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"

Andrea CARDINI[1-2],

[1]Dipartimento di Scienze Chimiche e Geologiche, Università di Modena e Reggio Emilia, Via Campi, 103 - 41125 Modena - Italy

[2]School of Anatomy, Physiology and Human Biology, The University of Western Australia, 35 Stirling Highway, Crawley WA 6009, Australia

E-mail address: alcardini@gmail.com, andrea.cardini@unimore.it

Tel. 0039 059 2058472

https://orcid.org/0000-0003-2910-632X

Abstract

The study of modularity and integration using Procrustes geometric morphometrics has become a prominent approach in evolutionary developmental biology. A most popular method is the 'within a configuration' approach, often used in combination with 'high density' morphometric data (i.e., large numbers of landmarks and semilandmarks). In 2019, I realized that this approach violates a basic assumption of shape analysis using superimposition methods and showed that this violation may increase the rate of false positives, beyond the nominal value, in statistical tests. A very recent study, however, argues that simulations in four different datasets indicate that the theoretical violation has a mostly negligible effect on tests, so that it "is not an obstacle to analyses of integration and modularity" (p. 167, Zelditch and Swiderski, 2023). Its authors also claim that I mischaracterized the methods, overstated the problem and published non-reproducible results. In this paper, I carefully compare their statements and mine, as well as our results, and demonstrate that: 1) the problem is always present; 2) the authors overlook the importance of statistical power and p/N ratios (p = number of variables; N = sample size); 3) results are case-specific but, in fact, perfectly congruent in the single dataset in common between both studies; 4) the impact is especially concerning, based on both mine and their recent findings, precisely in the 'high-density' morphometric analyses, claimed to be the state-of-the-art in the field; 5) unlike the recent study, that claims external validity and generalizes from a few cases, I explicitly stated multiple times that my findings were specific to the datasets I analysed, the parameters I used and the tests I explored. If confirmed by future research, the recent findings in fact fully corroborate my original suggestions that, despite the undeniable theoretical issue, its impact may vary, its assessment is complex and generalizations are unlikely to be easy. However, both my preliminary work and all their simulations, using semilandmarks slid according to the minimum bending energy criterion, strongly indicate that the most serious problems might affect precisely this highly advertised approach to the analysis of modularity and integration. Given its popularity, it is likely that dozens of studies in high impact journals have published results that may be little more than methodological artefacts.

Keywords

bending energy – overstatement – shape analysis – superimposition – type I error rate

Introduction

Zelditch and Swiderski (2023, henceforth abbreviated as ZS23) recently published in Evolutionary Biology a study on the effect of the Procrustes superimposition (including the possible sliding of semilandmarks) on morphometric analyses of modularity and integration at both a micro- and macro-evolutionary level. If my understanding of their design is correct, they use four different datasets to explore whether the superimposition produces results suggesting spurious modularity (assessed with the CR index (Adams 2016)) or integration (assessed with a PLS analysis (Rohlf and Corti 2000)). Thus, they add isotropic variation to a configuration obtained from real data, in which (i.e., in the real data) they also perform the same analyses. With these datasets, they test if there is modularity or integration before and after the superimposition and compare the results with those of the empirical, real, dataset. This is repeated many times for each dataset. They, then, argue that, even if there was a theoretical problem with splitting superimposed landmark data in subsets, there would be no analytical impact on findings (i.e., results are accurate) if:

a) spurious results, after the superimposition, appear about as often as in the raw data, which means that any evidence of modularity/integration occurs because, by chance, it was already present in the simulated data;

b) the range of spurious results does not include the empirically observed estimate of modularity or integration and, thus, the 'true' signal in the real data overcomes any spurious pattern introduced by the superimposition.

I clarify that, in this paper, I use the term 'superimposition' regardless of whether semilandmarks are slid or not, when present. If there is any specific comment that concerns 'sliding', however, I will explicitly mention that I am talking about findings with semilandmarks slid using a specific method (either by minimizing Procrustes distances – minPRD – or bending energy – minBEN, the two types of sliding used by ZS23).

I welcome this study, that follows my invitation for the morphometric community to explore the issue I found in 2019 (Cardini 2019, henceforth abbreviated C19). The problem concerns a bias, variable from case to case in degree and direction, in several of the most common tests of modularity and integration using the 'within a configuration' approach in Procrustes shape data. 'Within a configuration", in this context, means that landmarks and semilandmarks used to measure an organ or an organism are superimposed all together once, but, later, the resulting shape data are split into subsets ('modules'). The alternative approach, not considered in ZS23 but assessed in C19, re-superimposes, one at a time, the 'modules' and is generally referred to as 'separate superimpositions'.

Unfortunately, ZS23 provide a rather immodest and misleading conclusion (with the most relevant sentences emphasized in italics throughout my paper). They state that (p. 167), even if "all results of geometric morphometric analyses [using Procrustes methods] are invariant to rotation of the data" in the tangent space, which is "why subsets of data cannot be analysed separately", "*that definition of shape* makes shape data

'special', but that *is not an obstacle to analyses of integration and modularity*[1]". Thus, they argue (in the Abstract, p. 147) that "neither evolutionary modularity nor integration are significantly strengthened by superimposition followed by sliding" and, therefore, despite some problems they found with inaccurate results in specific cases (see below), they reassure users that the effect of the superimposition is slight on tests of modularity and integration in empirical datasets. This means that either there is no issue or, when there is, it is minor and somewhat exceptional. Nothing to worry much about, for ZS23.

ZS23 portray my previous work with inaccuracies, as well as with a tone of suspicion, as if I had been hiding something (e.g., p. 148, "We first revisit one example presented in Cardini, 2019 ... to report results that appear to be omitted from that work" or p. 166, "The relevant results are briefly and, we suspect, incompletely summarized in the text"[2]). It seems that I cherry picked findings and omitted results, that would disprove or weaken my conclusions. Regardless of what I really did, scientists are right to raise doubts and those in ZS23 are legitimate questions. I have never received a request for clarification by the authors and was not asked to review their paper, but these are options. In fact, on contentious issues, it might be better to have fully independent referees. Thus, as I am criticized in a published paper and, more importantly, because I strongly disagree with their conclusions, I am happy to answer in a publication. As I comment on ZS23, I take the chance to show that they have, in fact, corroborated with strong evidence a suggestion I cautiously made: one specific approach, considered by some (Goswami et al. 2019) the most effective for morphometric studies of modularity and integration, and employed in dozens of publications in high impact journals, is likely to be so seriously flawed, that all those results must be reassessed and the papers accordingly corrected or simply retracted.

[1]However, they also say (p. 148): "Nor do the methods [for testing modularity or integration] violate any fundamental assumption of Procrustes-based methods, which are that shapes are configurations of landmarks, invariant under the Euclidean transformations of translation, scaling, and rotation". Thus, first they say that these methods, which, using the within a configuration approach, separate landmarks after a common superimposition, do not violate any fundamental assumption. Yet, later, they seem to acknowledge that the superimposition requires that data are not analysed as subsets, but conclude that in practice (i.e., based on results of their simulations – i.e., p. 149 "the last, and most important aim of this study ... to determine whether covariances induced by the Procrustes superimposition are likely to compromise empirical studies") this is no obstacle at all for tests of modularity/integration. I will argue later why their conclusion is a misleading overstatement. For now, I just underscore a contradiction: they argue that I had mischaracterised the problem with those tests, when I wrote (C19) that they violate a fundamental assumption of shape analyses based on superimpositions; later, however, they acknowledge that we should not split superimposed landmarks in subsets. It is one or the other, right? Thus, regardless of whether the violation has an impact on results (which is a different issue), I restate that there is a violation and this is the same type of mistake one does with per-landmark variance, the interpretation of displacement vectors as landmarks moving in a given direction etc. (Cardini and Verderame 2022).

[2]Also, at p. 165: "These should be fully reproducible because the data and the hypothesis of modularity/integration are supplied by the geomorph package, and despite the lack of details regarding the simulations done in Cardini, 2019, both studies simulated isotropic variation (and presumably both used the variance of the landmarks in the data), both superimposed the data without sliding the semilandmarks, and both simulated populations of N = 100 and N = 500. Those results should be identical. Additionally, analyses of rodent mandibles fit to the simple, two-module model, ought to be at least comparable". Later, in the same page: "Unfortunately, none of the potentially comparable results are presented in the tables, which instead highlight the results for the simple geometric shapes (circle, polygon, hexagon) and the rodent mandible at N = 1000 (Table 2, Cardini, 2019), or those same geometric shapes and the rodent mandible at N = 500 (Table 3, Cardini, 2019".

Comments on ZS23

Raw vs superimposed data

I open my comments on ZS23 with a question, that is secondary in the context of this paper, but I hope it is explored by statisticians: is testing for modularity or integration in the raw data an appropriate null model to assess the impact of the superimposition on those tests? I cannot say and take Z&S23 results as genuine. The question is important[3], but the answer irrelevant for my specific criticisms of ZS23 misleading claims.

No concern for p/N and the 'rediscovery' of the impact of N on statistical power

ZS23 simulate N regardless of p (the number of variables). They claim smaller N (ca. 100) is realistic and consider samples ≥ 500 as "enormous" (p. 148). For microevolutionary studies, they seem unconcerned about the relationship between N and the size of the population being studied. But, is N = 100 a representative sample for both, say, the Vancouver Island marmot, whose surviving wild population on a few mountain tops in the Vancouver Island numbers between 100-200 individuals, and modern humans, with some 8 billion people distributed over the entire planet? Large samples not only increase power, when differences are small, but also make estimates of statistical parameters (means, variance and covariances) more accurate (Cardini et al. 2021).

[3]This issue has already been hinted at. For instance, Klingenberg wrote (Klingenberg 2021): "If the goal of the simulation is to extract shape variation, discrepancies between the simulated data and the Procrustes-superimposed configurations are not a problem. Instead, the investigator might want to consider using a different approach for the simulation, which provides pure shape variation without an additional non-shape component". ZS23 cite this same paper to argue (p. 167) that "those non-shape components do not contribute to the "true" covariance structure of shape, hence removing them does not introduce inaccuracies. The question is not whether superimposition induces covariances because it must, but rather, it is whether the methods used to analyze modularity and integration are misled to infer biologically-induced modularity and integration when superimposition, by itself, explains them". On this point, I agree with ZS23 (and Klingenberg), but ZS23 might have missed that, by comparing integration/modularity in the raw data with the same test in the superimposed shape data, they are precisely doing what Klingenberg argues is wrong: ZS23 are comparing modularity/integration *tested* in non-shape, raw data, with that found in superimposed shape data. I also used simulated isotropic data to demonstrate that the covariance introduced by the superimposition can bias the results of tests using the within a configuration approach, but did not test modularity/integration in the raw data (C19). Later I also stressed again (Cardini 2020) that the impact of the bias might vary and cannot be simply inferred using an isotropic model. However, in my analysis (C19), I did compare shape with shape, as the tests I performed were both on superimposed data, either using a common superimposition (within a configuration approach) or with separate superimpositions in each module.

ZS23 are also unconcerned about the recent attention in the morphometric community to a well known problem, that some (possibly many) statistical analyses might be negatively affected by large p/N ratios. For instance, patterns in the data might appear when there is none. This has been shown in simple principal component analyses (Björklund 2019; Bookstein 2017) and ordinations aimed at group separation (Cardini et al. 2019; Rohlf 2021), but Bookstein (2017) argued that the issue is broader. For instance, he wrote that (p. 532-533) "since PLS is a variant of principal components analysis, then the critique of PCA under conditions of 'too many uninformative variables' might be expected to apply in this setting as well … The pathology arises ... from ... the severe bias in the first eigenvalues of a covariance structure based on low-information high-p/N data". The PLS is the method of choice of ZS23 for testing integration. A quick look at p/N in some of ZS23's datasets suggests that high p/N ratios are common in their study: for instance, in the pupfish dataset, there are 56 2D points (landmarks and semilandmarks), which means about a dozen more variables than individuals when N = 100; in the squirrel hemi-mandible dataset, there are 96 points and, thus, almost twice as many variables as individuals with N = 100. Although I declare myself agnostic about the impact of p/N on ZS23 results and claims, it is an important aspect that should not be overlooked.

Indeed, a reviewer of the first version of C19 believed that (quoting from the original review) "the conclusions in this paper [C19] are not general, because they are based on point-estimate datasets [the first version of the paper, which only contained many examples – Table 2 of C19 – but no simulations] and not a formal type I error simulation: the latter of which properly addresses the question of performance. *When a proper type I error simulation is performed, there is no difference in statistical performance between separate vs. single GPA datasets* (though there will be differences in performance [type II error] when covariation is present). *Thus, the main conclusion of the paper is not correct.* Computer code is found below several specific comments". Using this code and the pupfish dataset, which is also the first study case of ZS23, with its 112 shape coordinates and N = 100, the reviewer demonstrated a correct type I error rate (ca. 0.05) in within a configuration tests of modularity/integration. Based on this simulation, he/she stated that *"the conclusions in this paper [C19, first version] are not general … [and] not correct"*. It is that code that allowed a completely unskilled R user like me to simulate type I errors and demonstrate that the issue, in fact, *is general* but easily missed, using isotropic variation, if N is not adequate in relation to p. While ZS23 argue that I found problems by using, in their view, unrealistic and suspiciously large sample sizes, they seem to have surprisingly missed that:
a) I made clear that N in the simulations was chosen in relation to p, to avoid having more variables than individuals (C19, p. 97): in "the circles and mandibles with unslid semilandmarks, … [N] was increased to 500 to take into account the much larger number of variables".

b) I also *explicitly* stated that (p. 98): “Also, as *an important cautionary note* in interpreting significant tests in this study, readers should bear in mind that simulated *samples were large in relation to the number of variables*, thus increasing power. Having *large samples is desirable but it might have overemphasized the importance of small covariances introduced by the superimposition*: in true biological data, where real covariance is expected and might be much larger than that due to the superimposition, the problem of spurious results might be less concerning”. To make this point more tangible, in the same page I added: when “run using N = 100 (instead of 500, as in Table 3), the proportion of false positives using the within a configuration approach would decrease to 0.07–0.11 in all tests except the mandible CR (that would be large − 0.53— but, nevertheless, almost half than found using N = 500)”.

I let the readers decide if the quotes above, from C19, suggest an intention to mislead by overstating the impact of the problem. Now, however, let me go back to *the first question* on within a configuration modularity/integration tests in Procrustes shape data: *do we have a general issue* or did I (C19) used ad-hoc examples of something that, as a paper put it (Goswami et al. 2019), is “Not-Really-a-Problem”? ZS23 (p. 167) definitely support the claim by Goswami and colleagues by also concluding that the problem is “not an obstacle to analyses of modularity and integration”. However, *splitting* landmarks/semilandmarks *after a common superimposition is problematic*, because of the biological arbitrariness of all currently employed superimposition methods (not just Procrustes). This arbitrariness prevents meaningful analyses and interpretations of subsets of points in a configuration, something known (Cardini & Verderame, 2022, and references therein) since before the GMM ‘revolution’ (Rohlf and Marcus 1993). Contrary to ZS23’s claim (p. 148) that “the methods for evaluating those hypotheses [of modularity]” do not violate this fundamental assumption of Procrustes shape analysis in biology, it is precisely in testing for modularity that one creates subsets of landmarks, adjusted by Procrustes to become correlated, even if originally they were not. It is, thus, the within a configuration tests that are problematic and not, of course, Procrustes in itself. Thus, with the within a configuration approach to modularity/integration, the problem is there all the time, including in ZS23, as long as statistical power is adequate to detect it. This is why one has to explore what happens in his/her data in relation p/N.

Whether the issue has a practical impact on results of modularity/integration tests is, in contrast, a separate question. It is a very relevant question, but one which is harder to answer, because internal validity (findings specific to a dataset – mandibles, fish, shells etc. with their putative modules – and the specific mathematical treatment of the data, e.g., if semilandmarks are slid or not and what sliding method is used) does not equate to external validity (i.e., whether those specific findings are generalizable). It is this second question that ZS23 contribute to answer. Yet, all they can soundly claim is internal validity, specific to their examples and the settings of their simulations. The question of the practical impact, however, is not really new, as I had already stated (C19, p. 102, as well as the quote, above, from p. 98): "It is also not impossible that this effect, as well as the more general ones on variance–covariance patterns ..., *may have been overemphasized in large simulated datasets purely made of isotropic variation. Yet, one cannot be sure, and indeed the impact of these issues may vary from case to case and generalizations might be difficult to make*". Regretfully, whereas the authors of ZS23 are unconcerned about p/N but ready to point out that I might have misled morphometricians by using (in their view and wording, p. 148) "enormous" samples, their 'discovery' of the impact of N on the magnitude of the bias had been acknowledged in full already in C19 and, in fact, no less than twice in the article.

Why pupfish results are not identical: what I hid … or ZS23 overlooked

ZS23 simulation using the geomorph pupfish dataset is said to follow C19. In fact, in that simulation, as plainly stated in my paper, I just recycled an R script a reviewer kindly made available, a script I did not publish online, for confidentiality, but wrote that was available upon request. Nobody asked me, however. The settings used for 'my' simulation of the pupfish (and other geomorph datasets) are those provided by the reviewer with a single change: he/she found an appropriate rate of type I errors using N = 100; I increased sample size and, when N was just a few times larger than p, the rate of type I errors in within a configuration modularity and integration tests became many times larger than the nominal 0.05.

As ZS23 complain about the poor reproducibility of my study in relation the pupfish dataset, they could have simply asked me the script to be sure they were doing precisely the same analyses. Or, in relation to the results I briefly reported for those tests, they could have dropped me an email for clarification and more specific information. Unfortunately, they did not contact me, but claimed (p. 165): "… two analyses ... should be fully reproducible: modularity and integration of pupfish body/opercle shape. These should be fully reproducible because the data and the hypothesis of modularity/integration are supplied by the geomorph package, and despite the lack of details regarding the simulations done in Cardini, 2019[4], both studies simulated isotropic variation (and presumably both used the variance of the landmarks in the data), both superimposed the data without sliding the semilandmarks, and both simulated populations of N = 100 and N = 500. Those results should be identical". To be accurate, results may not be identical unless: a) ZS23 are sure we used the same exact settings, which only required asking the R script of C19, which they did not do; b) the same geomorph version is used as in C19, which can be downloaded from the R depository. I confess I am guilty of not writing down the version of geomorph I used, but it is likely to be either https://cran.r-project.org/src/contrib/Archive/geomorph/geomorph_3.0.6.tar.gz or the one published just before or after this one. This is to be fully rigorous with reproducibility, although I suspect that our results are in fact very similar, differences in settings might be minor and probably updates in geomorph have not changed the specific functions I used.

ZS23 also accuse me (p. 165, for instance) of highlighting results of simple geometric shapes (circle, polygon, hexagon) and the marmot mandible, as if (as with the "enormous" N) I cherry picked those to prove my point. Indeed, they are right that the main focus was on those datasets, instead of the pupfish or the other datasets in geomorph, that are not shown in the tables. Why? Well, just because simple geometric figures and marmot mandibles were the study data I had used, since the very first version of the paper (the one without simulations) to explore the effect of using separate superimpositions compared to the within a configuration method. After I was gifted (first revision) by the reviewer with a script I could use to simulate type I error rates, I added, to the original point estimates from *my data* (Table 2 of C19), the simulations showing an unacceptably high rate of false positives (Table 3). The geomorph example datasets were mentioned briefly in the discussion, but were not part of my study. They were the data used by the reviewer to disprove my findings. I simply clarified that, even with those data, the problem was present as long as statistical power was adequate. Nothing to hide: the script, as mentioned, has always been available, since the publication of the paper (https://link.springer.com/article/10.1007/s11692-018-9463-x#Sec11): "*The R-script with the simulation by the anonymous reviewer is available upon request*"), for anyone who want to replicate the simulations on my data (also available as online supplementary material), any of the geomorph datasets or one's own personal data.

[4]From the SI of C19ZS (https://link.springer.com/article/10.1007/s11692-018-9463-x#Sec11): "Electronic supplementary material: *All simulated datasets are available online* as Supplementary Material. Files use the NTS format, commonly employed in morphmetrics, easy to import in MorphoJ or R, and described in details in the help manuals of the TPS Series. *The R-script with the simulation by the anonymous reviewer is available upon request*; the flaw in the original version is corrected with a single change (to be done for all example datasets): increasing sample size from 100 to 500 in order to avoid most unfavourable p/N ratios and thus low power."

ZS23 also state (p. 165) that "our results do not support the claims that the vast majority of tests for integration suggest a strong covariation between modules, that the frequencies of statistically significant integration in the simulated data range from 27%—100% and that 80%—100% of tests for modularity yield statistically significant results (Cardini, 2019)". The results I show are those from *my data*[5] analysed with the parameters stated in C19. The *geomorph datasets* are mentioned in a few lines of text in the discussion (a subsection called "Interpretations of Main Patterns", *after* the main results), *because they are not my study datasets*. They answer a reviewer's question. ZS23 are correct that I reported only briefly the range of PLS results and those are virtually identical to theirs in pupfishes. I selected the PLS because, unlike CR[6], one can perform the this test using both the within a configuration and separate superimposition approaches, and this seems enough to report for an example, that *does not belong to my main results*.

[5]From p. 96: "The *vast majority of these ten sets* of data, when analysed using the within a configuration approach, produced congruent results indicating a serious issue with type I errors". From p. 97: "*this small set of simple simulations* strongly supported the outcome of the point-estimates study: the separate blocks approach seems appropriate in terms of type I error rates, whereas the within a configuration analyses largely inflates the occurrence of false positives".

[6]In fact, Adams wrote me after this preprint was published (pers. comm. on July 6th 2023) that geomorph CR can be computed also using separate superimposition. This is not mentioned in the help file of the package and I suspect the option is not used by most morphometricians. When I wrote the 2019 paper, having found nothing in the help of geomorph and having 100% identical results using the referee's script, when CR was tested (according to the referee) using the within a configuration or separate superimposition method, I believed results for CR with separate superimpositions cannot be trusted. I am still most sceptical about those results, as far as that specific script is concerned: with separate superimpositions, CS is re-standardized within modules and the Procrustes shape coordinates are recomputed; shape data are therefore usually remarkably different from those of the within a configuration method (which is why, for instance, with the PLS they produce an appropriate type I error rate, unlike using the within a configuration method) and most unlikely to produce perfectly identical percentages of significant tests in the simulations. This suggests an error somewhere in the script. It will be interesting to implement the correct CR test with separate superimpositions and assess the type I error rates. It also means that some of the papers I quote in this preprint using CR and minBEN may have been using the separate superimposition approach, something I will carefully consider in the revision of the paper. Unfortunately, most GMM studies using CR do not explicitly mention which approach they are using. The within a configuration is the most likely, because it was advocated by Adams himself, who developed CR, as well as by Goswami et al. (2019) in their study claiming no problem with semilandmarks and modularity/integration tests in Procrustes shape data. Whenever this approach is used, however, the theoretical issue is there and its practical impact can be more or less important (probably very important using minBEN, as from the simulations of ZS23). On other issues one might have, using separate superimpositions, I refer readers to Cardini (2019).

ZS23 might have been confused because they seem to believe that is normal to make generalizations from a few examples (i.e., internal validity = external validity), but all I stated explicitly concerned just the results of my own few study cases. Unlike ZS23, however, who claim universal findings from a few examples, I concluded by stressing both the importance of the results, but also the preliminary nature of my work and the need of more research (Abstract of C19[7]): "*The study*, although *preliminary and exploratory in nature*, raises an important issue and indicates an avenue for future research. It also *suggests* that great caution should be exercised in the application and interpretation of findings from analyses of modularity and integration using Procrustes shape data, and that issues *might* be even more serious using some of the most common methods for handling the increasing popular semilandmark data used to analyse 2D outlines and 3D surfaces". It is curious that, while ZS23 misread and misreported the conclusions of my paper, they missed my cautious prediction, to which their analysis lends strong support: minBEN slid semilandmarks are likely to produce more serious artefacts in tests of modularity/integration.

Finally, in relation to how ZS23 describe my work, there is another inaccuracy: they hardly seem to mention that all my analyses (except for CR, as geomorph did not allow it[8]) had also been run using the separate superimposition method. Thus, with all the limitations of the isotropic model, but consistently with Klingenberg's (2021) suggestion, I compared tests of shape[9] with tests of shape, unlike ZS23, who compare tests of raw data – including non-shape parameters – with tests of shape. The separate superimposition approach has its own potential issues (briefly, but clearly mentioned in C19, but not an aim of that study). However, in terms of type I error rates, separate superimpositions did not produce any of the hugely inflated rates found with the within a configuration approach. This is also another demonstration that spurious significance, when it happens, is not just because of high power ("enormous samples", for ZS23): it is the within a configuration method that has a problem! Yey, I emphasize now, as in C19, that *I am not advocating for one or the other approach* and also suggest to read my warning (C19) on the possible 'cons' of separate superimpositions.

Missing in action: do ZS23 report results for minBEN slid data differently than for other data?

Table 3-4 of ZS23 show results of their simulations for the micro-evolutionary modularity/integration analysis of

[7]To be precise, I stressed multiple times that my work was just the first step, that it was exploratory, concerned a limited number of cases and parameters, that it could overemphasize the effect (although the effect is always present!) and required future studies. Besides other quotes already reported or mentioned below: p. 97-98: "Before discussing the main results, it is important to stress that the datasets and types of tests used in this paper were not aimed at thoroughly assessing the statistical properties of the methods. They are examples, analysed with common methods and used to explore whether there is a problem. If indeed they suggest a potential issue, that will require extensive studies to assess its importance and generality using simulations and a large number of different scenarios (e.g., landmark number and density, modules with large differences in number of landmarks, three or more modules, different proportions of landmarks and semilandmarks, a variety of sample sizes, different amounts of 'real' covariance etc.). It is also useful to emphasize that P values were not corrected for multiple testing and, more importantly, that they are used in this context mainly as a numerical aid to better appreciate possible misleading results …"; p. 102: "Bearing in mind its exploratory nature, the limited number of tests examined and a main focus on type I errors, and therefore the need of further in depth research considering a large variety of scenarios as well as type II errors, three main messages can be taken from this study, that might hopefully stimulate future investigations, as well as recommend a degree of caution in applications of Procrustes methods".

[8]See footnote 6 on this.

[9]Strictly speaking, although modules are obtained from superimposed data, with the within a configuration approach they no longer conform to the definition of shape in geometric morphometrics, because size and position are no longer standardized.

the squirrel mandible. In this dataset, semilandmarks are either not slid or slid using both minPRD and minBEN. Spurious results are found rather often, despite (if I am correct) a small N (97) in relation to p (≥ 96, without sliding). The percentage of spuriously significant results, however, varies depending on the parameters in the simulations (from about the same[10] to 50 or more times larger than in raw data). Yet, ZS23 reassuringly show that, despite spurious significance, the test statistics most of the time do not differ appreciably between raw and superimposed data. But something seems to be missing. The data most strongly impacted by the superimposition are those where semilandmarks are slid using minBEN, with most simulations producing 100% significant results in both CR and PLS tests. These percentages are reported at the top of both tables (section A). Unfortunately, the percentage of simulations where the test statistics differ significantly between the raw and superimposed data are shown for all types of superimposed data (section B) except minBEN. This asymmetry in the presentation of the results is puzzling. Why does it happen? If for minBEN the test statistics tended to be significantly different most of the time from those of the raw data (I am not sure about this and can only guess), one has a misleadingly optimistic impression by looking at the table: he/she sees a big problem with spurious significance of minBEN data (section A), but is not sure of its impact on effect size (section B), as the minBEN column is, sadly, absent.

It is, nonetheless, good that ZS23 at least call some of these findings troubling, but it is done concisely and no explicit justification is provided for omitting the relevant columns in Tables 3B-4B. Even if, and probably especially if, 100% of the times minBEN produces significantly different estimates compared to raw data, those columns are informative, because they bring attention to both sides of the problem: spurious significance and larger effect sizes. Showing only the columns for the superimpositions, where effect sizes rarely differ, is misleading.

Glass half full or half empty?

ZS23 main conclusion is that (from their Abstract), despite some "alarming results" such as "the extremely high frequency of significant integration in large samples (N = 500)", "the effect of superimposition is slight compared to the strength of variational or evolutionary modularity and integration found in empirical cases". This is the basis of their statement that (p. 167) tests using the within a configuration approach are "not an obstacle to analyses of integration and modularity". Thus, for ZS23 the glass is definitely much more than half full. But is it?

What we know for sure is that a certain amount of bias is always there with this approach. We have a theory that explains why and all analyses (theirs and mine) show the problem, despite differences in degree. This is a sound result. Then, there is the separate question of the practical effect on the results of the tests. Here, ZS23 claim it is almost always negligible, unless (mainly for the PLS) N is very large. I cannot say whether they are right or wrong, but believe their conclusion is at best premature:

[10]In fact, there is a curious outcome using minPRD slid semilandmarks and CR: about half of the times, minPRD slid data produce no significant modularity, despite raw data showing 2-5 cases of significant modularity. It seems, thus, that in rare cases the minPRD sliding makes the original modularity disappear.

1) Integration in large samples in not the only alarmingly spurious result. Sliding using minBEN has a massive effect in ZS23 on tests both at micro- and macro-evolutionary level, and this concerns not only integration (PLS) but also modularity (CR). Is this not an obstacle? Obviously it was not for all those (see next section) who have performed analyses of modularity/integration using minBEN slid semilandmarks, and they are many, despite a cautious but clear warning on this already present in C19. Unfortunately, ZS23 dismissively mention the huge effect of minBEN sliding, as if nobody would be so naive to fall into the trap (p. 164): "Sliding to minimize bending energy (BE) almost invariably yields highly significant modularity, but spatial covariances ought to be expected from that procedure".

2) If the bias really matters only with very large N, does that mean we are OK with samples with more variables than specimens? ZS23 mention neither the p/N issue, about which we need to learn more and, thus, should be cautious about, nor the importance of power and accurate estimates from samples.

3) What are the bases to generalize their conclusions to the universe of all possible GMM datasets in biology? The problem is complex and, indeed, ZS23 report a range of effects, from negligible to very serious, with variability depending on the configuration, number of points, treatment of semilandmarks, number of modules, and sample size, and this complexity emerges despite they simulated only four example datasets with a specific degree of variation in each.

What are the consequences for past and future studies if minBEN slid semilandmarks are as bad as ZS23 found for within a configuration tests of modularity/integration?

On June 12th 2023, a Google Scholar search for “Procrustes "bending energy" sliding modularity CR ratio” produced 172 results; if the same search is done replacing “CR ratio” with “PLS”, the number of entries is 261. This may or may not be studies using minBEN slid semilandmarks for testing modularity/integration with the within a configuration method. Some are, others are not, and there may be several which are but were not found. A quick search in Scholar is just a shortcut that helps to start appreciating how common this approach might be. minBEN sliding has often been presented by the Viennese School of morphometrics as the method of choice for treating this type of anatomical points ((Gunz et al. 2005), p. 25): “semilandmarks like these [i.e., slid using the minimum bending energy criterion] can then be treated as homologous, without artefact”; and also ((Gunz and Mitteroecker 2013), p. 107): “for larger shape variation and more extensive sliding, minimizing bending energy usually leads to better results that are in line with our notion of biological homology”. I do not comment on pure mathematics (bending energy) borrowed from the mechanics of thin metal plates creating homology in a biological context, as I have already done it elsewhere (Cardini 2020). On this, it is enough to ponder that, even when minBEN sliding gets geometry right (sometimes it does (Gunz et al. 2005) and sometimes it does not (McCane 2013)), there is no demonstration that geometric correspondence equates biological homology. Curves and surfaces can be biologically homologous, but we do not measure them as a whole: we discretize continuous variation using arbitrary sets of points, whose mapping on the same developmental and anatomical features across individuals and species cannot be guaranteed by pure mathematics devoid of any biological insight.

Semilandmarks sometime are really useful, but have limitations which are rarely acknowledged (Cardini 2020; Cardini and Verderame 2022). If important in a study, however, the decision of whether and how to slide them is not straightforward. Here too there are pros and cons (e.g., (Cardini 2019; Perez et al. 2006)). minBEN sliding might be a very poor choice especially in the study of modularity/integration, if results of C19 and ZS23 are found to be general. Yet, it is probably the most common treatment of semilandmarks, when they are used in this type of research. Indeed, many analyses of minBEN slid shape data using the within a configuration approach have been published over the years by leading morphometricians in high impact journals. For readers to truly appreciate how mainstream the combination of minBEN sliding has become in modularity/integration research, I quote from several papers not to misreport the approach. I focus mainly on those using CR, because, unlike the PLS, CR is to my knowledge only implemented as a within a configuration method (as of June 12th 2023 there seems to be no option for separate superimpositions in the geomorph function to test CR[11] - (Adams and Otárola-Castillo 2013)). The examples, of which I generally read only the relevant parts of the methods, are presented mentioning the authors, study group and journal:

- p. 184 of Botton-Divet et al. on mustelids in Evolutionary Biology (Botton-Divet et al. 2018): “Semi-landmarks on surfaces and curves were slid in order to minimize the bending energy of a thin plate spline (TPS) between each specimen and a common reference … To visualize the differences in the integration pattern depending on the ecology of the species, we computed the covariance ratio”.

[11]See footnote 6 on this.

- p. 476-477 of McLean et al. on ground-dwelling squirrels in Evolution (McLean et al. 2018): “with semilandmarks slid to minimize bending energy ... tests were performed on untransformed shape data using the covariance ratio (CR; Adams 2016) implemented in the phylo.modularity function in geomorph”.
- p. 1007 and 1009 of Sherratt and Kraatz in Evolution (Sherratt and Kraatz 2023): “semilandmarks of the cranial roof and zygomatic arches ... slide along their tangent directions in order to minimize bending energy … We tested the three a priori defined hypotheses of modularity and evaluated which was the most supported for the data, using the covariance ratio (CR)”.
- p. 46-47 of Sansalone et al. on hominins in Nature Ecology & Evolution (Sansalone et al. 2023): “semilandmarks were slid along the curves by minimizing the bending energy of a thin plate spine deformation … We defined four different modular configurations and evaluate between them by using the standardized test statistics based on the comparison of the CR measurement”.
- p. 4 of Larouche et al. on ray-finned fishes in Scientific Reports (Evans et al. 2021): “semi-landmarks were superimposed using the minimum bending energy criterion … a total of 24 a priori hypotheses of modularity were analyzed … Three approaches were used to test these hypotheses: (1) the covariance ratio (CR) …”.
- p. 7-12 Watanabe et al. on birds in eLife (Watanabe et al. 2021): “Lastly, we employed two methods for evaluating the pattern of integration—covariance ratio (CR) (Adams, 2016) and maximum likelihood (ML) … to calculate and test the strength of correlation between shapes of neuroanatomical regions … we subjected the coordinate data to a generalized Procrustes alignment ... minimizing total bending energy, while allowing semi-landmarks to slide on the mesh surface”.
- p. 3 and 5 of Bardua et al.’s guideline on semilandmarks in Integrative Organismal Biology (Bardua, Felice, et al. 2019): “The ability to retain correspondence between data points is important for many morphological studies, especially to compare morphology across different regions of a structure, as in studies of modularity, and thus sliding semilandmark approaches may be particularly useful for studies that are concerned with questions ... The surface points are then slid to minimize total bending energy of a thin plate spline (TPS) across all specimens”.
- p. 1898 of Rhoda et al. on wrasses and snakes in Integrative and Comparative Biology (Rhoda et al. 2021): “Semi-landmarks in both datasets were slid by minimizing bending energy … We measured modularity using the Covariance Ratio (CR) in geomorph”.
- p. 18 of Bardua et al. on caecilians in BMC Evolutionary Biology (Bardua, Wilkinson, et al. 2019): “This approach resulted in a total of 687 semilandmarks equidistantly placed along curves ... a total of 729 surface points were placed onto each specimen … the curve and surface points are slid to minimise bending energy … To assess patterns of modularity we used two methods, both implemented in R: a maximum likelihood approach (EMMLi) and the covariance ratio (CR)”.
- p. 3 and 5 of Hanot et al. on horses in BMC Ecology and Evolution (Hanot et al. 2021): “162 sliding semilandmarks placed on 20 curves and constrained by anatomical landmarks, and 1250 surface sliding semilandmarks … slid along their tangent vectors/planes to minimize bending energy … The degree of cranial modularity was assessed using the CR”.

- p. 11548-11549 of Hedrick et al. on birds in Ecology and Evolution (Hedrick et al. 2019): “Semilandmarks were slid according to the bending energy criterion … covariance ratio (CR) coefficient was calculated from the data”.
- p. 3 of Knapp et al. on ceratopsian in the Proceedings of the Royal Society B (Knapp et al. 2021): “Semilandmarks were then slid to minimize bending energy … we further assessed modularity by calculating the covariance ratio (CR) for both the original and allometry-corrected datasets”.
- p. 559-560 of Felice & Goswami on birds in PNAS (Felice and Goswami 2018) : “all surface landmarks [i.e., 734 semilandmarks] were slid to reduce bending energy … We also evaluated the seven-module hypothesis ... by calculating covariance ratios (CR) between all pairs of modules”.
- p. 8 of Evans et al. on flatfishes in PNAS (Evans et al. 2021): “semilandmarks were slid along their tangent directions while minimizing bending energy … Evolutionary modularity was also assessed using the CR coefficient”.

ZS23 provide evidence that the superimposition may have an impact on some analyses of modularity/integration using a within a configuration approach. They acknowledge problems, but find (end of the abstract) that “the most alarming results are the extremely high frequency of significant integration in large samples (N = 500) and the non-normal distribution of effect sizes (Z-scores)”. They may be right (p. 165) that it is especially “troubling ... that the standardized effect size for integration, $Z_{r.pls}$, increases with sample size and that neither Z_{CR} nor $Z_{r.pls}$ are normally distributed”. In terms of the impact of N, however, they do not seem particularly worried, because such large samples are uncommon in empirical applications. On p/N and why large samples are not bad news, I have already said enough. Unlike ZS23, however, but with due caution not to generalize from a limited number of examples (mine, in C19, and theirs), I find very worrying that “minimize bending energy almost invariably induces significant modularity and integration” (p. 147). Thus, I would not say that the special aspects of how shape is defined and derived in Procrustes GMM is not an obstacle, when it might affect precisely the ‘high-density’ analyses that are, for many, the state-of-the-art of GMM (Goswami et al. 2019). Others had argued that precisely this type of analysis, using within a configuration tests of modularity/integration with minBEN slid semilandmarks, is “Not-Really-a-Problem” (p. 1, (Goswami et al. 2019)). This conclusion, the opposite of the results of ZS23, was reached by generalizing from a few specific examples (p. 12 of the Goswami and colleagues): “Thus, on the question of whether the use of sliding semilandmarks exacerbates the effect of Procrustes superimposition on covariance structure (Cardini 2019), the results of our third experiment suggest that adding landmarks neither improves nor inhibits the recoverability of modules. The fact that the direction of variation in sliding semilandmarks tends to be fairly uniform as a result of their fitting procedure (e.g., Perez et al. 2006) suggests that they will not improve recoverability to the same extent as covarying landmarks (or non-sliding semilandmarks) whose direction varies with respect to one another. However, sliding semilandmarks improve representation of complex structures, such as surfaces, far beyond the abilities of landmarks, and thus the increased complexity, and added variation in directionality of variation, will constrain centroid variation, improve the Procrustes fit relative to the “natural superimposition,” and thus increase the accuracy of recovering modules for biological structures”.

If, beyond the specific datasets they analysed, we will find out that Goswami and colleagues are wrong and ZS23 right, the minBEN slid within a configuration approach to modularity/integration might have to be completely abandoned and quite a few papers will have to be retracted or amended with a published correction. In fact, especially in the light of the strong effect of sliding on patterns of variance and covariance, other types of evolutionary analyses, such as, for instance, those focusing on disparity or evolutionary rates of modules (for examples, see references in Cardini and Verderame 2022) could be impacted and need a careful reassessment.

Concluding remarks

When I first found the problem of spurious modularity/integration in tests of superimposed data using the within a configuration approach, I expected a degree of controversy. I learnt the very basic of GMM at the end of Nineties, when the field, at least in terms of biological studies, was still fairly young. That was the time when people were pioneering applications, Procrustes had not yet become the leading approach and almost everyone made mistakes. Back then, one of the most heated debates concerned the use and interpretation of partial warps (PW), a type of shape variables computed by minimizing bending energy (Bookstein 1989). Rohlf (Rohlf 1998) showed that any analysis of PWs one at a time, or as a subset, was meaningless. However, the developers of that type of approach (Zelditch et al. (1992, 1993, 1995), Swiderski (1993), Fink and Zelditch (1995), Zelditch and Fink (1995), cited in Rohlf 1998) vehemently argued against Rohlf's conclusions (Zelditch et al. 1998). Curiously, the problem was (not so) distantly related to the use of subsets of landmarks after a common superimposition, as summarized by Zelditch et al. (1998, p. 161): "According to Rohlf, the most fundamental problem with the studies by Z&F is that they interpret the partial warps as biologically meaningful rather than as a priori variables whose definition does not reflect covariance patterns in the data". Likewise, I argue that, after a superimposition, shape coordinates should be analysed as a whole and not by splitting them into subsets (within a configuration approach) or, even worse, one landmark at a time (Cardini and Verderame 2022, and references therein). This is sound theory, it is not my discovery and implies that the most common analyses using a within a configuration approach to test modularity/integration are indeed violating a basic assumption of superimposition methods. The problem is there and, after ZS23, I am more convinced that can often (if not always) be made more serious by certain types of mathematical 'massage' of the data, such as minBEN sliding semilandmarks.

A second question concerns whether the problem makes a practical difference, i.e. if, in empirical datasets, within a configuration tests inflate the rate of type I errors and, thus, tend to produce frequent false positives, when power is adequate. I suspect that most statisticians would not be happy with tests that have an intrinsic bias, regardless of the magnitude of the impact. However, I am a 'non-numerically oriented' biologist, who struggle with statistical theory and appreciate empirical studies. This why I am happy with ZS23 focus on the empirical impact of the problem with the within a configuration approach to modularity/integration. However, I am not happy with misreporting research by others (but acknowledge it may happen and I may have done it myself) and I am even less happy with overstating results and providing misleading conclusions. I have never claimed (see all the quotes in this paper) that the within a configuration approach is bound to produce spurious results all the time, especially in data with real covariance. Rather, I may have said the opposite. I had also already stated that not only the isotropic model, but also the large samples may have overemphasized the effect (and, yet, it is always there: "eppur si muove[12]"!).

I was also cautious in predicting that sliding, especially using minBEN, might make things worse, which ZS23 fully supports. Even now, despite seeing more evidence of how serious the problem might be in 'high-density' morphometric analyses (Goswami et al. 2019), I would say that a honest statement of what we understand up to this point is that there is a clear theoretical issue, that cannot be overlooked. Its impact, however, is complex and may vary widely from case to case: sometimes it will be negligible, sometimes it will be strong and lead to spurious findings. Even the direction of the bias is not trivial to predict (C19). In fact, I doubt we have or will have a universal conclusion on the effect on tests of modularity/integration using the within a configuration approach. Again, I urge caution in applications and interpretations, and hope to read, together with the results of this type of tests, at least a clear statement on the potentially serious issues with this approach. Also, even more importantly, I am not an evo-devo biologist, but I am inclined to believe that sound research on modularity/integration requires integrative approaches, that merge morphology, developmental biology and genetics: morphometrics on its own, even with reliable tests, can only be a small piece of evidence in a much more complex puzzle.

Acknowledgements

This paper is dedicated to the evolutionary biologists, and all other scientists and people in Ukraine, who are suffering for the invasion of their country, but also to those Russians, who are against the war, including most of my colleagues and friends, whose silence is the only protest they are allowed.

Funding

The study was supported by the Fondo Ateneo di Ricerca DSCG (Università degli Studi di Modena e Reggio Emilia), project TAXON.

[12] https://www.merriam-webster.com/dictionary/eppur%20si%20muove

Declarations Conflict of interest

The author declares no conflict of interest.

References

Adams, D. C. (2016). Evaluating modularity in morphometric data: challenges with the RV coefficient and a new test measure. *Methods in Ecology and Evolution*, *7*(5), 565–572. https://doi.org/10.1111/2041-210X.12511

Adams, D. C., & Otárola-Castillo, E. (2013). geomorph: an r package for the collection and analysis of geometric morphometric shape data. *Methods in Ecology and Evolution*, *4*(4), 393–399. https://doi.org/10.1111/2041-210X.12035

Bardua, C., Felice, R. N., Watanabe, A., Fabre, A.-C., & Goswami, A. (2019). A Practical Guide to Sliding and Surface Semilandmarks in Morphometric Analyses. *Integrative Organismal Biology*, *1*(1), obz016. https://doi.org/10.1093/iob/obz016

Bardua, C., Wilkinson, M., Gower, D. J., Sherratt, E., & Goswami, A. (2019). Morphological evolution and modularity of the caecilian skull. *BMC Evolutionary Biology*, *19*(1), 30. https://doi.org/10.1186/s12862-018-1342-7

Björklund, M. (2019). Be careful with your principal components. *Evolution*, *73*(10), 2151–2158. https://doi.org/10.1111/evo.13835

Bookstein, F. L. (1989). Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *11*(6), 567–585. https://doi.org/10.1109/34.24792

Bookstein, F. L. (2017). A Newly Noticed Formula Enforces Fundamental Limits on Geometric Morphometric Analyses. *Evolutionary Biology*, *44*(4), 522–541. https://doi.org/10.1007/s11692-017-9424-9

Botton-Divet, L., Houssaye, A., Herrel, A., Fabre, A.-C., & Cornette, R. (2018). Swimmers, Diggers, Climbers and More, a Study of Integration Across the Mustelids' Locomotor Apparatus (Carnivora: Mustelidae). *Evolutionary Biology*, *45*(2), 182–195. https://doi.org/10.1007/s11692-017-9442-7

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

Cardini, A., Elton, S., Kovarovic, K., Strand Viðarsdóttir, 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., 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., & 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

Evans, K. M., Larouche, O., Watson, S.-J., Farina, S., Habegger, M. L., & Friedman, M. (2021). Integration drives rapid phenotypic evolution in flatfishes. *Proceedings of the National Academy of Sciences*, *118*(18). https://doi.org/10.1073/pnas.2101330118

Felice, R. N., & Goswami, A. (2018). Developmental origins of mosaic evolution in the avian cranium. *Proceedings of the National Academy of Sciences*, *115*(3), 555–560. https://doi.org/10.1073/pnas.1716437115

Goswami, A., Watanabe, A., Felice, R. N., Bardua, C., Fabre, A.-C., & Polly, P. D. (2019). High-Density Morphometric Analysis of Shape and Integration: The Good, the Bad, and the Not-Really-a-Problem. *Integrative and Comparative Biology*, *59*(3), 669–683. https://doi.org/10.1093/icb/icz120

Gunz, P., & Mitteroecker, P. (2013). Semilandmarks: a method for quantifying curves and surfaces. *Hystrix, the Italian Journal of Mammalogy*, *24*(1), 103–109.

Gunz, P., Mitteroecker, P., & Bookstein, F. L. (2005). Semilandmarks in Three Dimensions. In D. E. Slice (Ed.), *Modern Morphometrics in Physical Anthropology* (pp. 73–98). New York: Kluwer Academic Publishers-Plenum Publishers. http://www.springerlink.com/content/r188217l01734877/. Accessed 22 December 2011

Hanot, P., Bayarsaikhan, J., Guintard, C., Haruda, A., Mijiddorj, E., Schafberg, R., & Taylor, W. (2021). Cranial shape diversification in horses: variation and covariation patterns under the impact of artificial selection. *BMC Ecology and Evolution*, *21*(1), 178. https://doi.org/10.1186/s12862-021-01907-5

Hedrick, B. P., Cordero, S. A., Zanno, L. E., Noto, C., & Dodson, P. (2019). Quantifying shape and ecology in avian pedal claws: The relationship between the bony core and keratinous sheath. *Ecology and Evolution*, *9*(20), 11545–11556. https://doi.org/10.1002/ece3.5507

Klingenberg, C. P. (2021). How Exactly Did the Nose Get That Long? A Critical Rethinking of the Pinocchio Effect and How Shape Changes Relate to Landmarks. *Evolutionary Biology*, *48*(1), 115–127. https://doi.org/10.1007/s11692-020-09520-y

Knapp, A., Knell, R. J., & Hone, D. W. E. (2021). Three-dimensional geometric morphometric analysis of the skull of Protoceratops andrewsi supports a socio-sexual signalling role for the ceratopsian frill. *Proceedings of the Royal Society B: Biological Sciences*, *288*(1944), 20202938. https://doi.org/10.1098/rspb.2020.2938

McCane, B. (2013). Shape Variation in Outline Shapes. *Systematic Biology*, *62*(1), 134–146. https://doi.org/10.1093/sysbio/sys080

McLean, B. S., Helgen, K. M., Goodwin, H. T., & Cook, J. A. (2018). Trait-specific processes of convergence and conservatism shape ecomorphological evolution in ground-dwelling squirrels. *Evolution*, *72*(3), 473–489. https://doi.org/10.1111/evo.13422

Perez, S. I., Bernal, V., & Gonzalez, P. N. (2006). Differences between sliding semi-landmark methods in

geometric morphometrics, with an application to human craniofacial and dental variation. *Journal of Anatomy*, *208*(6), 769–784. https://doi.org/10.1111/j.1469-7580.2006.00576.x

Rhoda, D., Segall, M., Larouche, O., Evans, K., & Angielczyk, K. D. (2021). Local Superimpositions Facilitate Morphometric Analysis of Complex Articulating Structures. *Integrative and Comparative Biology*, *61*(5), 1892–1904. https://doi.org/10.1093/icb/icab031

Rohlf, F. J. (1998). On Applications of Geometric Morphometrics to Studies of Ontogeny and Phylogeny. *Systematic Biology*, *47*(1), 147–158.

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.

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

Sansalone, G., Profico, A., Wroe, S., Allen, K., Ledogar, J., Ledogar, S., et al. (2023). Homo sapiens and Neanderthals share high cerebral cortex integration into adulthood. *Nature Ecology & Evolution*, *7*(1), 42–50. https://doi.org/10.1038/s41559-022-01933-6

Sherratt, E., & Kraatz, B. (2023). Multilevel analysis of integration and disparity in the mammalian skull. *Evolution*, *77*(4), 1006–1018. https://doi.org/10.1093/evolut/qpad020

Watanabe, A., Balanoff, A. M., Gignac, P. M., Gold, M. E. L., & Norell, M. A. (2021). Novel neuroanatomical integration and scaling define avian brain shape evolution and development. *eLife*, *10*, e68809. https://doi.org/10.7554/eLife.68809

Zelditch, M. L., Fink, W. L., Swiderski, D. L., & Lundrigan, B. L. (1998). On Applications of Geometric Morphometrics to Studies of Ontogeny and Phylogeny: A Reply to Rohlf. *Systematic Biology*, *47*(1), 159–167.

Zelditch, M. L., & Swiderski, D. L. (2023). Effects of Procrustes Superimposition and Semilandmark Sliding on Modularity and Integration: An Investigation Using Simulations of Biological Data. *Evolutionary Biology*, *50*(2), 147–169. https://doi.org/10.1007/s11692-023-09600-9