evolutionary divergence between the two lineages of *Petromarmota* is similar in terms of depth of coalescence (Rankin *et al.* 2019). Mandibular shape data are consistent with this similarity. If hoary marmots are pooled with Olympic marmots and VAN, variance in shape increases (0.00196) and becomes almost identical to the variance found in yellow-bellied marmots (0.00193). For CS, however, variance remains much higher in yellow-bellied marmots (31 mm^2) even when compared to the pooled samples of all other *Petromarmota* species (17 mm^2). However, this is not unexpected for a presumably more labile trait such as size, if size changes in relation to ecological conditions, which are likely to be more variable in yellow-bellied marmots (see above) compared to VAN, hoary and Olympic marmots. That shape variance is the same pooling these three species as in yellow-bellied marmots, in contrast, is surprising, given the distinctive mandibular shape of VAN. Yet, even with the additional variance brought by VAN, variation in mandible shape within the hoary marmot superspecies complex is comparable to variation found within the yellow-bellied marmot alone. The good congruence between findings from shape and genetic data seem to support the speculation that shape might be more informative for evolutionary inference and less plastic and labile than size.

Although the analysis of the magnitude of variation in VAN was done mainly to explain how to run this type of comparisons, results were, as I discussed above, more interesting than expected in a replica study. They confirmed the reduced variance of VAN, found in previous work (Nagorsen & Cardini 2009), but also showed how yellow-bellied marmots seem to have a slightly larger amount of intraspecific variation in shape and a much larger one in size, with shape being largely congruent with molecular data (Rankin *et al.* 2019).

Tests of differences in variance magnitude in taxonomic studies using morphometrics are, unfortunately rare, but should probably be considered more often as a potential source of useful information. Besides testing for founder effects in peripheral isolates using the phenotype as a proxy for genetic variance, the comparison of size and shape variance may help to provide clues to cryptic diversity. This might happen when, for instance, a taxon is shown to have unusually large variance compared to its closest relatives. One might then exclude the population(s) that he/she believes might be responsible for the increase in variance (e.g., a peripheral isolate or a population found in a different habitat) and repeat the comparison of variance magnitude with other species to assess whether it is really that population that makes a difference.

The test for the differences in variance magnitude is simple, despite some limitations when it is performed in user-friendly software. As anticipated in the methods, the sum of the univariate variance is only one of several ways of estimating the magnitude of total variance in a multivariate dataset. There are alternatives, some of which are simple to compute. For instance, it is easy to compute the average or a trimmed upper boundary (e.g., the 90th percentile) for the pairwise Procrustes shape distances in a species sample and then compare it with those of other species. The matrix of distances after the Procrustes superimposition can be obtained, one species at a time, from TPSSmall (*File, Save, Procrustes distances*) or, using the equivalent Euclidean distances, in PAST (select all the shape variables and click on *Statistics, Similarity and distance indices*, checking the box *Euclidean* in the window that is opened). Averages or percentiles are, then, simple to calculate by importing the matrix in a spreadsheet. However, compared to the sum of variances, these summaries of multivariate variance cannot be tested for differences between groups in user-friendly programs. Simple resampling tests are, nonetheless, fairly easy to implement in R (e.g., Milella *et al.* 2021).

Another limitation related to the adoption of user-friendly programs concerns the computation of Procrustes distances to the mean, one species at a time, in TPSSmall. This procedure might introduce a small inaccuracy, because individuals are separately superimposed within species instead of using all species together, as it would be more correct. However, this is negligible for samples that are not very small and show moderate shape variation, as typical in taxonomic studies.

I also stressed multiple times that accurate estimates of variances require large samples (Cardini *et al.* 2021). With large samples and detailed information on localities, comparisons can be made more precise and accurate. For instance, the comparative sample of hoary marmots could have been more specific, as I might have compared VAN first to the members of the coastal clade of *M. caligata* (Kerhoulas *et al.* 2015; Mills *et al.* 2023) and, maybe only later, to the total sample of this species. In general, in this as in all other types of tests, knowing the locality of origin of all specimens allow to better design the comparisons and, also, to check to what extent a taxonomic sample is representative or, in alternative, potentially biased and autocorrelated.

Conclusions

Guidelines: aims, readership, limitations and further readings

My aim, with this paper and its twin on preliminary analyses, is to provide a guideline to taxonomic comparisons using Procrustean GMM and user-friendly software. The style of the papers is informal and sometimes colloquial, as they are written like a series of lectures in a workshop. Indeed, the idea behind this project was to put on paper the experience I have made over the years by teaching introductory GMM courses to biologists. For taxonomists, my main target readership, the main appeal of GMM is that this family of methods allows to compare groups using fairly simple morphological measurements taken on low cost, and easy to obtain, digital images. Taxonomists generally do not aim at becoming professional morphometricians and use morphometrics as a complimentary approach to qualitative morphological analyses, ecology, genetics or other methods. Sometimes, they are professionals and sometimes they are amateurs. Rarely they have the time or interest to delve deep into the methods. Also, among the wide range of GMM and statistical methods, they typically need mostly landmark-based techniques for testing group differences.

The guidelines are long, and have some repetitions, but should provide the level of detail that allows to perform careful, in-depth, taxonomic analyses in zoology using Procrustean GMM. The advice I give is extensive, but at the same time focused on practical and relatively simple applications. There are some theoretical digressions. However, they are simplified and limited to those issues I found myself, as a biologist with no background in statistics, hard to fully understand. As the papers are organized in chapters, like a book, one might skip those topics he/she feels more confident on or less interested in. If this is done, the references to other chapters should help to quickly check if one is missing something important. For beginners who plan to extensively used Procrustean GMM, however, I strongly suggest to endure the effort of the long reading and, thus, slowly go through both papers, as well as their appendices. As one proceeds with the reading, playing with the data, I made available, and replicating the analyses could help to better understand the theory while practicing with the programs. It is for this reason that I decided to combine the methodological introductions with the short instructions on the software.

Readers who want to acquire a deeper knowledge of GMM can learn more in the 'Green Book' [28] (Zelditch *et al.* 2004, 2012), to my knowledge still the only extensive but comprehensible manual for biologists. The book is not perfect (Rohlf 2005); even the second version is no longer up to date with some recent developments and issues; and there may be disagreements with the view of other morphometricians (including myself, on some topics). Nonetheless, at least the first version, which I am more familiar with, is a good starting point for improving one's knowledge of Procrustean GMM.

Morphometrics with R (Claude 2008) is excellent as an introduction to applications of traditional and geometric morphometircs in the R statistical environment. The GMM literature is, in fact, vast and includes books on theory (e.g., Dryden & Mardia 1998), edited volumes (e.g., the free 'Yellow Book' (Cardini & Loy 2013)) and innumerable papers both on theory and applications. There is also an increasing number of reviews, among which, in my opinion, Adams *et al*. (2004, 2013) remain the best in terms of conciseness, completeness and, most of the time, balance. Although the newest developments are missing, the review by O'Higgins on Procrustes methods (O'Higgins 2000) and his previous work on techniques for the analysis of outlines (O'Higgins 1997) are also exemplar of clarity. These two papers, together with Oxnard & O'Higgins (2009), are probably unbeatable for the strength with which they stress the crucial connection between measurements and biology.

A limitation apparently shared by the books and review papers I know, however, is the lack of an explicit acknowledgment that Procrustean GMM, like all morphometric methods based on superimposition, prevents accurate analyses of per-landmark variation (Cardini & Verderame 2022) and, for the same reasons, implies serious problems for methods using subsets of landmarks within a configuration (Cardini 2019, 2020b, 2023). All these issues are related to the biological arbitrariness of the superimposition, which has been acknowledged in a slightly different context (Moyers & Bookstein 1979) since before the time of the 'morphometric revolution' (Rohlf & Marcus 1993), but had been emphasized mainly by the proponents of alternative GMM approaches (Lele 1991; Richtsmeier *et al.* 2002).

Even within the narrow context of taxonomic comparisons using GMM, my guidelines have their own big limitations that I am happy to acknowledge. First of all, there is an inevitable bias in relation to my research interests and experience in the '20 plus' years since I started learning methods and working in this field. In terms of taxa, for instance, although I have worked on a variety of groups and I am not a specialized theriologist, most of my research has been on mammals, which explains the choice of marmots in this study and a prevalence of mammalogical examples. The main scaffold for the research design, however, is that of Rohlf *et al.* (1996), which was one of the earliest applications of Procrustean GMM to taxonomy and has served as a model for many other studies (see Introduction for more references). The most important difference, in my papers, is that a few common mistakes of the early days of Procrustean GMM (e.g., the occasional misuse of partial warps, later acknowledged by Rohlf 1998) have hopefully been removed. Also, I have somewhat expanded the analyses of that pioneering paper by often using resampling methods, as well as by including the study of allometries and the comparisons of morphological variance.

[28] Most of the main GMM books are nicknamed using the colour of the cover.

Beyond free user-friendly software: why learning R

In terms of tools available for taxonomic applications of GMM, much more important than not having exemplified semilandmarks in this study (but see Appendix A for a comment on these 'special points') are the limitations imposed by adopting user-friendly software. I made this choice after originally considering to implement everything in R. I did not do it, in the end, because scientists who are already 'fluent in R' can easily write scripts for all the analyses I suggest. In contrast, taxonomists, who do not already use R, might find it hard to learn a method, such as GMM, they have little or no familiarity with, while also learning how to write code in a complex statistical environment (Eglen 2009). Unless a researcher already has some programming experience, learning first GMM and later R (or vice versa) may be, for most users, simpler than trying to do both at the same time. Learning R is, nonetheless, advisable, especially for younger scientists and taxonomists who plan to extensively use GMM (or other quantitative methods) in their work. For the first steps in R, there is a wide range of free online tutorials and guides and, although they reduce flexibility and restrict the analytical options, there are some user-friendly graphical interfaces such as R Commander (Fox 2005, see also https://r4stats.com/2022/02/09/r-graphical-user-interface-comparison/ for alternatives). Tips on R commands or even scripts (to be carefully checked and acknowledged!) might now be obtained also using chatbots, such as the popular ChatGPT (https://openai.com/chatgpt). With R, there is versatility and a constant development of new packages for virtually all types of analyses in biological research. It is open source, and thus everyone can check that algorithms and computations are correct. Yet, especially in applications that are less commonly used and in fields of research with a limited number of experts, one cannot exclude errors in packages and scripts (Claes *et al.* 2014). Nonetheless, assuming everything has been done accurately, once a script has been written for a study, it is easy to share and modify it, re-run it or recycle it for a different research project. Scripting may be slow, but performing the analyses becomes faster.

Among the methods I made a limited use of in this work, there are several examples of how R can help to improve and expand the analyses. For instance, R allows to perform cross-validated bgPCAs both for classification and ordination using the *groupPCA()* function of Morpho (Schlager 2017). Tests of group mean differences in size or shape can be run using permutations of distances with the *adonis2()* function, which also calculates the corresponding Rsq, in Vegan (Oksanen *et al.* 2022). Bootstrapping to estimate confidence intervals also becomes relatively easy using functions such as *sample()*, which is part of the base R package (R Core Team 2023). Using bootstraps in R, one can also assess the robustness of a phenogram of mean shapes (Caumul & Polly 2005; Cardini & Elton 2008), estimate confidence envelopes in ordinations (Nagorsen & Cardini 2009) and perform extensive randomized subsampling experiments to assess the sensitivity of results to small and heterogeneous N (e.g., Cardini *et al.* 2021). With R, it is also easy to subset landmarks and semilandmarks and, thus, explore the congruence of findings with different configurations (Adams *et al.* 2011; Watanabe 2018), as well as the impact of dimensionality when p / N is large. The number of packages for morphometric analyses in R is already fairly large and likely to increase. In April 2023, searching "morphometrics" in https://cran.r-project.org/web/packages/available_packages_by_name.html, I retrieved at least 10 packages specific to morphometric analyses, mostly using GMM and Procrustes methods. Among the R packages useful for morphometricians, there are also some, like StereoMorph (Olsen & Westneat 2015) and SlicerMorph (Rolfe *et al.* 2021), that facilitates the collection and visualization of 3D data using low-cost photogrammetry.

Main steps and results, in brief

The aim and limitations of the study have been mentioned. Let me, now, concisely go back to the results, as well as the main suggestions and most important steps of the guidelines (emphasized in bold in this summary subsection).

As in all scientific research, a careful inspection of the **literature** on the topic one is interested in is, of course, preliminary to all other steps and fundamental for a careful study design. For marmots, most of the systematic research using morphology is old (see references in Wilson & Reeder 2005) and, at least at the level of subspecies, a revision is likely needed. There are, however, molecular phylogenetic analyses (Steppan *et al.* 1999, 2011; Kerhoulas *et al.* 2015; Rankin *et al.* 2019; Mills *et al.* 2023) and several interspecific comparisons of phenotypic variation, including morphometric analyses using traditional (Hoffmann *et al.* 1979) or geometric morphometric methods (Cardini *et al.* 2009, and references therein).

The choice of the **study structure** is critical in morphometrics and taxonomy. Mandibular morphology is often investigated in rodents to assess evolutionary differences (Velhagen & Roth 1997; Michaux *et al.* 2007; Renaud *et al.* 2007; Álvarez *et al.* 2021), including in marmots (Hoffmann *et al.* 1979; Cardini 2003; Nagorsen & Cardini 2009). However, practical considerations, such as sample availability, are as important as anatomical and evolutionary knowledge. Likewise, a careful **selection of landmarks** is necessary. The landmark configuration must capture the aspects relevant to the study questions. On marmot mandibles, I used landmarks that describe the overall proportions of this structure. I started with a larger configuration, but later excluded the most imprecise landmarks. Indeed, after data collection, the **assessment of landmark precision and, more generally, of measurement error** are the first analytical steps, together with the **search for potential outliers**.

Prior to the data acquisition for the main project, however, a researcher might also want to perform a **pilot study** on a small sample of a few species (subspecies or populations), for which specimens are easy to find in her/his institution or a nearby museum. A pilot study also provides the opportunity for a **prospective power analysis**, which is more useful to plan the minimum desirable sample sizes than the mostly retrospective tests I performed. With marmots, simulations showed that power is adequate in interspecific comparisons of mandibular shape even when samples are small (N = 10). Yet, such small samples do not allow accurate tests of sex differences. Besides, despite adequate power in between species tests, **small samples** are likely to inaccurately estimate means, variances and covariances (Cardini *et al.* 2021), and might inflate group differences and prevent a robust validation of results using classification methods such as DA/CVAs or bgPCAs (Kovarovic *et al.* 2011; Cardini & Polly 2020; Rohlf 2021).

Once the dataset is collected and 'cleaned' (by removing potentially low precision landmarks and outliers), the **proper group comparison of size and shape** starts. As common in taxonomic comparisons, I chose **adults** for the North American marmot study. However, this still requires **assessing sexual dimorphism** before comparing groups. Tests of sexual dimorphism one species at a time showed that it is typically very small in marmot mandibles, except when it is likely inflated by comparisons of small numbers of females and males. That sexual dimorphism is negligible, and similar in magnitude and direction across species, was confirmed by the **species by sex ANOVA**. The ANOVA also allowed to explore the relative magnitude of sex and taxonomic differences, with the latter shown to be much larger. Thus, **sexes were pooled and species compared**.

Interspecific tests demonstrated large and significant differences, with a few exceptions for mandibular size. VAN turned out to have mandibles of the same average size as hoary marmots, which is unsurprising given their phylogenetic relatedness (Kerhoulas *et al.* 2015). Yet, this result is interesting, as it does not follow the island rule (Millien & Damuth 2004; Lomolino *et al.* 2013), which predicts smaller size on islands for large rodents. However, the Vancouver Island is a large island in a temperate-cold region. Thus, the advantages of a larger size for a better thermoregulation may have countered other selective pressures that generally promote a smaller size in medium and large insular mammals (Lomolino *et al.* 2013).

The other interesting exception, where differences in average mandibular size were almost inexistent, came from the comparison of woodchucks and Alaskan marmots. Woodchucks live at lower altitude, mostly on plains, and may be less strongly impacted by cold and a short growing season. Alaskan marmots, in contrast, live near the Arctic, in an extreme environment, but are unusual for their small-to-intermediate body mass, compared to other marmot species (Armitage 2014). The Alaskan marmot, however, is relatively poorly studied and specimens are difficult to find. Most of those in my small sample were collected by Rausch. Year and locality are unknown for these specimens, but it is likely that they represent closely related individuals, possibly from the same colony or from a group of neighbouring families. Thus, not only statistical power and accuracy are low in my Alaskan marmot sample, but chances are good that the majority of the specimens are strongly autocorrelated. Because the observations are not independent, there is probably a bias in the estimates for this sample. The small sample of Olympic marmots could have similar issues. However, the geographic range and population size of this species is much smaller than that of Alaskan marmots, which suggests that the sample could be less strongly **impacted by poor representativeness and autocorrelation**. Yet, both species provide good potential examples of how poor sampling should be carefully considered in the analyses and the interpretation of their results. Indeed, even with mandibular differences in shape, that are always fairly large and highly significant, one cannot exclude inaccuracies, especially in the smallest samples.

In terms of shape variation, the analysis confirms that, even using samples of woodchucks, yellow-bellied and hoary marmots larger than in previous studies (Cardini *et al.* 2009; Nagorsen & Cardini 2009), **VAN remains the most distinctive species** for mandibular morphology. The long, posteriorly curved coronoid is almost diagnostic for VAN, although it might occasionally occur in other species. This is analogous to what happens with its dark fur, which is consistently found in VAN but may be present at low frequency in some populations of other species (Armitage 2009). For VAN, it might have been **more accurate to separate the modern and subfossil samples**, but I did not do it to simplify the design. There is no appreciable change in average size between modern and subfossil mandibles, but mean shape differences are significant (not shown). However, the magnitude of this within-species time-related variation is about half of the magnitude of VAN average interspecific differences (Rsq 11.8% vs 19.4%). Consistent with this estimate of larger inter- than intra-specific variation, the large majority of VAN individuals (76%) clusters together, to the exclusion of all individuals of other species except two, in a UPGMA phenogram of shape (not shown). The tight clustering of most specimens in VAN explains why, even pooling modern and subfossil mandibles, this species has the highest cross-validated hit-rate in the DA/CVA of shape. That differences within VAN are much smaller than interspecific differences had already been shown using a slightly different configuration of mandibular landmarks (Nagorsen & Cardini 2009). The relative homogeneity of the VAN sample was also demonstrated by the 30% **smaller size and shape variance** in this species compared to hoary marmots. The lower phenotypic variation and distinctive mandibular shape of VAN are in agreement with the expectations of the hypothesis of a founder effect in the ancestral population that originated the modern and subfossil samples (Cardini 2003; Nagorsen & Cardini 2009). Regardless of the relatively negligible separation (compared to interspecific variation) between the modern sample and the VAN subfossils, I stress that it is **important to explore potential subgroups within a species**, before deciding whether or not to pool them in interspecific analyses. In a didactic study, I opted for pooling to increase N, but excluding the subfossils, or keeping them as a separate samples, are also potential options. In fact, using separate samples for modern and extinct populations is generally advisable, as results are potentially more interesting and accurate.

A main study question, which was not asked in our previous example of how to perform taxonomic comparisons using Procrustean GMM (V&C), concerns the **effect of size on shape**. With large differences in mandibular size among some marmot species, one might be interested to know if shape differences are mostly size-related (i.e., allometric). Allometry is pervasive in evolution and, in the context of morphological change, may be crucial to preserve function (Emerson & Bramble 1993).

In marmots, however, mandibular size was found to explain relatively little shape variation. Yellow-bellied marmots were the only exception. Their strong pattern of allometric change is likely related to their larger variability in size, probably, in turn, due to a broader ecological niche, in terms of habitat and altitudinal range (Armitage 2014). With a generally small effect of size on shape, controlling for allometry is unlikely to appreciably change results. This expectation is consistent with the totally negligible effect on hit-rates using 'size-corrected' shape in the DA/CVA. Yet, the use of 'size-corrected' shape was only exemplified in this study, because the interspecific divergence in allometric trajectories prevented an accurate application of allometric corrections. Interestingly, however, the comparison of allometries among species produced some preliminary evidence that suggests a degree of congruence between allometric divergence and phylogenetic separation, since divergence was larger between than within marmot subgenera.

What is new? From a broader assessment of measurement error to the sensitivity of results to sampling

Finally, from a methodological perspective, the protocol for taxonomic studies using GMM I have exemplified adds a few new possibilities for exploring group differences in more detail compared to V&C. As in the previous subsection, I emphasize the new analyses using characters in bold.

Power analysis, allometry and differences in the magnitude of variance have already been mentioned. In part A, however, I showed how to quantify and plot digitization error (part A: fig. 2, table 3) so that **low precision landmarks** are easier to spot. I also exemplified, using data from previous studies and a simple simulation (Part A: fig. 5), **how measurement error may bias results** when (a) there is a consistent systematic error across most or all landmarks in relation to a **time-lag in the data collection**, when (b) a **single landmark is highly imprecise** and when (c) **error is isotropic but very large**. In the first two cases, one observes error-related group structure that biases results, whereas in the latter measurement error might mask true group differences.

In both papers, I also demonstrated of how to very **preliminarily explore the impact of small sample size on results**. The basic idea is to extract **random subsamples** from the species with the largest samples and assess the sensitivity of results not only when N is smaller (as small as in the species with the lowest N), but also when all species have the same N (balanced design). Thus, I assessed power in tests of within-species mean sex differences, as well as tests of species mean differences, using N comparable to that of the smallest samples (Table 5). I also used randomized subsample means to re-estimate mean shapes in the largest samples and check if, despite the larger sampling error, they suggested the same similarity relationship (Figs 6–7 in part B). The assumption of these randomized subsampling experiments is that findings from random subsamples of the largest species apply to the species with the smallest samples. If this assumption is correct in my North American marmot dataset, power might be adequate and estimates of means fairly precise in interspecific comparisons involving the small samples of Alaskan or Olympic marmots, but not in within-species tests of sexual dimorphism in these same species (as well as in VAN, where most individuals are unsexed).

Then, can we be sure that samples of just ~ 15 individuals in Alaskan and Olympic marmots are adequate for studying between species differences? Probably we cannot because, even if findings from randomized subsamples were generalizable across species, the subsample of the larger species do not simulate a bias in the data collection. That such a bias is likely in Alaskan marmots, and cannot be completely ruled out in other species, has already been discussed. **Excluding small samples from the main analyses** is an option to, at least, check that small samples do not alter the conclusions for the species with the largest samples. For marmots, the effect of the smallest samples on the main comparisons was negligible. This result and similar ones in the analyses of subsamples are interesting. Nonetheless, a researcher might worry that all the various sensitivity analyses, I exemplified, not only require extra efforts, but also make results less easy to summarize. Yet, I consider them worth, because they help to demonstrate what findings are more robust. It is these robust results, that, I argue, should be mainly discussed by a researcher. Others may disagree or might simply prefer a compromise to publish more and faster, rather than investing time to increase confidence using sensitivity analyses.

Final remarks on GMM in taxonomy

GMM is likely to remain an important tool for taxonomists. It may even become more popular, as museum collections are digitized and become accessible online, reducing the economic and environmental costs of visiting many museums. This is nicely exemplified by projects such as https://www.dissco.eu/ or https://www.idigbio.org/. As technologies for obtaining 2D and 3D images go on improving, it may also become easier to collect data in the field. A fairly cheap smartphone, as long as one demonstrates its accuracy for the specific task, might sometimes be enough. In some cases, it may even be possible to obtain images of anatomical parts or entire animals from live individuals, without the need of sacrificing them. Fishes, for instance, can be captured, anesthetized, carefully positioned and photographed, and rapidly released (Herler *et al.* 2007). Photographs of live individuals have also been used, in combination with genetics, to uncover cryptic diversity in Hermann's tortoises (Djurakic & Milankov 2020). Similar approaches have a clear potential in other organisms with complete or partial exoskeletons, including arthropods, which represent the majority of living animal species (Zhang 2013).

With my two papers, I hope to have suggested a fairly detailed step-by-step protocol for GMM comparisons of groups in taxonomy, but also in other fields interested in morphological group differences, including forensic and biomedical applications. The analyses are extensive, but easy to replicate and explained in simple terms. I state the obvious, but must stress that, even if all steps are carefully followed using large samples that produce robust evidence of group differences, taxonomists should resist the temptation to name new species or subspecies exclusively on the basis of morphometrics. This might be somewhat inevitable in palaeontology, where morphology is the main and often only criterion to describe taxa. However, even in this field, when modern analogues or relatives exist for the fossils a researcher is studying, one could first build a model that approximately estimates the expected degree of interspecific morphological differences in a lineage. For instance, Harvati *et al.* (2004) showed that the magnitude of cranial differences between adults of modern humans and Neanderthals is comparable to that of well studied species, or even genera, of living primates. Yet, phenotypic and genetic data can tell different stories and we now know that the separation between us and Neanderthals was incomplete, and hybridization was not rare and might have involved other hominins as well (Lahr 2021, and references therein). Thus, morphometrics provides an important line of evidence on evolutionary separation, but, on its own, it can only be preliminary. For a sound taxonomic assessment, the approach must be integrative and ideally combine genetic, eco-ethological, meristic and morphometric data (Dayrat 2005).

Acknowledgements

This paper (part B), as the previous one (part A), is dedicated to the memory of my father Alberto (1939–2022) and of my friend and colleague Luigi Sala (1954–2022): I owe them both so much. I am also in debt with many colleagues and, among them, I especially thank Göran Arnqvist, Amro Daboul, Joe Felsenstein, Dan Franklin, Øyvind Hammer, Ardern Hulme-Beaman, Paula Jenkins (and all museum curators I met over the years), Philipp Mitteroecker, Dave Nagorsen, David Polly, Riccardo Poloni, Jim Rohlf, Tim Smith, Davide Tamagnini, as well as all the co-authors of previous studies on marmots. I am most grateful also to all EJT editors for their superb work and especially to Frank Zachos (as editor and reviewer), as well as Vida Jojić and two anonymous referees for all the excellent suggestions. The study was supported by the Fondo Ateneo di Ricerca (project TAXON), a grant of the Italian Ministero dell'Università e della Ricerca (PRIN Project 2022MAM9ZB), and stimulated by countless productive discussions with colleagues during many morphometric workshops, as well as by the always fruitful

interactions with museum curators and the amazing team of SYNTHESYS (Synthesys of Systematic Resources: https://www.synthesys.info/ and https://www.dissco.eu/synthesys/).

References

Abouheif E. & Fairbairn D.J. 1997. A comparative analysis of allometry for sexual size dimorphism: Assessing Rensch's rule. *The American Naturalist* 149 (3): 540–562. https://doi.org/10.1086/286004

Adams D.C. & Nistri A. 2010. Ontogenetic convergence and evolution of foot morphology in European cave salamanders (Family: Plethodontidae). *BMC Evolutionary Biology* 10 (1): 216. https://doi.org/10.1186/1471-2148-10-216

Adams D.C., Rohlf F.J. & Slice D.E. 2004. Geometric morphometrics: ten years of progress following the 'revolution'. *Italian Journal of Zoology* 71 (1): 5–16.

Adams D.C., Cardini A., Monteiro L.R., O'Higgins P. & Rohlf F.J. 2011. Morphometrics and phylogenetics: Principal components of shape from cranial modules are neither appropriate nor effective cladistic characters. *Journal of Human Evolution* 60: 240–243. https://doi.org/10.1016/j.jhevol.2010.02.003

Adams D.C., Rohlf F.J. & Slice D.E. 2013. A field comes of age: geometric morphometrics in the 21st century. *Hystrix* 24 (1): 7–14. https://doi.org/10.4404/hystrix-24.1-6283

Albrecht G. 1992. Assessing the affinities of fossils using canonical variates and generalized distances. *Human Evolution* 7 (4): 49–69.

Álvarez A., Ercoli M.D., Olivares A.I., De Santi N.A. & Verzi D.H. 2021. Evolutionary patterns of mandible shape diversification of caviomorph rodents. *Journal of Mammalian Evolution* 28 (1): 47–58. https://doi.org/10.1007/s10914-020-09511-y

Amaral A.R., Coelho M.M., Marugán-Lobón J. & Rohlf F.J. 2009. Cranial shape differentiation in three closely related delphinid cetacean species: Insights into evolutionary history. *Zoology* 112 (1): 38–47. https://doi.org/10.1016/j.zool.2008.03.001

Anderson T.W. 1996. R.A. Fisher and multivariate analysis. *Statistical Science* 11 (1): 20–34.

Armitage K.B. 1999. Evolution of sociality in marmots. *Journal of Mammalogy* 80 (1): 1–10. https://doi.org/10.2307/1383202

Armitage K.B. 2000. The evolution, ecology, and systematics of marmots. *Oecologia Montana* 9 (1–2): 1–18.

Armitage K.B. 2005. Intraspecific variation in marmots. *In*: Sánchez-Cordero V. & Medellín R.A. (eds) *Contribuciones mastozoológicas en homenaje a Bernardo Villa. Instituto de Biología, UNAM*: 39–48.

Armitage K.B. 2009. Fur color diversity in marmots. *Ethology Ecology & Evolution* 21 (3): 183–194. https://doi.org/10.1080/08927014.2009.9522474

Armitage K.B. 2013. Climate change and the conservation of marmots. *Natural Science* 5: 36–43. https://doi.org/10.4236/ns.2013.55A005

Armitage K.B. 2014. *Marmot Biology: Sociality, Individual Fitness, and Population Dynamics*. Cambridge University Press, Cambridge UK.

Armstrong R.A. 2014. When to use the Bonferroni correction. *Ophthalmic and Physiological Optics* 34 (5): 502–508. https://doi.org/10.1111/opo.12131

Ateş C., Kaymaz Ö., Kale H.E. & Tekindal M.A. 2019. Comparison of test statistics of nonnormal and unbalanced samples for multivariate analysis of variance in terms of type-I error rates. *Computational and Mathematical Methods in Medicine* 2019: e2173638. https://doi.org/10.1155/2019/2173638

Berns C.M. & Adams D.C. 2013. Becoming different but staying alike: patterns of sexual size and shape dimorphism in bills of hummingbirds. *Evolutionary Biology* 40 (2): 246–260. https://doi.org/10.1007/s11692-012-9206-3

Blumstein D. 1999. Alarm calling in three species of marmots. *Behaviour* 136 (6): 731–757. https://doi.org/10.1163/156853999501540

Bookstein F.L. 1996. Landmark methods for forms without landmarks: localizing group differences in outline shape. *In*: Kavanaugh M.E. (ed.) *Proceedings of the Workshop on Mathematical Methods in Biomedical Image Analysis (IEEE), 1996*: 279–289.

Bogdanović A.M., Ivanović A., Tomanović Ž., Žikić V., Starý P. & Kavallieratos N.G. 2009. Sexual dimorphism in *Ephedrus persicae* (Hymenoptera: Braconidae: Aphidiinae): intraspecific variation in size and shape. *The Canadian Entomologist* 141 (6): 550–560. https://doi.org/10.4039/n09-029

Brashares J.S., Werner J.R. & Sinclair A.R.E. 2010. Social 'meltdown' in the demise of an island endemic: Allee effects and the Vancouver Island marmot. *Journal of Animal Ecology* 79 (5): 965–973. https://doi.org/10.1111/j.1365-2656.2010.01711.x

Campbell N. & Mahon R. 1974. A multivariate study of variation in two species of rock crab of the genus *Leptograpsus*. *Australian Journal of Zoology* 22 (3): 417–425.

Cardini A. 2003. The geometry of the marmot (Rodentia: Sciuridae) mandible: phylogeny and patterns of morphological evolution. *Systematic Biology* 52 (2): 186–205. https://doi.org/10.1080/10635150390192807

Cardini A. 2019. Integration and modularity in procrustes shape data: Is there a risk of spurious results? *Evolutionary Biology* (46): 90–105. https://doi.org/10.1007/s11692-018-9463-x

Cardini A. 2020a. Modern morphometrics and the study of population differences: Good data behind clever analyses and cool pictures? *The Anatomical Record* 303 (11): 2747–2765. https://doi.org/10.1002/ar.24397

Cardini A. 2020b. Less tautology, more biology? A comment on "high-density" morphometrics. *Zoomorphology* 139 (4): 513–529. https://doi.org/10.1007/s00435-020-00499-w

Cardini A. 2022. As fast as a hare: Did intraspecific morphological change bring the Hallands Väderö Island population of *Lepus timidus* close to interspecific differences in less than 150 years? *Zoology* 152: 126014. https://doi.org/10.1016/j.zool.2022.126014

Cardini A. 2023. Shall we all adopt, with no worries, the 'within a configuration' approach in geometric morphometrics? A comment on claims that the effect of the superimposition and sliding on shape data is "not an obstacle to analyses of integration and modularity". *EcoEvoRxiv*, unpublished preprint. https://doi.org/10.32942/X2002C

Cardini A. & Elton S. 2007. Sample size and sampling error in geometric morphometric studies of size and shape. *Zoomorphology* 126 (2): 121–134. https://doi.org/10.1007/s00435-007-0036-2

Cardini A. & Elton S. 2008. Does the skull carry a phylogenetic signal? Evolution and modularity in the guenons. *Biological Journal of the Linnean Society* 93 (4): 813–834. https://doi.org/10.1111/j.1095-8312.2008.01011.x

Cardini A. & Elton S. 2009. The radiation of red colobus monkeys (Primates, Colobinae): morphological evolution in a clade of endangered African primates. *Zoological Journal of the Linnean Society* 157 (1): 197–224. https://doi.org/10.1111/j.1096-3642.2009.00508.x

Cardini A. & Loy A. 2013. On growth and form in the computer era: from geometric to biological morphometrics. *Hystrix, the Italian Journal of Mammalogy* 24 (1): 1–5. https://doi.org/10.4404/hystrix-24.1-8749

Cardini A. & O'Higgins P. 2005. Post-natal ontogeny of the mandible and ventral cranium in Marmota species (Rodentia, Sciuridae): allometry and phylogeny. *Zoomorphology* 124 (4): 189–203. https://doi.org/10.1007/s00435-005-0008-3

Cardini A. & Polly P.D. 2020. Cross-validated between group PCA scatterplots: A solution to spurious group separation? *Evolutionary Biology* 47 (1): 85–95. https://doi.org/10.1007/s11692-020-09494-x

Cardini A. & Tongiorgi P. 2003. Yellow-bellied marmots (*Marmota flaviventris*) 'in the shape space' (Rodentia, Sciuridae): sexual dimorphism, growth and allometry of the mandible. *Zoomorphology* 122 (1): 11–23. https://doi.org/10.1007/s00435-002-0063-y

Cardini A. & Verderame M. 2022. Procrustes shape cannot be analyzed, interpreted or visualized one landmark at a time. *Evolutionary Biology* 49 (2): 239–254. https://doi.org/10.1007/s11692-022-09565-1

Cardini A., Thorington R.W. & Polly P.D. 2007. Evolutionary acceleration in the most endangered mammal of Canada: speciation and divergence in the Vancouver Island marmot (Rodentia, Sciuridae). *Journal of Evolutionary Biology* 20 (5): 1833–1846. https://doi.org/10.1111/j.1420-9101.2007.01398.x

Cardini A., Nagorsen D., O'Higgins P., Polly P.D., Thorington R.W. & Tongiorgi P. 2009. Detecting biological distinctiveness using geometric morphometrics: an example case from the Vancouver Island marmot. *Ethology Ecology & Evolution* 21 (3): 209–223. https://doi.org/10.1080/08927014.2009.9522476

Cardini A., Filho J.A.F.D., Polly P.D. & Elton S. 2010. Biogeographic analysis using geometric morphometrics: clines in skull size and shape in a widespread African arboreal monkey. *In*: Elewa A.M.T. (ed.) *Morphometrics for Nonmorphometricians*: 191–217. Springer Berlin Heidelberg, Berlin, Heidelberg.

Cardini A., Seetah K. & Barker G. 2015. How many specimens do I need? Sampling error in geometric morphometrics: testing the sensitivity of means and variances in simple randomized selection experiments. *Zoomorphology* 134 (2): 149–163. https://doi.org/10.1007/s00435-015-0253-z

Cardini A., O'Higgins P. & Rohlf F.J. 2019. Seeing distinct groups where there are none: Spurious patterns from between-group PCA. *Evolutionary Biology* 46 (4): 303–316. https://doi.org/10.1007/s11692-019-09487-5

Cardini A., Elton S., Kovarovic K., Strand Viđarsdóttir U. & Polly P.D. 2021. On the misidentification of species: sampling error in primates and other mammals using geometric morphometrics in more than 4000 individuals. *Evolutionary Biology* 48 (2): 190–220. https://doi.org/10.1007/s11692-021-09531-3

Cardini A., de Jong Y.A. & Butynski T.M. 2022. Can morphotaxa be assessed with photographs? Estimating the accuracy of two-dimensional cranial geometric morphometrics for the study of threatened populations of African monkeys. *The Anatomical Record* 305 (6): 1402–1434. https://doi.org/10.1002/ar.24787

Castiglione S., Serio C., Tamagnini D., Melchionna M., Mondanaro A., Febbraro M.D., Profico A., Piras P., Barattolo F. & Raia P. 2019. A new, fast method to search for morphological convergence with shape data. *PLOS ONE* 14 (12): e0226949. https://doi.org/10.1371/journal.pone.0226949

Caumul R. & Polly P.D. 2005. Phylogenetic and environmental components of morphological variation: skull, mandible, and molar shape in marmots (Marmota, Rodentia). *Evolution* 59 (11): 2460–2472. https://doi.org/10.1111/j.0014-3820.2005.tb00955.x

Claes M., Mens T. & Grosjean P. 2014. On the maintainability of CRAN packages. *In*: *2014 Software Evolution Week - IEEE Conference on Software Maintenance, Reengineering, and Reverse Engineering (CSMR-WCRE)*: 308–312. 2014 Software Evolution Week - IEEE Conference on Software Maintenance, Reengineering, and Reverse Engineering (CSMR-WCRE).

Claude J. 2008. *Morphometrics with R*. Springer Verlag, New York (US).

Corti M. & Rohlf F.J. 2001. Chromosomal speciation and phenotypic evolution in the house mouse. *Biological Journal of the Linnean Society* 73 (1): 99–112. https://doi.org/10.1111/j.1095-8312.2001.tb01349.x

Damuth J. 1993. Cope's rule, the island rule and the scaling of mammalian population density. *Nature* 365 (6448): 748–750. https://doi.org/10.1038/365748a0

Dayrat B. 2005. Towards integrative taxonomy. *Biological Journal of the Linnean Society* 85 (3): 407–417. https://doi.org/10.1111/j.1095-8312.2005.00503.x

de Groot C., Perdigao A.L. & Deurenberg P. 1996. Longitudinal changes in anthropometric characteristics of elderly Europeans. SENECA Investigators. *European Journal of Clinical Nutrition* 50 Suppl 2: S9–15.

de Moura Bubadué J., Cáceres N., dos Santos Carvalho R. & Meloro C. 2016. Ecogeographical variation in skull shape of South-American canids: Abiotic or biotic processes? *Evolutionary Biology* 43 (2): 145–159. https://doi.org/10.1007/s11692-015-9362-3

de Queiroz K. & Good D.A. 1997. Phenetic clustering in biology: A critique. *The Quarterly Review of Biology* 72 (1): 3–30. https://doi.org/10.1086/419656

Djurakic M.R. & Milankov V.R. 2020. The utility of plastron shape for uncovering cryptic diversity in Hermann's tortoise. *Journal of Zoology* 310 (2): 145–157. https://doi.org/10.1111/jzo.12736

Drake A.G. & Klingenberg C.P. 2010. Large-scale diversification of skull shape in domestic dogs: Disparity and modularity. *The American Naturalist* 175 (3): 289–301.

Dryden I.L. & Mardia K.V. 1998. *Statistical Shape Analysis*. John Wiley & Sons New York.

Eglen S.J. 2009. A quick guide to teaching R programming to computational biology students. *PLoS Computational Biology* 5 (8): e1000482. https://doi.org/10.1371/journal.pcbi.1000482

Elton S., Dunn J. & Cardini A. 2010. Size variation facilitates population divergence but does not explain it all: an example study from a widespread African monkey. *Biological Journal of the Linnean Society* 101 (4): 823–843. https://doi.org/10.1111/j.1095-8312.2010.01504.x

Emerson S. & Bramble D. 1993. Scaling, allometry and skull design. *In*: Hanken J. & Hall B.K. (eds) *The Skull: Volume 3, Functional and Evolutionary Mechanisms*: 384-421.

Evin A., Cucchi T., Cardini A., Strand Vidarsdottir U., Larson G. & Dobney K. 2013. The long and winding road: identifying pig domestication through molar size and shape. *Journal of Archaeological Science* 40 (1): 735–743. https://doi.org/10.1016/j.jas.2012.08.005

Felsenstein J. 2004. *Inferring Phylogenies*. Sinauer Associates, Incorporated, Sunderland, Massachusetts.

Fontaneto D., Panisi M., Mandrioli M., Montardi D., Pavesi M. & Cardini A. 2017. Estimating the magnitude of morphoscapes: how to measure the morphological component of biodiversity in relation to habitats using geometric morphometrics. *The Science of Nature* 104 (7): 55. https://doi.org/10.1007/s00114-017-1475-3

Foote M. 1997. The evolution of morphological diversity. *Annual Review of Ecology and Systematics* 28: 129–152.

Foster J.B. 1964. Evolution of mammals on islands. *Nature* 202 (4929): 234–235. https://doi.org/10.1038/202234a0

Fox J. 2005. The R commander: A basic-statistics graphical user interface to R. *Journal of Statistical Software* 14: 1–42. https://doi.org/10.18637/jss.v014.i09

Fox J. & Weisberg S. 2019. *An R Companion to Applied Regression*. Third. Sage, Thousand Oaks, CA.

Frost S.R., Marcus L.F., Bookstein F.L., Reddy D.P. & Delson E. 2003. Cranial allometry, phylogeography, and systematics of large-bodied papionins (primates: Cercopithecinae) inferred from geometric morphometric analysis of landmark data. *The Anatomical Record Part A: Discoveries in Molecular, Cellular, and Evolutionary Biology* 275A (2): 1048–1072. https://doi.org/10.1002/ar.a.10112

Gastwirth J.L., Gel Y.R. & Miao W. 2009. The impact of Levene's test of equality of variances on statistical theory and practice. *Statistical Science* 24 (3): 343–360. https://doi.org/10.1214/09-STS301

Gayon J. 2000. History of the concept of allometry. *American Zoologist* 40 (5): 748–758. https://doi.org/10.1093/icb/40.5.748

Gidaszewski N.A., Baylac M. & Klingenberg C.P. 2009. Evolution of sexual dimorphism of wing shape in the *Drosophila melanogaster* subgroup. *BMC Evolutionary Biology* 9 (1): 110. https://doi.org/10.1186/1471-2148-9-110

Glorfeld L.W. 1995. An improvement on Horn's parallel analysis methodology for selecting the correct number of factors to retain. *Educational and Psychological Measurement* 55 (3): 377–393. https://doi.org/10.1177/0013164495055003002

González S., Álvarez-Valin F. & Maldonado J.E. 2002. Morphometric differentiation of endangered pampas deer (*Ozotoceros bezoarticus*), with description of new subspecies from Uruguay. *Journal of Mammalogy* 83 (4): 1127–1140. https://doi.org/fw65d4

Goossens B., Graziani L., Waits L.P., Farand E., Magnolon S., Coulon J., Bel M.-C., Taberlet P. & Allainé D. 1998. Extra-pair paternity in the monogamous Alpine marmot revealed by nuclear DNA microsatellite analysis. *Behavioral Ecology and Sociobiology* 43 (4): 281–288. https://doi.org/10.1007/s002650050492

Grossnickle D.M. 2020. Feeding ecology has a stronger evolutionary influence on functional morphology than on body mass in mammals. *Evolution* 74 (3): 610–628. https://doi.org/10.1111/evo.13929

Hair J.F., Black W.C., Babin B.J. & Anderson R.E. 2013. *Multivariate Data Analysis*. Pearson Education Limited.

Hallgrímsson B., Katz D.C., Aponte J.D., Larson J.R., Devine J., Gonzalez P.N., Young N.M., Roseman C.C. & Marcucio R.S. 2019. Integration and the developmental genetics of allometry. *Integrative and Comparative Biology* 59 (5): 1369–1381. https://doi.org/10.1093/icb/icz105

Harvati K., Frost S.R. & McNulty K.P. 2004. Neanderthal taxonomy reconsidered: Implications of 3D primate models of intra- and interspecific differences. *Proceedings of the National Academy of Sciences* 101 (5): 1147–1152. https://doi.org/10.1073/pnas.0308085100

Hayssen V. 2008. Patterns of body and tail length and body mass in Sciuridae. *Journal of Mammalogy* 89 (4): 852–873. https://doi.org/10.1644/07-MAMM-A-217.1

Herler J., Kerschbaumer M., Mitteroecker P., Postl L. & Sturmbauer C. 2010. Sexual dimorphism and population divergence in the Lake Tanganyika cichlid fish genus *Tropheus*. *Frontiers in Zoology* 7 (1): 4. https://doi.org/10.1186/1742-9994-7-4

Herler J., Lipej L. & Makovec T. 2007. A simple technique for digital imaging of live and preserved small fish specimens. *Cybium* 31 (1): 39–44.

Hoffmann R.S., Koeppl J.W. & Nadler C.F. 1979. The relationship of the amphiberigian marmots (Mammalia, Sciuridae). *Occasional Papersof the Museum of Natural History of the University of Kansas* 83: 1–56.

Howell D.C. 2013. *Statistical Methods for Psychology (Eight Edition).* Wadsworth Cengage Learning, Wadsworth (US).

Ivanović A., Sotiropoulos K., Džukić G. & Kalezić M.L. 2009. Skull size and shape variation versus molecular phylogeny: a case study of alpine newts (*Mesotriton alpestris*, Salamandridae) from the Balkan Peninsula. *Zoomorphology* 128 (2): 157–167. https://doi.org/10.1007/s00435-009-0085-9

Jackson C., Baker A., Doyle D., Franke M., Jackson V., Lloyd N., McAdie M., Stephens T. & Traylor-Holzer K. 2015. *Vancouver Island Marmot Population and Habitat Viability Assessment Workshop Final Report*. IUCN SSC Conservation Breeding Specialist Group, Apple Valley, MN: 70.

Jarman P.J. 1974. The social organisation of antelope in relation to their ecology. *Behaviour* 48 (1–4): 215–267. https://doi.org/10.1163/156853974X00345

Kerhoulas N.J., Gunderson A.M. & Olson L.E. 2015. Complex history of isolation and gene flow in hoary, Olympic, and endangered Vancouver Island marmots. *Journal of Mammalogy* 96 (4): 810–826. https://doi.org/10.1093/jmammal/gyv089

Klenovšek T. & Kryštufek B. 2013. An ontogenetic perspective on the study of sexual dimorphism, phylogenetic variability, and allometry of the skull of European ground squirrel, *Spermophilus citellus* (Linnaeus, 1766). *Zoomorphology* 132 (4): 433–445. https://doi.org/10.1007/s00435-013-0196-1

Klingenberg C.P. 1996. Multivariate allometry. *In*: Marcus L.F., Corti M., Loy A., Naylor G.J.P. & Slice D.E. (eds) *Advances in Morphometrics*: 23–49. Plenum Press, New York.

Klingenberg C.P. 1998. Heterochrony and allometry: the analysis of evolutionary change in ontogeny. *Biological Reviews* 73 (1): 79–123. https://doi.org/10.1111/j.1469-185X.1997.tb00026.x

Klingenberg C.P. 2011. MorphoJ: an integrated software package for geometric morphometrics. *Molecular Ecology Resources* 11 (2): 353–357. https://doi.org/10.1111/j.1755-0998.2010.02924.x

Klingenberg C.P. 2013. Visualizations in geometric morphometrics: how to read and how to make graphs showing shape changes. *Hystrix, the Italian Journal of Mammalogy* 24 (1): 15–24. https://doi.org/10.4404/hystrix-24.1-7691

Klingenberg C.P. 2016. Size, shape, and form: concepts of allometry in geometric morphometrics. *Development Genes and Evolution* 226 (3): 113–137. https://doi.org/10.1007/s00427-016-0539-2

Klingenberg C.P. 2022. Methods for studying allometry in geometric morphometrics: a comparison of performance. *Evolutionary Ecology* 36 (4): 439–470. https://doi.org/10.1007/s10682-022-10170-z

Klingenberg C.P. & Monteiro L.R. 2005. Distances and directions in multidimensional shape spaces: Implications for morphometric applications. *Systematic Biology* 54 (4): 678–688. https://doi.org/10.1080/10635150590947258

Klingenberg C.P., Barluenga M. & Meyer A. 2002. Shape analysis of symmetric structures: Quantifying variation among individuals and asymmetry. *Evolution* 56 (10): 1909–1920. https://doi.org/10.1111/j.0014-3820.2002.tb00117.x

Kovarovic K., Aiello L.C., Cardini A. & Lockwood C.A. 2011. Discriminant function analyses in archaeology: are classification rates too good to be true? *Journal of Archaeological Science* 38 (11): 3006–3018. https://doi.org/10.1016/j.jas.2011.06.028

Kruckenhauser L., Bryant A.A., Griffin S.C., Amish S.J. & Pinsker W. 2009. Patterns of within and between-colony microsatellite variation in the endangered Vancouver Island marmot (*Marmota vancouverensis*): implications for conservation. *Conservation Genetics* 10 (6): 1759–1772. https://doi.org/10.1007/s10592-008-9779-7

Krzywinski M. & Altman N. 2014. Comparing samples—part II. *Nature Methods* 11 (4): 355–356. https://doi.org/10.1038/nmeth.2900

Kyle C.J., Karels T.J., Davis C.S., Mebs S., Clark B., Strobeck C. & Hik D.S. 2007. Social structure and facultative mating systems of hoary marmots (*Marmota caligata*). *Molecular Ecology* 16 (6): 1245–1255. https://doi.org/10.1111/j.1365-294X.2006.03211.x

Lahr M.M. 2021. The complex landscape of recent human evolution. *Science* 372 (6549): 1395–1396. https://doi.org/10.1126/science.abj3077

Larson J.R., Manyama M.F., Cole J.B., Gonzalez P.N., Percival C.J., Liberton D.K., Ferrara T.M., Riccardi S.L., Kimwaga E.A., Mathayo J., Spitzmacher J.A., Rolian C., Jamniczky H.A., Weinberg S.M., Roseman C.C., Klein O., Lukowiak K., Spritz R.A. & Hallgrimsson B. 2018. Body size and allometric variation in facial shape in children. *American Journal of Physical Anthropology* 165 (2): 327–342. https://doi.org/10.1002/ajpa.23356

Lele S. 1991. Some comments on coordinate-free and scale-invariant methods in morphometrics. *American Journal of Physical Anthropology* 85 (4): 407–417. https://doi.org/10.1002/ajpa.1330850405

Lindenfors P., Gittleman J. L., Jones K. E. (2007). Sexual size dimorphism in mammals. *In*: Fairbairn D.J., Blanckenhorn W.U. & Szekely T. (eds) *Sex, Size, and Gender Roles: Evolutionary Studies of Sexual Size Dimorphism*: 16–26. Oxford University Press.

Loison A., Gaillard J.-M., Pélabon C. & Yoccoz N.G. 1999. What factors shape sexual size dimorphism in ungulates? *Evolutionary Ecology Research* 1 (5): 611–633.

Lomolino M.V., van der Geer A.A., Lyras G.A., Palombo M.R., Sax D.F. & Rozzi R. 2013. Of mice and mammoths: generality and antiquity of the island rule. *Journal of Biogeography* 40 (8): 1427–1439. https://doi.org/10.1111/jbi.12096

Machado F.D.A. & Hingst-Zaher E. 2009. Investigating South American biogeographic history using patterns of skull shape variation on *Cerdocyon thous* (Mammalia: Canidae). *Biological Journal of the Linnean Society* 98 (1): 77–84. https://doi.org/10.1111/j.1095-8312.2009.01274.x

Maher C.R. & Duron M. 2010. Mating system and paternity in woodchucks (*Marmota monax*). *Journal of Mammalogy* 91 (3): 628–635. https://doi.org/10.1644/09-MAMM-A-324.1

Marcus L.F., Bello E. & García-Valdecasas A. 1993. *Contributions to Morphometrics*. Editorial CSIC, CSIC Press, Madrid.

Marcus L.F., Hingst-Zaher E. & Zaher H. 2000. Application of landmark morphometrics to skulls representing the orders of living mammals. *Hystrix, the Italian Journal of Mammalogy* 11 (1): 24–47. https://doi.org/10.4404/hystrix-11.1-4135

Marroig G. & Cheverud J.M. 2005. Size as a line of least evolutionary resistance: diet and adaptive morphological radiation in new world monkeys. *Evolution* 59 (5): 1128–1142. https://doi.org/10.1111/j.0014-3820.2005.tb01049.x

Matějů J. & Kratochvíl L. 2013. Sexual size dimorphism in ground squirrels (Rodentia: Sciuridae: Marmotini) does not correlate with body size and sociality. *Frontiers in Zoology* 10 (1): 27. https://doi.org/10.1186/1742-9994-10-27

Maurer B.A., Brown J.H. & Rusler R.D. 1992. The micro and macro in body size evolution. *Evolution* 46 (4): 939–953. https://doi.org/10.1111/j.1558-5646.1992.tb00611.x

Meloro C., Guidarelli G., Colangelo P., Ciucci P. & Loy A. 2017. Mandible size and shape in extant Ursidae (Carnivora, Mammalia): A tool for taxonomy and ecogeography. *Journal of Zoological Systematics and Evolutionary Research* 55 (4): 269–287. https://doi.org/10.1111/jzs.12171

Michaux J., Chevret P. & Renaud S. 2007. Morphological diversity of Old World rats and mice (Rodentia, Muridae) mandible in relation with phylogeny and adaptation. *Journal of Zoological Systematics and Evolutionary Research* 45 (3): 263–279. https://doi.org/10.1111/j.1439-0469.2006.00390.x

Milella M., Franklin D., Belcastro M.G. & Cardini A. 2021. Sexual differences in human cranial morphology: Is one sex more variable or one region more dimorphic? *The Anatomical Record* 304 (12): 2789–2810. https://doi.org/10.1002/ar.24626

Millien V. 2006. Morphological evolution is accelerated among island mammals. *PLoS Biology* 4 (10): e321. https://doi.org/10.1371/journal.pbio.0040321

Millien V. & Damuth J. 2004. Climate change and size evolution in an island rodent species: New perspectives on the island rule. *Evolution* 58 (6): 1353–1360. https://doi.org/10.1111/j.0014-3820.2004.tb01713.x

Mills K.K., Everson K.M., Hildebrandt K.B.P., Brandler O.V., Steppan S.J. & Olson L.E. 2023. Ultraconserved elements improve resolution of marmot phylogeny and offer insights into biogeographic history. *Molecular Phylogenetics and Evolution* 184: 107785. https://doi.org/10.1016/j.ympev.2023.107785

Mitteroecker P., Gunz P., Windhager S. & Schaefer K. 2013. A brief review of shape, form, and allometry in geometric morphometrics, with applications to human facial morphology. *Hystrix, the Italian Journal of Mammalogy* 24 (1): 59–66. https://doi.org/10.4404/hystrix-24.1-6369

Moore D.S. & McCabe G.P. 2005. *Introduction to the Practice of Statistics*. WH Freeman & Co., New York.

Moyers R.E. & Bookstein F.L. 1979. The inappropriateness of conventional cephalometrics. *American Journal of Orthodontics* 75 (6): 599–617.

Nagorsen D.W. & Cardini A. 2009. Tempo and mode of evolutionary divergence in modern and Holocene Vancouver Island marmots (*Marmota vancouverensis*) (Mammalia, Rodentia). *Journal of Zoological Systematics and Evolutionary Research* 47 (3): 258–267. https://doi.org/10.1111/j.1439-0469.2008.00503.x

Neff N.A. & Marcus L.F. 1980. *A Survey of Multivariate Methods for Systematics*. American Museum of Natural History, New York.

O'Higgins P. 1997. Methodological issues in the description of forms. *In*: Lestrel P. (ed.) *Fourier Descriptors and their Applications in Biology*: 74–105. Cambridge: Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511529870.005

O'Higgins P. 2000. The study of morphological variation in the hominid fossil record: biology, landmarks and geometry. *Journal of Anatomy* 197 (1): 103–120. https://doi.org/10.1046/j.1469-7580.2000.19710103.x

O'Keefe F.R., Meachen J.A. & Polly P.D. 2022. On information rank deficiency in phenotypic covariance matrices. *Systematic Biology* 71 (4): 810–822. https://doi.org/10.1093/sysbio/syab088

Oksanen J., Simpson G.L., Blanchet F.G., Kindt R., Legendre P., Minchin P.R., O'Hara R.B., Solymos P., Stevens M.H.H., Szoecs E., Wagner H., Barbour M., Bedward M., Bolker B., Borcard D., Carvalho G., Chirico M., De Caceres M., Durand S., Antoniazi Evangelista H.B., Fitzjohn R., Friendly M., Furneaux B., Hannigan G., Hill M.O., Lahti L., McGlinn D., Ouellette M.-H., Ribeiro Cunha E., Smith T., Stier A., Ter Braak C.J.F. & Weedon J. 2022. *vegan: Community Ecology Package*. Available from https://cran.r-project.org/web/packages/vegan/index.html [accessed 9 Apr. 2024].

Olsen A.M. & Westneat M.W. 2015. StereoMorph: an R package for the collection of 3D landmarks and curves using a stereo camera set-up. *Methods in Ecology and Evolution* 6 (3): 351–356. https://doi.org/10.1111/2041-210X.12326

Oxnard C. & O'Higgins P. 2009. Biology clearly needs morphometrics. Does morphometrics need biology? *Biological Theory* 4 (1): 84–97. https://doi.org/10.1162/biot.2009.4.1.84

Pearson A., Groves C. & Cardini A. 2015. The 'temporal effect' in hominids: Reinvestigating the nature of support for a chimp-human clade in bone morphology. *Journal of Human Evolution* 88: 146–159. https://doi.org/10.1016/j.jhevol.2015.06.012

Pérez-Barbería F.J., Gordon I.J. & Pagel M. 2002. The origins of sexual dimorphism in body size in ungulates. *Evolution* 56 (6): 1276–1285. https://doi.org/10.1111/j.0014-3820.2002.tb01438.x

Perry J.S. 1954. Some observations on growth and tusk weight in male and female African elephants. *Proceedings of the Zoological Society of London* 124 (1): 97–104. https://doi.org/10.1111/j.1096-3642.1954.tb01481.x

Polly P.D., Cardini A., Davis E.B. & Steppan S.J. 2015. Marmot evolution and global change in the past 10 million years. *In*: Hautier L. & Cox P.G. (eds) *Evolution of the Rodents: Advances in Phylogeny, Functional Morphology and Development*: 246–276. Cambridge University Press, Cambridge.

R Core Team 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ralls K. 1977. Sexual dimorphism in mammals: Avian models and unanswered questions. *The American Naturalist* 111 (981): 917–938.

Rankin A.M., Schwartz R.S., Floyd C.H. & Galbreath K.E. 2019. Contrasting consequences of historical climate change for marmots at northern and temperate latitudes. *Journal of Mammalogy* 100 (2): 328–344. https://doi.org/10.1093/jmammal/gyz025

Renaud S., Chevret P. & Michaux J. 2007. Morphological vs. molecular evolution: ecology and phylogeny both shape the mandible of rodents. *Zoologica Scripta* 36 (5): 525–535. https://doi.org/10.1111/j.1463-6409.2007.00297.x

Richtsmeier J.T., Deleon V.B. & Lele S. 2002. The promise of geometric morphometrics. *American Journal of Physical Anthropology* 119 (S35): 63–91. https://doi.org/10.1002/ajpa.10174

Rohlf F.J. 1970. Adaptive hierarchical clustering schemes. *Systematic Zoology* 19 (1): 58–82. https://doi.org/10.2307/2412027

Rohlf F.J. 1993. Relative warp analysis and an example of its application to mosquito wings. *In*: Marcus L.F., Bello E. & Garcia-Valdecasas A. (eds) *Contributions to Morphometrics*: 131–159. Museo Nacional de Ciencias Naturales, Madrid 8.

Rohlf F.J. 1998. On applications of geometric morphometrics to studies of ontogeny and phylogeny. *Systematic Biology* 47 (1): 147–158. https://doi.org/10.1080/106351598261094

Rohlf F.J. 2015. The tps series of software. *Hystrix, the Italian Journal of Mammalogy* 26 (1): 9–12. https://doi.org/10.4404/hystrix-26.1-11264

Rohlf F.J. 2021. Why clusters and other patterns can seem to be found in analyses of high-dimensional data. *Evolutionary Biology* 48 (1): 1–16. https://doi.org/10.1007/s11692-020-09518-6

Rohlf F.J. & Corti M. 2000. Use of two-block partial least-squares to study covariation in shape. *Systematic Biology* 49 (4): 740–753. https://doi.org/10.1080/106351500750049806

Rohlf F.J. & Marcus L.F. 1993. A revolution morphometrics. *Trends in Ecology & Evolution* 8 (4): 129–132. https://doi.org/10.1016/0169-5347(93)90024-J

Rohlf F.J. & Slice D. 1990. Extensions of the Procrustes method for the optimal superimposition of landmarks. *Systematic Zoology* 39 (1): 40–59. https://doi.org/10.2307/2992207

Rohlf F.J., Loy A. & Corti M. 1996. Morphometric analysis of Old World Talpidae (Mammalia, Insectivora) using partial-warp scores. *Systematic Biology* 45 (3): 344–362. https://doi.org/10.1093/sysbio/45.3.344

Rolfe S., Pieper S., Porto A., Diamond K., Winchester J., Shan S., Kirveslahti H., Boyer D., Summers A. & Maga A.M. 2021. SlicerMorph: An open and extensible platform to retrieve, visualize and analyse 3D morphology. *Methods in Ecology and Evolution* 12 (10): 1816–1825. https://doi.org/10.1111/2041-210X.13669

Ruckstuhl K.E. & Neuhaus P. 2002. Sexual segregation in ungulates: a comparative test of three hypotheses. *Biological Reviews* 77 (1): 77–96. https://doi.org/10.1017/S1464793101005814

Salvidio S., Crovetto F. & Adams D.C. 2015. Potential rapid evolution of foot morphology in Italian plethodontid salamanders (*Hydromantes strinatii*) following the colonization of an artificial cave. *Journal of Evolutionary Biology* 28 (7): 1403–1409.

Schlager S. 2017. Morpho and Rvcg – Shape Analysis in R. *In*: Zheng G., Li S. & Szekely G. (eds) *Statistical Shape and Deformation Analysis*: 217–256. Academic Press.

Schutz H., Polly P.D., Krieger J.D. & Guralnick R.P. 2009. Differential sexual dimorphism: size and shape in the cranium and pelvis of grey foxes (*Urocyon*). *Biological Journal of the Linnean Society* 96 (2): 339–353. https://doi.org/10.1111/j.1095-8312.2008.01132.x

Sheets H.D. & Zelditch M.L. 2013. Studying ontogenetic trajectories using resampling methods and landmark data. *Hystrix, the Italian Journal of Mammalogy* 24 (1): 67–73. https://doi.org/10.4404/hystrix-24.1-6332

Slice D.E. 1999. Morpheus et al. *Ecology and Evolution. State University of New York, Stony Brook*. Available from https://sbmorphometrics.org/morphmet/morpheus_vienna_2006.zip [accessed 23 Apr. 2024].

Smith G.R. 1990. Homology in morphometrics and phylogenetics. *In*: *Proceedings of the Michigan Morphometrics Workshop*: 325–338. University of Michigan Museum of Zoology, Ann Arbor.

Smith O.A.M., Duncan C., Pears N., Profico A. & O'Higgins P. 2021. Growing old: Do women and men age differently? *The Anatomical Record* 304 (8): 1800–1810. https://doi.org/10.1002/ar.24584

Sneath P.H.A. & Sokal R.R. 1973. *Numerical Taxonomy. The Principles and Practice of Numerical Classification.* Freeman & Co., San Francisco.

Sokal R.R. & Rohlf F.J. 2009. *Introduction to Biostatistics Second Edition*. Dover Publications Inc., N.Y.

Solow A.R. 1990. A randomization test for misclassification probability in discriminant analysis. *Ecology* 71 (6): 2379–2382. https://doi.org/10.2307/1938650

Steppan S.J., Akhverdyan M.R., Lyapunova E.A., Fraser D.G., Vorontsov N.N., Hoffmann R.S. & Braun M.J. 1999. Molecular phylogeny of the marmots (Rodentia: Sciuridae): Tests of evolutionary and biogeographic hypotheses. *Systematic Biology* 48 (4): 715–734.
https://doi.org/10.1080/106351599259988

Steppan S.J., Kenagy G.J., Zawadzki C., Robles R., Lyapunova E.A. & Hoffmann R.S. 2011. Molecular data resolve placement of the Olympic marmot and estimate dates of trans-Beringian interchange. *Journal of Mammalogy* 92 (5): 1028–1037. https://doi.org/10.1644/10-MAMM-A-272.1

Tafani M., Cohas A., Bonenfant C., Gaillard J.-M., Lardy S. & Allainé D. 2013. Sex-specific senescence in body mass of a monogamous and monomorphic mammal: the case of Alpine marmots. *Oecologia* 172 (2): 427–436. https://doi.org/10.1007/s00442-012-2499-1

Thioulouse J., Renaud S., Dufour A.-B. & Dray S. 2021. Overcoming the spurious groups problem in between-group PCA. *Evolutionary Biology* 48 (4): 458–471.
https://doi.org/10.1007/s11692-021-09550-0

Varón-González C., Whelan S. & Klingenberg C.P. 2020. Estimating phylogenies from shape and similar multidimensional data: Why it is not reliable. *Systematic Biology* 69 (5): 863–883.
https://doi.org/10.1093/sysbio/syaa003

Velhagen W.A. & Roth V.L. 1997. Scaling of the mandible in squirrels. *Journal of Morphology* 232 (2): 107–132. https://doi.org/c6p5kd

Verhoeven K.J.F., Simonsen K.L. & McIntyre L.M. 2005. Implementing false discovery rate control: increasing your power. *Oikos* 108 (3): 643–647. https://doi.org/10.1111/j.0030-1299.2005.13727.x

Viscosi V. & Cardini A. 2011. Leaf morphology, taxonomy and geometric morphometrics: A simplified protocol for beginners. *PLoS ONE* 6 (10): e25630. https://doi.org/10.1371/journal.pone.0025630

Voje K.L., Hansen T.F., Egset C.K., Bolstad G.H. & Pélabon C. 2014. Allometric constraints and the evolution of allometry. *Evolution* 68 (3): 866–885. https://doi.org/10.1111/evo.12312

Warne R. 2019. A primer on multivariate analysis of variance (MANOVA) for behavioral scientists. *Practical Assessment, Research, and Evaluation* 19 (1): e17. https://doi.org/10.7275/sm63-7h70

Watanabe A. 2018. How many landmarks are enough to characterize shape and size variation? *PLoS ONE* 13 (6): e0198341. https://doi.org/10.1371/journal.pone.0198341

Waterman J., Wolff J. & Sherman P. 2007. Male mating strategies in rodents. *In*: Wolff J.O. & Sherman P.W. (eds) *Rodent Societies: An Ecological and Evolutionary Perspective*: 27–41. University of Chicago Press, Chicago.

Willig M.R., Owen R.D. & Colbert R.L. 1986. Assessment of morphometric variation in natural populations: The inadequacy of the univariate approach. *Systematic Biology* 35 (2): 195–203.
https://doi.org/10.1093/sysbio/35.2.195

Willmore K.E., Zelditch M.L., Young N., Ah-Seng A., Lozanoff S. & Hallgrímsson B. 2006. Canalization and developmental stability in the Brachyrrhine mouse. *Journal of Anatomy* 208 (3): 361–372.
https://doi.org/10.1111/j.1469-7580.2006.00527.x

Wilson D.E. & Reeder D.M. 2005. *Mammal Species of the World: A Taxonomic and Geographic Reference*. Johns Hopkins University Press, Baltimore.

Zachos F.E. 2016. *Species Concepts in Biology: Historical Development, Theoretical Foundations and Practical Relevance*. Springer International Publishing.

Zelditch M., Swiderski D., Sheets D. & Fink W. 2004. *Geometric Morphometrics for Biologists: A primer*. Elsevier Academic Press. Waltham, MA.

Zelditch M.L., Swiderski D.L. & Sheets H.D. 2012. *Geometric Morphometrics for Biologists: A Primer*. Elsevier Academic Press.

Zhang Z.Q. 2013. Animal biodiversity: An outline of higher-level classification and survey of taxonomic richness (Addenda 2013). *Zootaxa* 3703: 5–11. https://doi.org/10.11646/zootaxa.3703.1.3

Manuscript received: 30 July 2023
Manuscript accepted: 9 February 2024
Published on: 15 May 2024
Topic editor: Tony Robillard
Section editor: Frank Zachos
Desk editor: Kristiaan Hoedemakers

Supplementary files

Formats: mj.txt = MorphoJ (with 'id', the specimen label*, in the first column); past.txt = PAST (with its own label in the first column, which for multivariate analyses also contains the group colour code); nts = TPS Series.

*(bro_mle_MVZ_8360 is a specimen that should be renamed as cal_mle_MVZ_8360, because it is a hoary marmot, as correctly reported in the species classifiers. In the label, which was not used for any analysis, I kept the wrong abbreviation (bro) used in the original jpg image name. However, in general, it is better to have accurate, descriptive labels, as discussed in V&C - see main text).

Supp. file 1. ALL_RAW_15L_N462by2.mj.txt: this is the main data file with the raw coordinates of all 15 landmarks and 462 individuals (including possible outliers), each with its two digitizations. It is the main morphometric dataset, from which all others can be obtained. It should be used in MorphoJ for assessing ME, but also, once low precision landmarks and outliers are removed, it can be used for all main analyses with averaged individuals (*Preliminaries, Average observations by* ... using the classifier 'indiv'). https://doi.org/10.5852/ejt.2024.934.2529.11383

Supp. file 2. ALL_CLASSIFIERS.mj.txt: the file contains the following variables: subgenus, species, modern_paleo (which is relevant only for *M. vancounverensis* in order to distinguish recent and subfossil specimens), sex, indiv (an integer used as a simple individual identifier useful to recognize duplicates), side (of the mandible), OUTLIER (marks the 17 potential outliers, which were excluded from the main analyses), collection (where specimens originated), catalogue_number (in the corresponding museum), year_coll (year when the specimen was captured), Country, Province_State, Locality (with these last three variables containing the information, if available, on geographical origin of a specimen). https://doi.org/10.5852/ejt.2024.934.2529.11385

Supp. file 3. ALL_COVARIATES.mj.txt: this file can be created from the previous one, as it simply recodes a few variables using integer numbers (species, sex - with 0 for females and 1 for males - and year_coll, which was already numeric). For instance, it can be useful to test sex using regressions on dummy variables after removing outliers and unsexed individuals (in MorphoJ: Preliminaries, include or exclude observations) and splitting data by species (in MorphoJ: *Preliminaries, Subdivide dataset by* ...). HOWEVER, the species covariate cannot be used for similar purposes without being modified. For ANOVAs/MANOVAs using the regression approach, one needs a design matrix. For pairwise tests of species differences using regressions, one has first to subset the data (e.g., select VAN and woodchucks by first splitting by species and then combining these two species in MorphoJ) and probably replace the species code with that conventionally employed for dummy variables (say, VAN = 1 and woodchuck = -1). https://doi.org/10.5852/ejt.2024.934.2529.11387

Supp. files 4–8. EXAMPLE FILES FOR SPECIFIC ANALYSES (potential outliers excluded; 12 landmark configuration).

Supp. file 4. TESTING_SDM_CS_12L_N356.past.txt: this is to run the species by sex ANOVA of CS in PAST. It only includes the 356 specimens of known sex. The main variables are sp.n (coding species with an integer as required in PAST for the two-way ANOVA), sex.n (coding males as 1 and females as 2) and CS (centroid size). Other variables are described above and can be ignored. Some (species, sex and indiv) are included only as an aid to identify specimens; sp.nBigN codes species with largest samples with an integer if one wants to repeat the ANOVA after excluding small samples (drag this variable so that it replaces sp.n and then select only the rows with woodchucks, hoary and yellow-bellied marmots). https://doi.org/10.5852/ejt.2024.934.2529.11389

Supp. file 5. TESTING_species_CS_12L_N445.past.txt: this is an example of how to organize data, after pooling sexes, for most univariate analyses/plots in PAST. For instance, it can be used for a one-way ANOVA testing species differences in CS or for drawing a box-plot of CS. https://doi.org/10.5852/ejt.2024.934.2529.11391

Supp. file 6. TESTING_SDM_SH_12L_N356_sexed.nts: this is the landmark data to run the MANOVA in TPSRegr as explained in the main text of the second paper (B2). It only includes the 356 individuals of known sex. https://doi.org/10.5852/ejt.2024.934.2529.11393

Supp. file 7. TESTING_SDM_SH_12L_N356_dummy_variables_MANOVA.nts: this is the design matrix to run the MANOVA in TPSRegr as explained in the main text of the second paper (B2). It only includes the 356 individuals of known sex. https://doi.org/10.5852/ejt.2024.934.2529.11395

Supp. file 8. TESTING_SLOPES_etc_STATIC_ALLOMETRY_12L_N445.nts: this is the landmark data to run the MANCOVA in TPSRegr as explained in the main text of the second paper (B6). https://doi.org/10.5852/ejt.2024.934.2529.11397

Supp. file 9. TESTING_SLOPES_etc_STATIC_ALLOMETRY_12L_N445_dummy_variables_MANCOVA.nts: this is the design matrix to run the MANCOVA in TPSRegr as explained in the main text of the second paper (B6). https://doi.org/10.5852/ejt.2024.934.2529.11399

Appendix B

Checklist of the main steps

This is a non-exhaustive checklist to help beginners not to miss important steps in their GMM analysis. It should be used in combination with Fig. 1 of part A.

1. Did you draft a series of clear, well defined study questions?

2. Following the hourglass model of scientific research, the study questions conclude the introduction that starts with the broader reasons to perform the research, goes on with a brief review of the relevant literature and ends with the specific details of the current study. In relation to the research questions, I recommend to check that the study structure rigorously follows the same order as the main questions in all sections (i.e., methods for hypothesis 1 explained before those for hypothesis 2 etc.; same reasoning for the results and discussion, with the discussion starting from the specific results of your study and ending with broader implications, generalizations and future directions).

3. Did you do a detailed search in the literature for studies related to your main study questions?

4. Is GMM a good option for your research or could you get more accurate answers with other approaches?

5. Have you selected a life stage or age group for the taxonomic comparison? Do you know if sexual dimorphism is important in the study group?

6. Have you considered whether enough specimens may be available? The literature might help guessing an approximate number, but a prospective power analysis is also an option.

7. Have you gathered enough information on the study groups, including their distribution range, to be sure that not only sample size is adequate but also that the sample is representative of the populations being studied. In this respect, if information on the locality of origin is available, plotting the specimens on a map, and comparing it to published maps of the distribution range for the study taxon, may be of help to spot gaps as well as clusters of observations.

8. Did you consider carefully what the most informative anatomical structure is for your taxonomic study? Practical issues may also be important: the study structure should be easy to obtain and measure using a standardized protocol. For instance, for tetrapods, post-cranial material is usually less abundant in museum collections; skulls may be more common (and important, if used to describe type specimens); mandibles are less informative than crania, but may be a good compromise for a preliminary 2D analysis etc.

9. On the selected study structure, did you consider what the most informative landmarks are? Here too one might have some practical considerations (certain landmarks may be potentially useful, but hard to see or not available in all groups; others may be on parts that are easily broken or missing; too many landmarks might lead to problematic p / N ratios etc.).

10. Have you tried a small pilot study to better understand possible issues with both the data collection protocol and the analysis?

11. Typically, you should be using both size and shape in taxonomy: if you focus only on shape, is this justified? Bear in mind that most of the time it is not, and size is as important as shape and, in fact, usually easier to analyse and interpret.

12. Have you tried to have replicates (at least duplicates) of each step of the data collection in a representative subsample, if not in all individuals, to assess measurement error?

13. After collecting data, and making backup copies on portable hard-drives or the cloud, have you checked evident errors (wrongly labelled or duplicated specimens; inconsistent scale factors or unit of measurements – e.g., inches or cm etc.; issues with using semicolon or commas in spreadsheets, depending on the language of the operating system; missing information – e.g., sex, locality, year of collection etc.).

14. Having digitized the data, including possible replicates, did you check if obvious outliers are evident?

15. After removing obvious outliers or correcting errors (e.g., swapped landmarks; missing or wrong scale factor etc.), have you assessed measurement error at least in a subsample?

16. Did you consider a 'time-lag' effect that may introduce a bias in the measurements, if collected at multiple times with long interruptions (months or years)? Remember that one can mitigate against this risk by having a representative sample on which to re-train herself/himself to improve precision and which is re-measured any time the data collection is restarted, after a long interruption, to check for differences.

17. Biases are also likely when data are collected by multiple operators and/or using different instruments or techniques: have you thought about this and included this likely important source of error in the assessment of measurement error?

18. In the assessment of measurement error, as well as in all other analyses, have you considered common assumptions such as independent observations, homoscedasticity and, if required, normality? Have you at least explored these assumptions in the largest groups?

19. If very low precision landmarks are found, that might impact results or just make data noisier, have you reassessed measurement error and outliers after excluding those points?

20. With a cleaned dataset, the main analysis starts: have you considered what graphical techniques to use for complementing the numerical analyses? In the next points, I focus on the latter, but give for granted that they are always coupled with graphical approaches. Plots are as important as numerical tests and, with small samples, may in fact be the only option to at least very preliminarily explore the data.

21. In gonochoric species, the assessment of sexual dimorphism is a fundamental first step: is it large relative to interspecific differences? Does it vary across taxa (i.e., there is a significant interaction between taxon and species)? I stress that I am assuming that age-related variability is controlled by selecting a homogeneous age group (e.g., adults).

22. Having provided sound evidence for the decision of pooling sexes or using separate sex analyses (or, even if less desirable, sex-correct the data), the taxonomic comparison can start. If the measurement error ANOVA and/or the taxon by sex ANOVA have already demonstrated overall significant and large differences among study taxa, did you explore specifically which pair of taxa differ and how much? This is the series of pairwise tests of mean size and shape differences, where estimates of effect size (Rsq) are, as in all other analyses, as important as P values. In this, and similar cases of multiple tests of a specific hypothesis, bear in mind that type I error rates could be inflated.

23. To complement tests, have you explored group separation in ordinations and phenograms, as well as computing cross-validated classification accuracy?

24. With DA/CVAs and similar techniques, did you consider typicality probabilities?

25. If differences are large, have you summarized mean differences and, for shape, visualized them using shape diagrams?

26. Have you explored if it is worth/possible to assess, and maybe control for, allometry in shape comparisons? In doing this, bear in mind the assumption of parallel allometries.

27. Is it interesting to also compare the magnitude of size and shape variance?

28. Have you checked the sensitivity of results to the inclusion of small samples?

29. Have you checked the sensitivity of results to strongly unbalanced sample sizes? The question before this one might already provide a first answer. Sometimes, it may be interesting to replicate the main analyses using balanced (with N equal to N in the smallest sample) randomized subsamples.

30. If p/N is large, can this affect your results? For instance, in ANOVAs and CVA/DAs, have you explored if the inclusion of a different number of PCs appreciably changes results?

31. Finally, have you carefully revised all analyses, assessed (and acknowledged) potential limitations, and stressed which findings are more robust and which ones, in contrast, are uncertain and preliminary?