



Less tautology, more biology? A comment on “high-density” morphometrics

Andrea Cardini^{1,2}

Received: 5 June 2020 / Revised: 31 August 2020 / Accepted: 4 September 2020 / Published online: 22 September 2020
© Springer-Verlag GmbH Germany, part of Springer Nature 2020

Abstract

In the context of geometric morphometric analyses of modularity and integration using Procrustes methods, some researchers have recently claimed that “high-density geometric morphometric data exceed the traditional landmark-based methods in the characterization of morphology and allow more nuanced comparisons across disparate taxa” and also that, using “high-density” data (i.e., with dozens or hundreds of semilandmarks), “potential issues [with tests of modularity and integration] are unlikely to obscure genuine biological signal”. I show that the first claim is invalidly tautological and, therefore, flawed, while the second one is a speculation. “High-density” geometric morphometrics is a potentially useful tool for the quantification of continuous morphological variation in evolutionary biology, but cannot be said to represent absolute accuracy, simply because more measurements increase information, but do not by default imply that this information is accurate. Semilandmarks are an analytical expedient to break the continuity of regions devoid of clearly corresponding landmarks, but the shape variables which they generate are a function of the specific choice of the placement and possible mathematical manipulation of these points. Not only there are infinite ways of splitting a curve or surface into discrete points, but also none of the methods to slide the semilandmarks increases the accuracy of their mapping onto the underlying biological homology: indeed, none of them is based on a biological model, and the assumption of universal equivalence between geometric and biological correspondence is unverified, if at all verifiable. Besides, in the specific context of modularity and integration using Procrustes geometric morphometrics, the limited number of scenarios simulated until now may provide interesting clues, but do not yet allow strong statements and clear generalizations. The Procrustes superimposition does alter the ‘true’ covariance structure of the data and sliding semilandmarks further contributes to this change. Although we hope that this might only add a negligible source of inaccuracy, and simulations using landmarks (but no semilandmarks yet) suggest that this might be the case, it is too early to confidently share the view, expressed by the promoters of high-density methods, that this is “Not-Really-a-Problem”. The evidence is very preliminary and the dichotomy may not be this simple, with the magnitude (from negligible to large) and direction (inflation of modularity, integration, or both) of a potential bias in the tests likely to vary in ways specific to the data being analysed. We need more studies that provide robust and generalizable evidence, without indulging in invalid tautology and over-interpretation. With both landmarks and semilandmarks, what is measured should be functional to the specific hypothesis and we should be clear on where the treatment of the data is pure mathematics and where there is a biological model that supports the maths.

Keywords Accuracy · Homology · Modularity · Procrustes methods · Shape analysis · Semilandmarks

✉ Andrea Cardini
alcardini@gmail.com; andrea.cardini@unimore.it

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

Introduction: Geometric morphometrics (GMM) and “high-density” analyses of modularity and integration

Over the last three decades, Procrustes GMM has become the dominant morphometric approach to study shape differences in biology (O’Higgins 2000; Adams et al. 2004, 2013; Cardini and Loy 2013). To obtain shape data, the first step is to measure morphology (e.g., a vertebrate

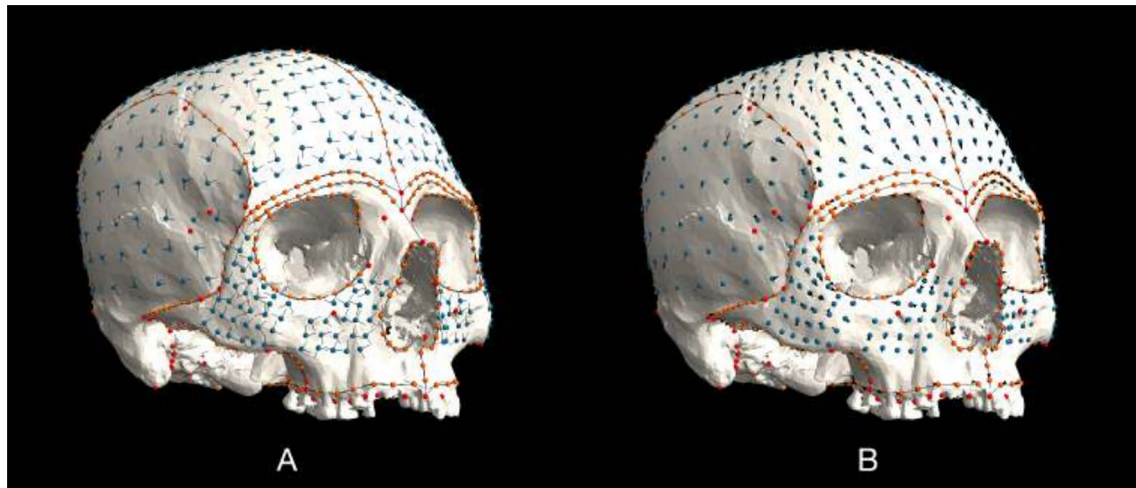


Fig. 1 Semilandmarks on curves (red) and surfaces (blue) before (a) and after (b) sliding: arrows show the positions of the slid semilandmarks relative to the original unslid ones: on the frontal bone, they clearly shift in a consistent direction, suggesting directionality in the

bone or an insect wing) using a set of anatomical points (called landmarks) with a precise one-to-one correspondence across all individuals and taxa. This correspondence, which is often called “homology”, establishes the anatomical and evolutionary equivalence of the traits being measured (Klingenberg 2008). After landmarking, size and meaningless positional differences are ‘standardized’ using a least-squares procedure called generalized Procrustes analysis (GPA) or simply Procrustes superimposition (Rohlf and Slice 1990; Adams et al. 2013). The GPA allows us to numerically describe morphology in terms of the geometry captured by a specific set of landmarks, and therefore, this method belongs to a larger family of geometric morphometric techniques (Rohlf and Marcus 1993).

Using semilandmarks (Gunz and Mitteroecker 2013), Procrustes GMM has been extended to the analysis of curves and surfaces. Semilandmarks are simply points placed on curves (e.g., along a suture) or surfaces (e.g., the cranial vault) to obtain shape coordinates in regions lacking precise landmarks. The number, density, and positions of the semilandmarks can be optimized, but they are fundamentally arbitrarily decided by an operator or an algorithm. In particular, it is a customary procedure to slide semilandmarks (as exemplified in Fig. 1) to mathematically improve (without any biological model behind the numerical manipulation) the topographic correspondence of these points (DeQuardo et al. 1996). Finally, the matrix of landmark and semilandmark shape coordinates, taken all together, is analysed using multivariate statistics. Because of the lack of clear anatomical correspondence, and the large number of semilandmarks used in this type of analyses, these applications have been sometimes called

covariance introduced by sliding (reprinted from Fig. 3 of (Gunz and Mitteroecker 2013) under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Italy License)

“homology free” approaches (Polly 2008a) or “high-density” GMM (Goswami et al. 2019).

Procrustes GMM and its “high-density” extensions make important assumptions that require a specific attention in both the statistical analyses, as well as in the interpretation and visualization of results (O’Higgins 2000; Klingenberg 2008, 2013; Viscosi and Cardini 2011). Violations of these assumptions produce inaccuracies that can range from misinterpretations of shape differences (Viscosi and Cardini 2011; Klingenberg 2013) and serious analytical artifacts in phylogenetic applications (Rohlf 1998; Adams et al. 2011) to more recently discovered (Cardini 2019) and less well-understood spurious results and uncertainties in some of the commonly used tests of modularity and integration (TEMI). Besides, and somewhat unsurprisingly, having very large numbers of variables relative to sample size, as common in “high-density” GMM, can create other potential issues with subsequent applications of particular statistical analyses. These are not specific to GMM and include, for instance, fairly simple ordination methods, such as canonical variate analysis (Kovarovic et al. 2011) and principal component analysis (Bookstein 2017; Cardini et al. 2019), in which patterns of shape variation can be distorted or inflated. As with TEMI, the degree of inflation of differences might vary from negligible to large or even completely spurious, but it is never good news: as pointed out by Cardini et al. (2019), in relation to the problems with between-group principal component analyses, “showing a consistent spurious degree of separation between groups is not a promising property for a method that was proposed to analyze data with large numbers of variables [i.e., ‘big data’ in genetics and morphometrics] and small samples”.

In this context, the points I am going to discuss are broad and of general interests for scientists whose aim is an accurate quantification of phenotypic variation in biological forms: as a reviewer pointed out, “the paper could have been written and would be a valuable contribution even without Goswami et al. (2019) [see below] to structure its content”. Nevertheless, it is a fact that my main inspiration came from a specific paper (Goswami et al. (2019), henceforth, for brevity, referred to as G19) entitled “High-Density Morphometric Analysis of Shape and Integration: The Good, the Bad, and the Not-Really-a-Problem”. G19 was, in turn, largely prompted by my recent article (Cardini 2019) on spurious results in Procrustes shape coordinates from random isotropic data analysed using ‘within a configuration’ methods for testing modularity and integration. From there, the authors broadened their scope to conclude not only that the issues I found may be negligible in semilandmark data, but that, in fact, high density is the tool to achieve accuracy in modern morphometrics. Thus, they simulated data with different numbers of landmarks and patterns of covariance, and demonstrated that there are, indeed, potential issues related to how the Procrustes superimposition alters the patterns of variance and covariance (“The Bad”), but they span a range of severity (from negligible to serious). However, they also explored if, in general, high-density data provide a more accurate description of morphological variation, and reported that this is, indeed, the case (“The Good” in the title of the paper), making them ‘superior’ to landmark data (p. 669, “*exceed traditional landmark-based methods in characterization of morphology*”—here and elsewhere, I use italics to stress important points). Finally, because the number of anatomical points used to measure form does not seem to have a strong impact on TEMI and because real data are unlikely to have the type of strong directional covariation that increases the severity of the errors in the tests, they stated that “these *potential issues are unlikely to obscure genuine biological signal*” (p. 669). This, therefore, means that the altered pattern of variance and covariance in Procrustes GMM is “Not-Really-a-Problem” and we can go on safely with the “high-density morphometric approaches ... immense potential to propel a new class of studies of comparative morphology and phenotypic integration” (G19, p. 669).

I found several interesting results in G19, but also major flaws, which risk to exacerbate an already existing trend towards misinterpretation, overstatement, and abuse of, otherwise, useful techniques. Accordingly, I organized the paper in terms of what I see as: ‘The Good’ (i.e., simulations and new analyses exploring the issue with TEMI); ‘The Bad’ (i.e., circular reasoning, or invalid tautology, in demonstrating the ‘superiority’ of “high-density” data); and ‘The not-so-accurate’ (i.e., inaccurate representations of issues, as well as overstatements). Nevertheless, my criticisms on

circular reasoning and inaccuracies in describing what methods can or cannot do are broader (and do not only concern G19): invalid (i.e., based on wrong premises) tautology (<https://www.merriam-webster.com/dictionary/tautology>: “a statement that is true by virtue of its logical form alone”) has no place in morphometrics and more generally in science; thus, and more specifically, methodological advancements and analytical solutions cannot be ‘biology-free’ in evolutionary studies (Klingenberg 2008; Oxnard and O’Higgins 2009).

‘The Good’: new simulations and more attention to the problem

I start briefly with ‘The Good’, in my view. The simulation by G19 is intriguing, although I confess that my understanding of what was done is limited and I cannot assess the accuracy and, more importantly, the generalizability of the conclusions. It seems to me that the reality of the issues discovered by my study (Cardini 2019) is confirmed: the covariance introduced by the Procrustes superimposition can lead to spurious results in testing integration and modularity (for instance, increasing false positives when modules are not re-superimposed within a configuration). However, G19’s results suggest that highly variable directions in patterns of covariance across landmarks (as expected for real biological data) might moderate the effect of the superimposition and allow the correct inference of modules. In fact, more accurately, as a referee pointed out: “both papers showed that sometimes the effect [of the inaccuracies caused by the superimposition] is severe, sometimes it is not. ... [G19 suggests that the problem might be] most severe when the configuration of landmarks and the direction of their variation has a strong effect on the ... location of the centroid ... an effect ... independent of the number of landmarks, but highly dependent on the overall shape of the object and the pattern of covariation ... [Because uniformly] increasing the density of semilandmarks does not necessarily affect the position of the centroid point nor the pattern of directional variation [which may be strongly dependent on the landmarks to which they are anchored] ..., higher density will not normally fix the problem. Furthermore, if the density is increased in some areas more than others, it could exacerbate the problem if the increase is near landmarks that contribute strongly to centroid position”. Therefore, although G19’s findings look promising, they do not indicate that we can be confident about a negligible impact on TEMI and it is probably *premature* (based on a relatively small set of simulations, using simple made-up configurations with just two modules and apparently equal numbers of landmarks within each of the two modules) *to generalize and conclude that the “Procrustes superimposition does not mislead analyses of*

integration in biologically realistic scenarios ... regardless of number of landmarks or semilandmarks” (p. 680, G19).

There is certainly a need for more work and some questions are potentially unanswered. I am not sure, for instance, that, besides exploring whether simulated modules are correctly retrieved in the analysis, G19 quantified the magnitude of the distortion of the real pattern: one could approximate the answer regarding whether modules are present or not, but not be able to accurately estimate the strength of modularity (or integration). Among other factors, which can affect results, there might be how many modules are described (with alternative schemes being differentially impacted by the superimposition), as well as the heterogeneity in the number, density, and types of points within each module (plus, as suggested by a referee, the relative number and distribution of the landmarks to which semilandmarks are anchored). Also, the recoverability of modules seemed unaffected by the number of landmarks in G19, but sliding was not simulated and this may be important (see below).

That, in the third and last set of analyses on real samples (G19), TEMI produced congruent results regardless of whether using only landmarks or also including semilandmarks sounds reassuring, but we do not know the real covariance structure of those data and at best one could argue that this suggests precision [“the way in which repeated observations conform to themselves” (Kendall et al. 1957)], but does not say much about accuracy (closeness to the truth). Sliding of semilandmarks adds a layer of complexity, because it not only alters the covariance structure of the data by aligning the specimens in a sample as in a standard GPA, but also because it moves the semilandmarks to mathematically increase their correspondence, thus further contributing to the alteration of the pattern of variances and covariances (Cardini et al. 2019, and references therein). As mentioned, regardless of arguments about differences and more or less desirable properties of the different approaches to sliding (Perez et al. 2006; Gunz and Mitteroecker 2013), the choice of the mathematical treatment of the semilandmarks is arbitrary, and, in fact, even the decision to slide is optional. Indeed, this operation maximizes the topographic correspondence of semilandmarks, but the two most common sliding methods (DeQuardo et al. 1996) minimize either the sum of squared shape distances or *bending energy* (<https://life.bio.sunysb.edu/morph/glossary/gloss1.html>: “a metaphor borrowed for use in morphometrics from the mechanics of thin metal plates”), with none of these quantities having any obvious biological meaning. How much these choices matter in terms of variance–covariance patterns is easily seen by comparing variance–covariance matrices with unslid GPA data, slid data using the minimum Procrustes distance approach or slid by minimizing bending energy, as well as, often, by simply looking at the differences in the scatter of superimposed configurations, as in Figs. 2, 3, as

well as in Fig. 3 of (Perez et al. 2006) and Fig. 2 of (Cardini 2019). Yet, despite the weaknesses of G19’s assertions about the value of high-density semilandmark schemes, their simulations provide useful insights to factors that contribute to covariance artifacts during Procrustes superimposition. Indeed, it is important that morphometricians are beginning to pay more attention to problems with TEMI, which, above all, was the main message of my 2019 paper (p. 102), with its “... exploratory nature, ... limited number of tests examined and ... main focus on type I errors, and therefore the need of further in depth research considering a large variety of scenarios”.

Thus, going back to the observation that patterns of modularity in real datasets were similar when tested using reduced and “high-density” configurations (G19), although I call this precision, it is an interesting result. On this, G19 makes some contradictory statements (see the end of the second to the last section of this paper), but, in reality, their results seem largely in agreement with my observations (Cardini 2019), and our interpretations, although different, are not necessarily mutually exclusive. In terms of the congruence of results from landmark-only and “high-density” data, for instance, G19 finds that this result further motivates users to employ “high-density” data. In contrast, I am reassured that, using a presumably well-designed configuration, one may need many less variables to achieve about the same conclusion as with ‘big data’. A good agreement between smaller and much larger landmark datasets has been found by others, such as a QTL analysis (Navarro and Maga 2016) that discovered 23 loci using 13 landmarks and almost 600 semilandmarks on the mouse mandible, but managed to detect 19 of them even including only those 13 landmarks. In a probably more balanced conclusion than in many papers using ‘big data’, the authors of this study stated (p. 1153): “It appears that finer phenotypic characterization of the mandibular shape with 3D landmarks, along with higher density genotyping, yields better insights into the genetic architecture of mandibular development. Most of the main variation is, nonetheless, preferentially embedded in the natural 2D plane of the hemi-mandible, reinforcing the results of earlier influential investigations”, with those earlier investigations using about 2% of the total number of shape variables employed in the “high-density” study.

‘The Bad’: tautology does not demonstrate accuracy of “high-density” data

Flawed assumption and flawed conclusions

I will now focus mostly on ‘The Bad’, which is where I am in disagreement with the authors of G19. The approach of (Watanabe 2018) (henceforth, abbreviated as W18), followed

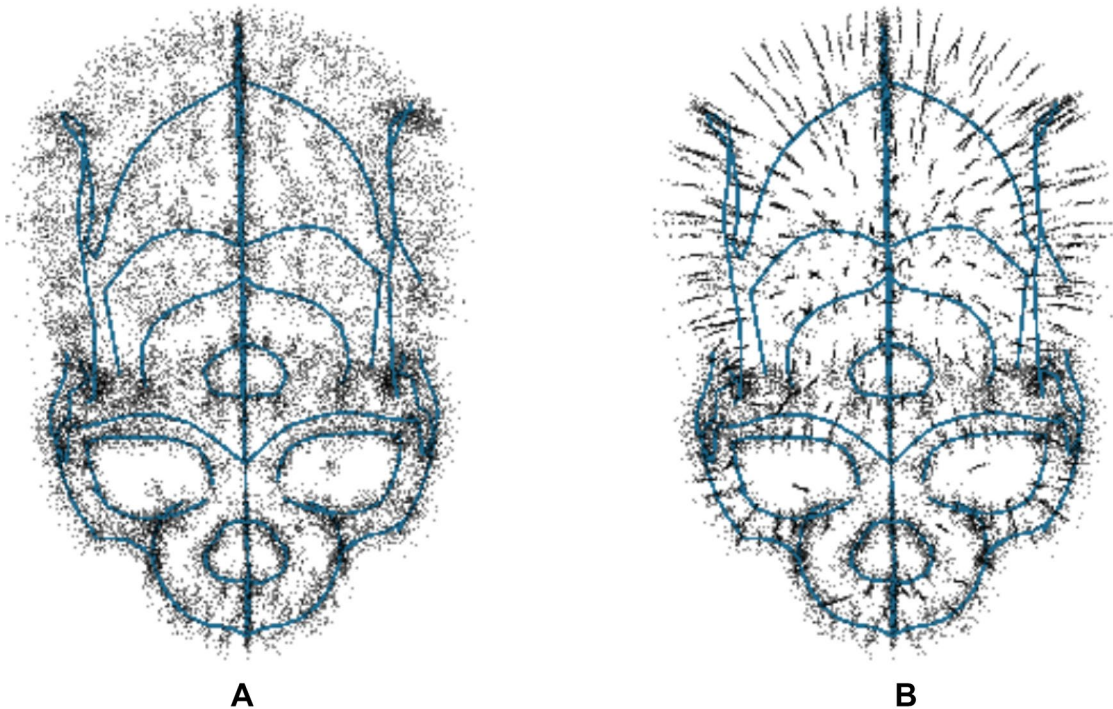


Fig. 2 Semilandmarks on a sample of crania slid using the minimum bending energy criterion (**a**) or the one minimizing Procrustes shape distances (**b**). The figure clearly shows differences due to the impact on covariance of the choice of treatment of semilandmarks

using functions which are elegant mathematics but have no biological model behind (reprinted from Fig. 5 of (Gunz and Mitteroecker 2013) under a Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Italy License)

by G19, is an invalid tautology: they judge the accuracy of less dense configurations using the most dense dataset as the reference for the comparison, but there is no demonstration in the first place that the highest density equates with accuracy, which is as nonsensical as arguing that measuring more is always better, regardless of what I measure and why. That inevitably makes results and conclusions, based on flawed epistemological premises, misleading, and inaccurate. More precisely, the misleading idea is to decide the appropriate number of landmarks (and/or semilandmarks) based on the assessment of the correspondence between similarity relationships calculated using the full “high-density” landmark configuration compared to its random subsets, which are reduced configurations made of randomly selected points from the full configuration. Thus, (p. 673) they use a “function [that] generates a sampling curve ..., where a plateau in the curve signifies stationarity in characterization of shape variation and fewer landmarks than the plateau indicates inadequate characterization”. With this, they arrive at the conclusion that “semilandmarks provide far more comprehensive, as well as accurate, characterizations of morphological variation than analysis of landmarks alone” (p. 681).

It seems convincing at first that covering all anatomical regions with points, getting measurements out of them, and, maybe, optimizing their number using W18 provides

a better quantification of form. Unfortunately, if a reader thinks carefully about what was done, the invalid tautology becomes obvious: it is logically flawed to argue for how many landmarks/semilandmarks one needs for absolute accuracy (“characterization” or sampling of “morphology” in the authors’ words), after one has already and arbitrarily decided that the full configuration (i.e., the one with the highest number of points) is the most accurate. But can we be sure about this? Indeed, do we really know that the “highest density” configuration, used to decide if more or less points are better, represents accuracy? This is a given in W18-G19, but the truth is that we do not know: it could be or it could not, but, because we do not know, we cannot rely on this assumption to decide if more or less points are necessary. Since the assumption is wrong, and *the reasoning circular (premise: “high-density” is the best proxy for accuracy; conclusion: more points are better than less), the conclusion has no heuristic value and is invalid.*

I will say more on this point later, but, just to make it less abstract, let us imagine that we are comparing cranial shape among adult living and fossil monotremes to elucidate how morphology diverged during the evolution of this group. Because their crania have extensive fusions of the sutures, we may find few clearly corresponding landmarks (say, just 10). Therefore, we might opt for high-density

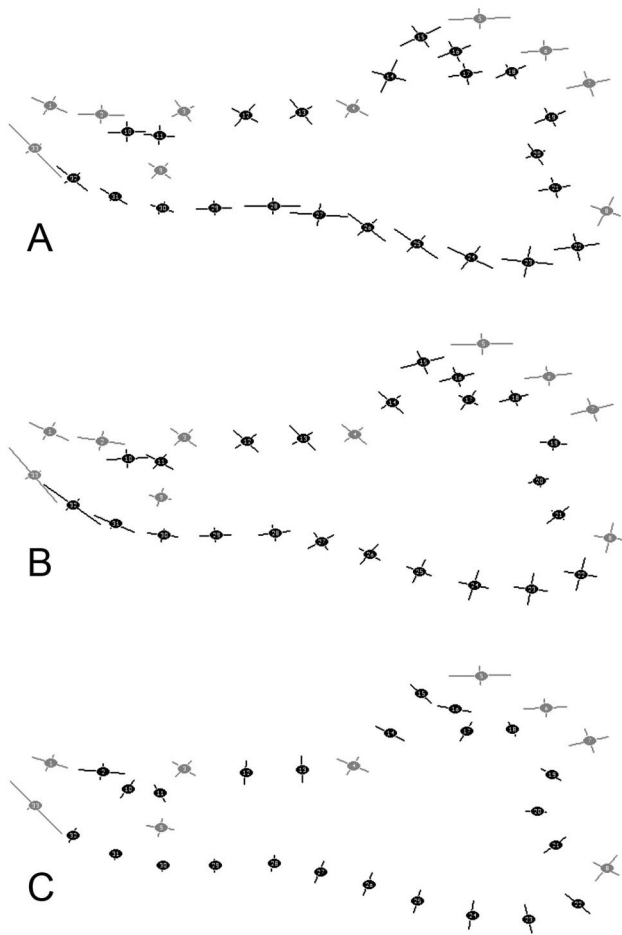


Fig. 3 Hemi-mandible mean shape of a sample of 33 adult yellow-bellied marmots analysed using 2D GMM; vectors, computed in GRF-ND (Slice 1994), show the main directions of variation at each landmark (grey) and semilandmark (black), with a fourfold magnification in all plots to emphasize differences: **a** GPA superimposition with unslid semilandmarks; **b, c** GPA superimposed data with semilandmarks slid using either **(b)** the minimum bending energy or **(c)** minimum Procrustes distance method

morphometrics and cover the cranium with an initial set of 10,000 semilandmarks (uniformly distributed and then maybe slid). We then apply W18-19's approach to deciding whether the dense semilandmark description is more 'accurate' than the one based on just 10 landmarks. Indeed, we find that a minimum of several hundred points is required to get a shape description close to the full configuration, whereas with just 10 landmarks shape differences (to the same full configuration of more than 10000 points) are huge: thus, we argue that high density provides better accuracy and go on with that to find clusters that suggest similarities and evolutionary patterns (with a most impressive visualization as a bonus!). Unfortunately, looking at the development of the bones that make up the cranium in the monotremes, we might later realize that

we did measure an overall homologous cranial surface (as argued by those who side-lines the shortcomings of semilandmarks by saying that points cannot be homologous but curves and surfaces are), but, across species, most of the semilandmarks of each bone clearly map on wrong non-homologous bones. Thus, regardless of the seemingly reassuring results of W18-G19's approach, showing that we have chosen the most accurate configuration for the monotreme comparison, the high dense description is evolutionarily and developmentally highly inaccurate and probably worse than the quantitatively smaller, but qualitatively superior information embedded in the 10 anatomically equivalent landmarks.

This made-up example is extreme and unlikely (although not impossible, as in the rodent example I will consider later), but it helps to emphasize that the same problem (a consequence of our ignorance of what corresponds to what) occurs when, in a study of human evolution, we cover the cranial vault of a fossil hominin with a dense set of semilandmarks: we cannot claim biological accuracy and 'homology' of shape descriptors, because we do not actually know if the points are truly mapping on equivalent features. Compared to the hypothetical example with the monotremes, the error will be likely less in degree but similar in nature: even if the whole bone is homologous, high density does not guarantee the equivalence of the features being measured and does not equate by default to accuracy.

Thus, what is assessed by G19 using the method of W18 is purely precision, which, as anticipated, is about how close estimates are among themselves, regardless of whether they are close to the truth. An archer whose arrows end up all more or less in the same spot is precise, but, if that spot is not the center of its target, the archer is inaccurate. In morphometrics, applied to evolutionary questions, the target is the aspect we are interested to measure, but the measurements must be biologically equivalent and thus 'homologous'. I am using single quotes for this attribute, as homology has a variety of definitions and, certainly, there is a degree of fuzziness and ambiguity in its use in morphometrics (Smith 1990; Klingenberg 2008). I am not claiming absolute rigour in its application to landmarks in evolutionary analyses, but there is a range that goes from less to completely ambiguous: bregma, the meeting point of coronal and sagittal sutures, is not the same as *i*th semilandmark on the cranial vault, treated as a single bone and covered in an arbitrary number of semilandmarks. To get a good intuition of what the term 'homologous' approximates in this context, consider also the most ambiguous extreme of this range, which is to measure everything regardless of its biological and evolutionary comparability, as in a narrow-sense phenetic approach that focuses exclusively on quantity: then, the more characters or measurements I get, the better, and it is irrelevant if they mean the same across all specimens and taxa.

Because our understanding of morphological evolution and developmental change is still very approximate most of the time, the number of clearly ‘homologous’ features is limited and cannot *simply* be naively increased using a ‘blanket’ of “high-density” points. That could be a good approach for computer animation, digital art, or even some kind of ‘brute force’ biometric identification,¹ but much less well targeted at comparing homologous aspects of form in evolutionary studies and probably, at best, ‘noisy’ as a source of the biological signal. Thus, *the evidence produced by G19 can be informative on precision, but not on accuracy and cannot be used to conclude in absolute terms that “for many skull regions, 20–30 landmarks and/or semilandmarks are needed to accurately characterize their shape variation, and landmark-only analyses do a particularly poor job of capturing shape variation in vault and rostrum bones”*. As anticipated in this section, we all know that the cranial vault, for instance, offers a few ‘homologous’ landmarks to measure its shape, and we can say, probably without any analysis, that those few ones do a poor job at capturing its complexity (if that is truly necessary for our scientific investigation!). Yet, we cannot say that comparing the match of shape distances obtained using those few landmarks to those from the full configuration with hundreds of semilandmarks demonstrates the inaccuracy of landmarks-only data and the superiority of having 20–30 points per anatomical region: first, we failed to relate the choice of the measurements to the study question (Oxnard and O’Higgins 2009), as discussed in more depth below; crucially, we have provided no evidence whatsoever that the full configuration is accurate, except if our meaning of accuracy is moving towards the progressively more inaccurate and naive end of the narrow-sense phenetic approach (Klingenberg 2008; Oxnard and O’Higgins 2009). I am not claiming that either landmarks-only is accurate or that landmarks plus dozens or hundreds of semilandmarks is inaccurate: simply, we do not know, and this is why the full configuration cannot be used as the ‘meter of comparison’.

Besides the obvious issue of the complete lack of evidence in G19 for claiming that “high-density” equates by definition to absolute accuracy in morphological “characterization”, there are other important questions to ask before deciding how and what to measure in morphometrics (Oxnard and O’Higgins 2009). For instance, even if

sampling everything is crucial, can we exclude that the configuration with the highest number of points has not actually increased the noise in the data more than the signal we were interested in? Is there a way to discern this inaccuracy, if the term of comparison for accuracy is the highest density configuration itself? To start, what is the signal we are looking for? Does one size really fits all or is it possible that the measurements relate to the question we are asking (Oxnard and O’Higgins 2009)?

The approach of W18-G19 gives little importance to all these questions and the decision on how many landmarks/semilandmarks are needed is based on a single main criterion: the largest number of points represents absolute accuracy, from which it follows that everything getting closer to this target is better. However, besides the problematic tautology, why one is measuring something is a question that cannot be ignored and it is in fact the first step to decide what to measure (Oxnard and O’Higgins 2009). Would I be wise to decide what the best car is by exclusively focusing on speed? Then, any car whose speed is closer to the fastest is better. Unfortunately, I may not be a racing driver and many other aspects of a car could be more interesting for the use I intend to do: for instance, space, because I may have a large family; a smaller carbon footprint, because I am concerned about the environment; and robustness and flexibility, because I often drive on slippery roads full of potholes (yes, I live in Italy!). Thus, the fastest car may not be the best by definition. With car speed, however, at least there is an objective quantity to measure and the problem is whether this estimator of performance works well regardless of why I need a car. In contrast, with “high-density” landmarking, whether one configuration is more appropriate in relation to my aim is only part of the problem, and this is discussed at greater length in the section on ‘why measure’. The other, and probably even clearer, issue is the one I have already presented: even if my study question does require the capturing of as many details as possible of a structure, because we do not know if semilandmarks are really located on equivalent features across all specimens and taxa (Klingenberg 2008; Oxnard and O’Higgins 2009), “high-density” cannot be assumed a priori to be the most accurate representation of morphology.

We can agree that “*it is uncontroversial that semilandmarks can sample more morphology than ... landmarks*” (p. 673, G19), but *how and what they sample is less clear*. Using another example, borrowed this time from genetics, it is true that a sequence including all bases in a genome samples much more DNA than a group of phylogenetically informative genes or polymorphisms. Yet, it is unlikely that anyone would employ this most complete sequence (coding and non-coding regions, neutral sequences, and regions with highly variable evolutionary rates, junk DNA, etc.) to infer phylogeny. Yet, at least with DNA, the alignment of the sequences employs models that incorporate knowledge from

¹ As a referee remarked “In many studies that use dense correspondences the aim is simple identification, which ... does not require points to be equivalent in any biologically meaningful sense (because differences are not interpreted).[However,] how we measure affects relative differences among specimens in complex ways—arbitrary measurements give arbitrary distances ... [which] may not tally with the everyday experience of many workers, especially in computer science, and increasingly in biology, where the seeking of dense correspondences between surfaces of objects by algorithm is common”.

decades of research in genetics and molecular evolution. In contrast, Procrustes (plus the potential treatment of semilandmarks) aligns specimens based purely on mathematics and the desirable statistical properties of the shape space that the superimposition generates (Rohlf 2000a, b; O’Higgins 2000; Adams et al. 2004).

Assuming, a priori, that “high-density” is best is unfounded and can weaken one’s power to detect an effect of interest. For instance, in a paper on murid cranial evolution which I reviewed a few years ago, the authors (coming from a strong genetic background but with probably less experience on GMM) treated the whole outline of a cranium, photographed in ventral view, as if it was a single bone with no landmarks at all except for its opposite ends (rhinion and opisthion). This means that, for instance, many of the semilandmarks mapping on the maxilla in long-faced taxa might happen to be on the zygomatic arch in a species with a short rostrum, and viceversa. However, if the premise is that a “high-density” characterization of the cranial outline represents accuracy, one starts with the densely sampled cranial outline I described above and, following W18-G19, tests the ‘accuracy’ (sensu W18-G19) of reduced configurations with fewer semilandmarks. Almost certainly, that would demonstrate that more semilandmarks are better than fewer, without unfortunately providing any clue to the most important problem, which is that, because the “high-density” configuration was nonsensical in the first place for a comparative evolutionary study, the result of the application of W18-G19 is meaningless.

What semilandmarks ‘cannot know’ and more on why precision is not accuracy

G19 makes a distinction (p. 671) between semilandmarks (“those whose initial position is relative to landmarks with biological homology”) and pseudolandmarks (“automatically placed without reference to ... landmarks”), but the difference is subtle, as semilandmarks are often placed automatically and the boundary landmarks, that provide the reference for their definition, may be questionable and subject to a variable degree of arbitrariness. In the rodent example of the previous paragraph, the points used on the outline are consistent with G19 definition of semilandmarks (as they are defined in relation to end points which are ‘homologous’ landmarks). However, the points are placed automatically and ignore many other landmarks that could have split the outline in a series of ‘bone-specific curves’. Thus, the uncertain distinction between semilandmarks and pseudolandmarks may not allow a simple discrimination between what, following G19, seems acceptable for comparative evolutionary studies (i.e., semilandmarks) and what is more problematic (i.e., pseudolandmarks).

To better stress this issue, I take inspiration from the introduction to semilandmark methods of Gunz and Mitteroecker (2013). To start, they remark that digitizing semilandmarks may not be so easy, especially on surfaces where (p. 105) “one approach is to measure a mesh of surface semilandmarks on a single template specimen, and project this mesh onto all other forms in the sample”, which is a type of automatic placement as in G19 definition of pseudolandmarks (p. 671: “sampling uniformly from a surface mesh”). Then, to illustrate the technique (Gunz and Mitteroecker 2013) say that, apparently regardless of whether some semilandmarks potentially are on different bones (as not unlikely at “high-density”, especially for those close to the sutures in species with large anatomical differences), they “treat the outer shell of the braincase in its entirety as homologous between” a human and a young gorilla. They do acknowledge that this decision may not be appropriate for all research questions (e.g., when focusing on development compared to supra-generic analyses), but are vague about how they decide that this is fine for “many comparative purposes”. For instance, if treating the vault as a whole and ignoring boundaries between bones was appropriate for their case study looking at doming of the braincase in hominin evolution, is it also necessarily appropriate in the rodent case study, where the focus was on the association between the evolution of morphology and genomic change and the whole cranial outline was analysed as a single structure (again regardless of boundaries between bones)? Overall, this shows again that, even with *semilandmarks defined as in G19, in reality, these points may have aspects that fit in the definition of pseudolandmarks, and viceversa.*

Having shown that the distinction between the two types of ‘non-homologous’ points is operationally not as clear as one wishes (which is why I typically avoid using it), I find that the cases cited as problematic, by G19 (p. 671), for landmark-only analyses, which are “caecilian amphibians” with their “large degree of variation in bone presence and suture patterns” or “birds” with their “high degree of bone fusion”, may not be too dissimilar in their implications from the rodent study which I mentioned. The fact that some bones may be missing, and/or sutures not evident, cannot be a straightforward and convincing reason to act as if anatomical, developmental, and evolutionary regions had no identity and can be blanketed with points regardless of biological meaning, as in the ambitious statement in the Introduction (G19, p. 671)) on “the promise these methods offer for quantifying regions that are poorly characterized by use of only discrete landmarks, due to the lack of unambiguous homology across specimens or the presence of large areas without any appropriate structures at which to place landmarks”. If one takes this literally, it could be argued that we are at the frontier of a no-limit “high-density” GMM future, where an evolutionary comparison of caecilians and, say, earthworms

and cucumbers is perfectly sensible (Fig. 4): cover them in a “high-density” set of points, get shape coordinates with Procrustes GMM (plus sliding, if you wish), maybe decide the optimal density with W18 method, and finally summarize shape using a principal component analysis or a cluster analysis. However, do not be too surprised if caecilians group closer to earthworms and cucumbers than, say, to frogs.

Later in the paper (p. 674), the authors moderate the initial statement on “high-density” being a promising almost ‘homologous-free’ solution and acknowledge: “although curves may capture much of the morphological variation of the full landmark, curve, and surface dataset for many structures, they can be problematic and inapplicable in some of the most interesting, highly variable regions, particularly as comparisons expand across increasingly disparate taxa”. But is this really the only issue one should have before deciding whether “high-density” is a simple fix to a “lack of points of unambiguous homology”? It seems that, for G19, “high-density” is fine, but only as long as the curve or surface is homologous, which, in their examples, means placing points along or within the same bone (and maybe even not exactly the same as, in fact, the vault is a composite bone). This is definitely more reasonable, but still unsatisfactory because, in reality, even within homologous structures, such as a specific bone, the placement of semilandmarks is based on the same but less obvious type of reasoning as in the evolutionary absurd comparison of non-homologous traits (as in the made-up example on monotremes, as well as the real one on rodents). This is because, within, for instance, the frontal bones of a mammal, we do not actually know if semilandmark number 1 (or 2 or any of them) is actually mapping on the same ‘homologous’ subregion of these bones in all specimens or species: it could be that a small bump on an Indian elephant cranium is a completely new auto-apomorphic character that was not present in mastodonts and mammoths, and, therefore, even as an approximation, semilandmarks mapping on that frontal bone bump cannot map on the same feature in other species, as this is simply missing.²

Shifting from biological homology to geometric homology [“point-to-point, curve-to-curve, or surface-to-surface correspondence”, p. 104, (Gunz et al. 2005; Gunz and Mitteroecker 2013)] does not solve this problem either: the bump is still not there and, in a biological evolutionary study, we are not comparing like with like (as it happens with the more extreme but analogous ‘no-limit GMM’ example of caecilians, earthworms, and cucumbers, all geometrically similar when measured by a dense set of points, even if their shape similarities are at best the product of convergence).

² A similar hypothetical example can be found in Polly (Polly 2017) and why GMM is ill suited to describing the appearance of new structures is discussed in Oxnard and O’Higgins (2009).

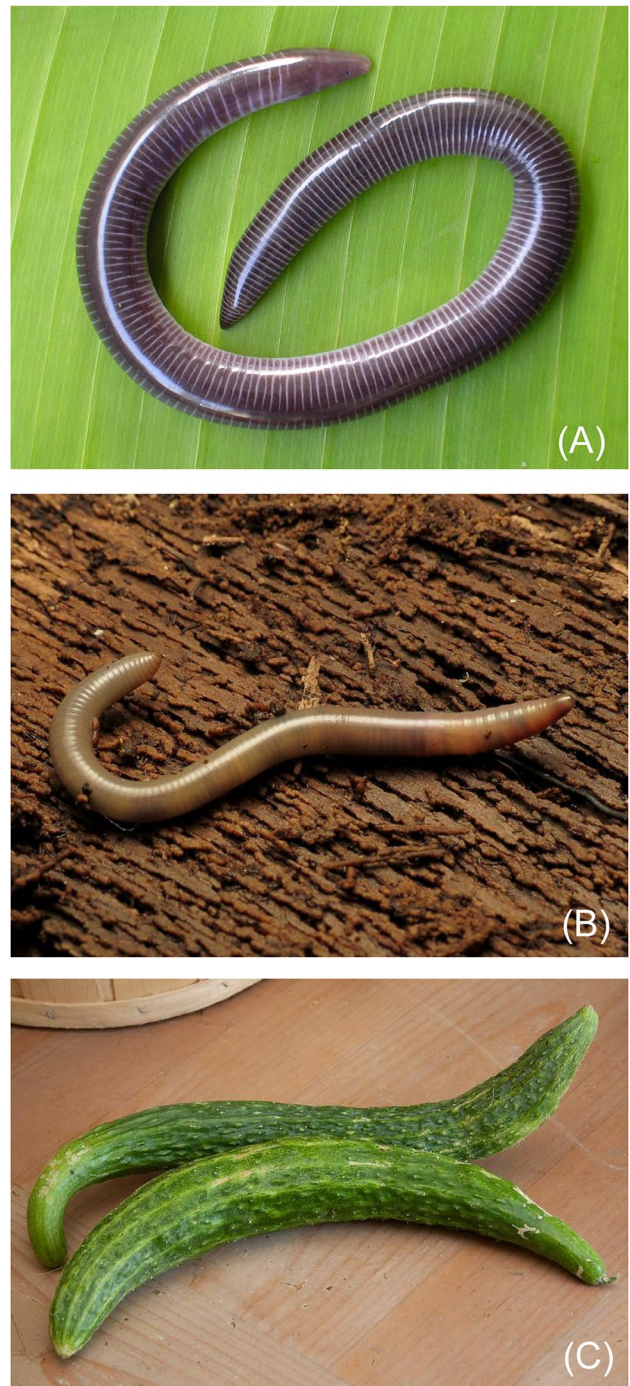


Fig. 4 A provocative suggestion (to be taken with a smile) for evolutionary comparisons with ‘no limits- “high-density” GMM’: **a** Narayan’s caecilian, **b** earthworm, and **c** cucumbers (images, respectively, from: **a** https://commons.wikimedia.org/wiki/File:Uraeotyphlus_narayani_habitus.JPG, by Venu Govindappa under Creative Commons Attribution 3.0 Unported; **b** <https://www.flickr.com/photos/treegrow/38799468075>, by Katja Schulz under Creative Commons 2.0 Generic; **c** <https://www.flickr.com/photos/twisted-string/7507861822>, by twistedstringknits under Creative Commons 2.0 Generic)

Besides, where is the universal demonstration that morphological evolution follows geometry? Sometimes it might, sometimes it might not, but, even in the apparently convincing case, used to promote sliding semilandmarks using the minimum bending energy criterion, that correctly depicts shape changes from a shorter to a more elongated rectangle showing parallel transformation grids, unlike the twisted ones of equally spaced points [Fig. 2.1, (Gunz et al. 2005)], the demonstration is invalidly tautological. It assumes the identity of biological homology and geometry, but the truth is, again, that we do not know if this is the case. Slid semilandmarks could correctly place the same semilandmark (say, number 10) in the same corner of the two rectangles, but, if we later find out that the second rectangle was a bone made from the fusion of two smaller ones, that corner could be measuring totally non-homologous features in spite of the elegant maths and beautiful geometry. Thus, one simply cannot conclude ((Gunz et al. 2005), p. 25) that “semilandmarks like these [i.e., slid using the minimum bending energy criterion] can then be treated as homologous, without artifact”.

Accuracy in ‘homology’, and the correspondence of morphometric descriptors, is an issue that, when one looks into the details (Oxnard and O’Higgins 2009), might affect most of GMM (including apparently precise landmarks), but, as anticipated, there is a range of severity from ‘homologous’ landmarks to semilandmarks, all the way to the extreme of completely “homology free” phenetic analyses. Semilandmark methods have mitigated but not solved some of the limits of older, and often strongly criticized (O’Higgins 1997), techniques for the analyses of outlines (Perez et al. 2006; McCane 2013). As I have already stressed using examples, even the rhetoric of glossing over these issues for semilandmark data by insisting on the homology of the curve or surface in its entirety (and/or in terms of geometric correspondence) does not avoid the problem: the ambiguity in homology is simply shifted within a suture or bone (or any other structure described by semilandmarks). Besides, from a purely analytical perspective, the shape distances that capture the morphological relationships are a function of the coordinates of the points and not of the curve or surface as a whole. Thus, where the points are, not just their number or density, is fundamental to establish accurate shape relationships in terms of biological equivalence among specimens in a sample.

Once all these problems are honestly acknowledged, one can probably go on using semilandmarks, when they are really important in relation to its study aims. In this context, W18 approach to decide a parsimonious set of points becomes potentially informative. However, it may be useful for precision, while achieving a degree of parsimony in the number of morphometric descriptors, but does not by default “improve the accuracy of morphological characterization” (p. 1, (Watanabe 2018)). Using another analogy, saying that

the configuration with the largest number of points (landmarks and semilandmarks) represents accuracy in an evolutionary study is like claiming that a tree based on all possible traits, as in old fashion phenetics (Sokal and Sneath 1963), is the true phylogeny. Occasionally that may be true, but more often than not the phenetic tree will be a very poor proxy for phylogeny, although its topology may be robustly supported (like the “stationarity” peak of W18) and less unstable than estimates based on fewer characters. Yet, a tree built using a careful selection of phylogenetically informative characters and a valid evolutionary model (Felsenstein 2004) is more likely to produce answers closer to the truth. Again, precision and accuracy are not the same.

Why I measure is more important than how much I measure

The importance of relating the measurements (landmarks, semilandmarks, or traditional distances) to the study aim is crucial both in general, as well as in the context of “high-density” analyses, as already but briefly mentioned. Measuring a lot (many landmarks and/or semilandmarks) not only is not a synonym of accuracy, but does not even guarantee statistical power, because much of what is measured could be irrelevant or totally inappropriate for the specific study hypothesis. For instance, using the beautiful example of Oxnard and O’Higgins (2009, p. 86): “in a study of bat and bird wings if one is interested in function, landmarks at wing tips and along the leading and trailing edges may be functionally equivalent; they might embody the question in being related to functionally relevant aspects of form. However, these landmarks may lie on structures that are not equivalent in other ways; for a study of growth or evolution, alternative landmarks may be the most suited ones”. Yet, according to G19 (p. 673), “landmark data alone are insufficient to fully characterize morphological variation for many datasets” and for “characterizing variation, the addition of curve sliding semilandmarks alone is a vast improvement on landmark only analyses”. This statement fails to relate what is being measured to the hypothesis being investigated and taken literally is, again, misleading, as it implies that a set of densely spaced points along the outline of bats and bird wings is appropriate and always superior to a few anatomically corresponding landmarks capturing the proportions of the homologous bones, which make up these evolutionarily convergent structures. Thus, G19 conclusively but inaccurately state (p. 669): “high-density geometric morphometric data exceed traditional landmark-based methods in characterization of morphology and allow more nuanced comparisons across disparate taxa”.

Readers may think that no evolutionary biologist, asking an evolutionary rather than a functional question, would be so naive as to compare bird and bat morphology using

a series of equidistant points (slid or not) on their wings, but the rodent cranium example I made should warn that this is not as remote a possibility as it might look. However, let us be less naive and see if, using a well-designed mix of landmarks and semilandmarks, I can increase the amount anatomical details in a comparison. For instance, in a study of growth, together with ‘homologous’ landmarks, I could place the semilandmarks within bones, such as along the shaft of the radius. However, I would argue again that semilandmarks may add something but, rigorously speaking, are at best trying to provide a geometric and mathematical answer, that may or may not be a good proxy for the underlying biological model. Useful as they appear to and sometimes might be, semilandmarks represent ‘known unknowns’. Let us consider another hypothetical scenario in which, in the future, developmental biologists find out that the radius develops asymmetrically, so that the ossification center grows faster towards the humerus in birds but does the opposite in bats. As this knowledge was not available at the time of our GMM study, how likely is it that the semilandmarks on the shaft are mapping on developmentally corresponding regions in the two taxa? Even if the radius is clearly homologous and the geometric correspondence is accurate, the roughly central set of semilandmarks will be to one side of the ossification center in birds and to the opposite one in bats: this shape change, in relation to evolution and development, has gone missing.

Similar considerations have been made by Klingenberg (2008) and Oxnard and O’Higgins (2009) in relation to the semilandmark approach of Polly (Polly 2008b) other putatively ‘homology-free’ methods (Polly 2008a). Thus, semilandmarks do offer a simple and potentially useful expedient to discretize curves and surfaces with no clear landmarks. And those curves and surfaces may be homologous, but the shape variables produced by the semilandmarks are a function of the relative distances between the corresponding points (Klingenberg 2008), and the correspondence is, by definition, unclear for semilandmarks and potentially inaccurate, regardless of a possible mathematical ‘massage’ of those points (such as by sliding (Gunz et al. 2005; Gunz and Mitteroecker 2013)).

On this, and more generally on the argument about ‘why vs how much to measure’, a reviewer suggested another interesting point of view. For him, the W18-G19’s approach is flawed, because their test of alternative landmarking density presupposes that one wants to measure the precise shape specified by the high-density scheme that they use. Unless fine-scale features covary precisely with larger scale features, then altering the density of surface sampling simply changes what aspects of shape one measures. Just as their lower density sampling does not capture the same shape as their original configuration, neither would a higher density sampling. Crucially, if one has a hypothesis about the overall

shape and relationships of bones independent of their surface features, then a low-density scheme will be better able to test that hypothesis than W18/G19’s high-density schemes, because the surface details would likely be noise that weakens the statistical power to test the hypothesis. In general, any test of whether one landmarks scheme is “better” than another is likely to be flawed without first specifying precisely which aspects of shape one needs to measure (Oxnard and O’Higgins 2009). Thus, the reviewer wrote: “One could cover each bone in high density semilandmarks or ... put a small number of fixed landmarks at junctions between each bone and another. G19 argue that the former is “better” than the latter. ... Neither scheme is better or worse, they simply measure different things. The simple landmarking scheme captures the basic shape of the bone and its topographical relationship to other bones. The high-density scheme captures the convexities and concavities of its surface, and perhaps even its texture. If one is studying, for example, the contribution of brain development to skull shape, the simple scheme will be “right” because the high-density scheme contains a mixture of properties that are relevant to the question (bone location on the skull) with those that aren’t (bone surface texture, horns, armouring, etc.). Conversely if one is studying the relation of skull morphology to sexual dimorphism, the texturing is likely to be more important than the bone placement itself. ... If we know we want to measure covariances in surface ornamentation between bones, we can use G19’s downsampling procedure to show that, if we decrease the density of points, we lose the power to measure those covariances; and conversely, if we know we want to measure the basic topographical contribution of the bone to the overall skull, we could use the same strategy to up-sample the density and eventually show that the “noise” of the surface texture overpowers the signal we want to measure. In other words, ...the downsampling (or upsampling) scheme cannot show us that a configuration is “better” in an absolute sense, [but] only that alternatives are worse at measuring something that we a priori want to measure.”

The reviewer’s view ties well with my conclusion to this section on “The Bad” in G19, in which I wish to stress again what my criticisms are about: they are against the notion that ‘high-density’ data are generally better than fewer well-considered homologous landmarks in studies of development and evolution, but *not against a cautious and parsimonious use of these techniques*, which I employed myself more than once, including a totally and truly ‘homology-free’ GMM study with no landmarks at all (Ferretti et al. 2013). The same holds for W18 approach, which can be fruitfully used to explore the sensitivity of results to the number of points, but not to test accuracy. In both cases, semilandmarks and W18 approach, one should be clear about what can and cannot be said using a certain type of data and methods. Indeed, with due acknowledgement of potential limitations

and inaccuracies (Rohlf 1998; O’Higgins 2000; Adams et al. 2004, 2011; Klingenberg 2008, 2013; Oxnard and O’Higgins 2009; Viscosi and Cardini 2011; Cardini 2019), using semilandmarks, and parsimoniously selecting their number and position (Gunz and Mitteroecker 2013), can be of help when there is no other option [e.g., (Ferretti et al. 2013)], as well as for taxonomic or forensic identification [e.g., (Schlager and Rüdell 2015)], biomechanics (O’Higgins et al. 2011), to reconstruct fossils (León and Zollikofer 2001; Gunz et al. 2009) and in many other cases. However, invalid tautologies should be avoided, and “what should I measure?” should be “sensibly responded to by another [question:] what is the hypothesis you are testing? ...[thus,] measure according to the question rather than relying on subsequent statistical analyses and interpretations to disentangle” (Oxnard and O’Higgins 2009, p. 87).

‘The not-so-accurate’

Too early for strong statements on either side

I conclude with ‘The not-so-accurate’. G19 summarized my 2019 paper on spurious results in TEMI because of the superimposition by saying (p. 671): “It has been recently asserted that this effect may be exacerbated in larger geometric morphometric datasets, such as those generated through the application of semilandmarks, although such an effect was not demonstrated, and assumed that the effects would reduce the ability to detect biological modularity in data (Cardini 2019). Second, and more specifically, it has also been asserted that closely packed semilandmarks may falsely inflate the pattern of modularity ... For these reasons, it has been suggested that “big data” is not necessarily better data when it comes to geometric morphometric analyses, especially analyses of phenotypic integration and modularity (Cardini 2019)”. It is inevitable that, when one summarizes a complex study, most of the complexity and nuances are lost. This is everyone’s problem, and I am sure I missed many subtleties in both W18 and G19, the latter having likely faced also the problem of compromising among views in a multi-authored paper. However, as my 2019 article provided both the stimulus for the study by G19 and the main context in which they interpret their findings, it is important to accurately reference that initial work. The summary of G19 mixes two issues, that have relationships but are better considered first separately: one is the problem of spurious results in TEMI; the second is the effect of semilandmarks on TEMI.

I start with the inflation of type I error rates in TEMI using the ‘within a configuration approach’ (i.e., analyses of modularity/integration with modules obtained by splitting the Procrustes superimposed full configuration). The first

sentence says that I asserted that the detection of modularity may be reduced using “big data” with many semilandmarks. However, the second sentence says that I also asserted that closely packed semilandmarks inflate modularity, which is confusing as it seems to contradict the previous statement. This is partly clarified later (p. 680), but now the distinction is between “proximal semilandmarks”, whose covariance (increased by the superimposition and sliding) may spuriously inflate modularity, and landmarks, which “may suffer from boundary bias, exaggerating the apparent integration of those elements”. On the same page, the potentially opposite effect of landmarks and semilandmarks is stated again: “both approaches suffer from statistical artifacts due to the nature of the data collection approach and may have opposing biases in reconstructing trait integration and modularity”. Thus, even if moderated by the verb “may”, the authors suggest multiple times that the consequences of using one or the other type of anatomical point (with the different mathematical treatment and higher density of semilandmarks) might be opposite, with landmarks inflating integration and semilandmarks spuriously suggesting modularity.

In fact, my conclusions were cautious and more nuanced: “The study, although preliminary and exploratory in nature... indicates an avenue for future research ... [and] 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” (from the abstract of Cardini 2019, p. 90). As in the introductory summary of my study by G19, I also had to squeeze a range of complex results and considerations in a few sentences. However, I made no strong conclusive assertions and stressed: a) the preliminary and exploratory nature of my study, as well as the need for more work; b) the importance of cautious applications of TEMI; c) the possibility (“might”) that the issues may be more serious using semilandmarks. In more detail:

1. By extrapolating results and interpretations from their context, G19 might have created a misleading impression of strong claims about what I suggested may or may not happen in relation to the problems with Procrustes shape data in TEMI. In fact, I did the opposite and my explanations were openly uncertain from the title of the paper (“Integration and modularity in Procrustes shape data: is there a risk of spurious results?”) and the abstract (see above) to the paragraph introducing the main conclusions (p. 102: “Bearing in mind its exploratory nature, the limited number of tests examined and a main focus on type I errors ... three main messages can be taken from this study, that might hopefully stimulate

- future investigations”). Thus, there are no strong assertions
2. G19 (p. 671) reported, citing my work, that “questions have been raised about their [semilandmarks/pseudolandmarks] necessity and applicability for the study of phenotypic integration”. Indeed, I raised questions on the effect of slid semilandmarks on TEMI, and these are not answered conclusively by the simulations of G19, as well as by their small set of examples of congruent findings with or without semilandmarks. However, I did not suggest giving up on either TEMI or semilandmarks. In fact, I urged caution but wrote, in the third of my three main conclusions (p. 103): “Clearly, this does not mean that one should never use semilandmarks ... However, it does suggest that there is no guarantee that more points necessarily lead to increased accuracy and, if and when semilandmarks are really crucial, one should also acknowledge their limitations and potential issues”. Thus, I said back then, as I repeated in this paper, that semilandmarks have *pros* (allow measurements where no landmarks are available) and *cons* (‘homology’ issues and proliferation of variables); can be used when they are important (with clear acknowledgement of the limitations); but cannot be assumed a priori to be more accurate and, as more general with any landmark configuration, their use should be functional to the specific hypothesis. This definitely means that they are not better ‘by default’, but they could be better, although this cannot be assessed by the tautological analysis of W18-G19, that only addresses precision and has nothing to do with accuracy
 3. What the effect of semilandmarks might be in the specific context of TEMI is definitely the most complex issue and, therefore, the one I will discuss in more depth. On this, I did say that semilandmarks “might” make the problem with spurious results more serious, but did not say that this happens for sure and offered an additional example of the complexity of this problem in the very last paragraphs. This example came immediately after my third conclusion, saying that slid semilandmarks are not the same as standard landmarks and that sliding does alter the covariance structure in non-trivial ways with potentially important consequences on results. Thus, using an example configuration of landmarks and semilandmarks and simulated isotropic variation, I showed that the choice of the mathematical treatment of semilandmarks can lead to inaccurate and incongruent results in exploring integration using the ‘integration/disintegration’ approach (Bookstein 2015), with both unslid and ‘Procrustes-slid’ semilandmarks correctly suggesting isotropic noise, and minimum-bending energy-slid semilandmarks misleadingly suggesting ‘disintegration’ (i.e., the opposite of integration). This

is the context in which my concerns on the complex effect of sliding semilandmarks on TEMI (expressed as a possibility: “issues might be even more serious”, p. 90) should be understood

As rightly argued by G19, and indeed already stated in my own 2019 paper, isotropic variation is not what we normally find in biological data. Thus, we cannot exclude that the inaccuracies in TEMI are inflated and we might hope that, with a strong ‘real’ covariance, the effect of the superimposition (plus a possible sliding of semilandmarks) will be less dramatic. Yet, there is an effect, whose importance will vary from case to case in ways that are not easy to predict. This is why, it is unwise to extrapolate beyond the specific configurations and set of parameters used in simulations, and analyses of example datasets, in both my work (Cardini 2019) as well as in G19. Especially in microevolutionary analyses (but probably not exclusively in these³), with smaller differences among specimens, the impact of sliding on real data could be particularly severe (Perez et al. 2006). Figure 3 shows superimposed configurations from a sample of adult yellow-bellied marmot hemi-mandibles. Instead of showing the scatter of points around the sample mean (as in Fig. 2), I computed here the main vectors of variation at each landmark and semilandmark. Landmarks show very similar directions of variation, but semilandmarks show a range of scatters depending on whether they are superimposed without sliding (A) or slid using either the minimum-bending energy (B) or minimum-Procrustes distance (C) criterion. Are these differences just a visual impression? Probably not: if A is compared to B and C, respectively, the matrix correlations of pairwise Procrustes shape distances are 0.830 and 0.884 and those between the corresponding variance–covariance matrices are even lower: 0.778 and 0.631. The total sample variance (the sum of the variances of all shape coordinates) is, respectively, 0.00263 (A), 0.00224 (B), and 0.00145 (C), which means 15% and 45% smaller, depending on the choice of sliding, compared to the superimposed unslid data. If one computes *variance at each landmark/semilandmark* (which

³ In fact, a reviewer argued that, even if most of the time the ‘noise’ might be relatively more important in microevolutionary studies, where the ‘signal’ tends to be smaller, there might be interesting exceptions: “Microevolutionary differences tend to be small. That said, depending on the shapes of the objects, the problem could be bigger in macroevolutionary studies if one takes the simulations in G19 at face value: large differences in points that are far from the center of the object—for example, tips of bird beaks—will strongly affect the position of the centroid and, therefore, the artifactual covariance introduced by Procrustes. That effect will be even more pronounced if the beak has more surface area than the cranium and semilandmarks are placed with equal density in both regions”.

is biologically totally meaningless, as it depends on the superimposition, and used here only to compare results), the correlation of landmark variances between A and, respectively, B and C is 0.688 and 0.799, which again confirms the differences (here in magnitude) seen by eye, as well as using matrix correlations. That there is a substantial modification of the variance–covariance structure is indicated also by the variability in the variance accounted for by the first principal component of shape, which is 22.6% (A), 36.0% (B), and 31.8% (C). All these descriptive statistics can be rapidly and easily obtained, and readers are encouraged to try them on their own semilandmark data: sometimes, they will find high congruence, and sometimes not; but any difference is always purely a consequence of the arbitrariness in the choice of sliding.

Before moving to what the current evidence is regarding the practical importance of Procrustes and sliding on TEMI, I briefly go back to the apparently contradictory claims of semilandmarks both worsening the detection of modularity and inflating modularity (p. 671 of G19). Neither of these was a simple conclusion of my 2019 paper, as I made clear that predicting the direction of the bias introduced by the superimposition may not be straightforward. In this context, I speculated whether (p. 102) “the ‘contiguity effect’⁴ induced by a common superimposition might explain ambiguous results suggesting both integration and modularity”; and added: “This is definitely an issue that requires more study to be better understood ... and indeed the impact of these issues may vary from case to case and generalizations might be difficult to make”. Thus, I warned users not to assume that, because the superimposition creates covariance in isotropic data, one can confidently conclude that a finding of significant modularity, despite this bias, guarantees the accuracy of results and makes the evidence for modules stronger”. This may or may not be true depending on the data, as one cannot exclude that covariance within modules is increased more than between them, which seemed to me the only explanation for the examples in which I found spuriously both significant integration as well as significant modularity. Assuming that my speculation was correct, the “contiguity effect” may or may not happen more often with semilandmarks than landmarks, and, although I suspect that it might, indeed, be more common with semilandmarks, this is far from sure and will depend on their relative position, number, and density, as well as the strength of the real covariance.

⁴ (i.e., the likely stronger covariance between contiguous landmarks, such as those within a ‘module’).

More agreement than disagreement?

Having clarified what I wrote on semilandmarks and TEMI, with explicit acknowledgements of the huge uncertainties of all those findings, I can proceed to discuss what G19 discovered on this issue. The paper contains some potentially misleading or even contradictory statements (more below), which does not help with interpretations. However, it seems that their conclusions on the effect of semilandmarks on TEMI are not that dissimilar from mine. G19 reported a generally good congruence in results of TEMI using a few example datasets with only landmarks or both landmarks and semilandmarks. However, they also stated (p. 680): “Despite supporting a modular pattern, the landmark-only dataset returned a CR much closer to one, indicating relatively more integration among modules ...”, which they assumed was mostly due to a lack of within-region landmarks in the cranial vault. Later (p. 681) they remarked again that landmark-only data suggest less modularity with allometry contributing to integration: “As before, the landmark-only analyses returned CR values closer to one, suggesting more integration than the analysis of the full dataset, and removing allometric effects resulted in reduced CR values, supporting a more modular pattern”. Since these specific results are discussed in terms of what landmarks-only suggest and why this may vary (also in the abstract), and probably also because of a misleading statement in the last paragraph of the previous section on simulations in G19 (see below), one almost has the impression that landmarks are less reliable to infer modularity/integration than semilandmarks. However, this is clearly not something we can say, since we do not know the true pattern of modularity/integration in real data, and one could have presented these same results with an emphasis on semilandmarks, surprisingly unaffected, for instance, by allometry, which sounds almost counter-intuitive “because allometry is a known factor contributing to integration” (p. 125–126, (Klingenberg 2016) and references therein). Similarly, the cranial vault, with its frequently fused bones, could be truly more strongly integrated, making results from landmarks more credible than those from semilandmarks. Yet, I am not stating that this is the case: once more, we simply do not know and, indeed, assuming that measurements are accurate in relation to TEMI, it could well be that there are truly different answers depending on what is measured.

In the very final section of the paper, G19 finally clarifies that, as in my interpretation of their results, we cannot really say whether landmarks or semilandmarks get the right answer in TEMI: “... analyses incorporating semilandmarks may be less influenced by boundary bias and allometric effects, which may exaggerate the degree of integration across regions in landmark-only analyses, while analyses of sliding semilandmark may exaggerate within-region

integration and between-region modularity” (G19, p. 681). This is not unlike what I suggested (as a speculation) in my 2019 paper, when I warned that the increased overall covariance due to the superimposition cannot be simply used to argue that significant modularity must be accurate (as it happens despite the ‘extra’ covariance due to the Procrustes fit). Thus, I wrote, not only one can find spurious or inaccurate evidence for integration but also for modularity (p. 102): “for instance, if the number of points within a module is much larger than that marking the boundary of two contiguous modules, the superimposition (plus the possible sliding of semilandmarks) could increase intra-modular covariance even more strongly than inter-modules, thus spuriously inflating the evidence for modularity”. If this is compared to the main conclusion of G19, the only one which is truly supported by the data, their results do not seem *contra* Cardini (2019). In fact, they support my original guess on this, as they acknowledge that, if landmarks-only may inflate integration, semilandmarks could do the opposite and spuriously increase modularity. Nevertheless, although supported by G19’s findings, I stress that the evidence for the ‘contiguity effect’ (Cardini 2019), and its relation to the density and mathematical treatment of the points, is still preliminary. It may turn out that an opposite bias of landmarks vs semilandmarks is really a prevalent finding in TEMI, but I suspect that it is not a sharp and simple dichotomy: there might be cases where this is reversed; and there are definitely examples, at least using simulations with isotropic noise, where one can find both spuriously significant integration as well as modularity with both landmarks-only or landmarks plus semilandmarks (Cardini 2019).

Everyone agrees that the superimposition (plus possible sliding) alters the covariance structure, which is what is tested in TEMI, and that this happens in complex ways that make tests potentially inaccurate. Then, if this leads to inevitable uncertainties about the strength of the evidence for either integration or modularity, how can *G19 state as facts* (p. 680) that “... sliding semilandmarks ... improve the Procrustes fit relative to the “natural superimposition,” and thus increase the accuracy of recovering modules for biological structures”? Not only can we not say if sliding semilandmarks make the Procrustes superimposition more ‘biologically accurate’, since we do not know in the first place what the “natural superimposition” is, and semilandmarks were not even included in the simulations, but also, and more importantly, we cannot say that they increase accuracy in recovering modules. This is in contradiction to G19’s own findings in simulations, showing that “adding landmarks [not semilandmarks but still more points!] neither improves nor inhibits the recoverability of modules”. And it is even more in contradiction with their own main conclusions (presented in my previous paragraph) on the comparison of TEMI with or without semilandmarks, that

“analyses of sliding semilandmark may exaggerate within-region integration and between-region modularity”. It is either one (semilandmarks increase the accuracy in recovering modules) or the other (adding more points does not improve finding modules or even, possibly, semilandmarks inflate modularity).

Leaving the confusing statements of some of G19’s paragraphs aside, I think that, for the time being, we can safely conclude that there is no evidence that semilandmarks improve accuracy in TEMI. Possibly, in many cases, but most likely not all of them, they might, in fact, inflate the evidence for modularity (as hypothesized by Cardini (2019) and now potentially supported by G19). However, although very preliminary, there is definitely evidence, from both simulated data (Cardini 2019) and the real sample of Fig. 3, that sliding semilandmarks adds a further layer of complexity: this is because, even using identical configurations of points, the simple choice of the mathematical treatment of the semilandmarks further changes the variance–covariance structure of the data, with a difference that can be large enough to produce fairly low correlations of covariance matrices and contradictory conclusions on integration and modularity using Bookstein’s (2015) approach.

Conclusions

Overall, G19, like Cardini (2019), shows that the effects of the Procrustes superimposition, sliding semilandmarks, and the landmarking scheme, interact in complex ways but definitely alter the pattern of real variance–covariance. This happens with modalities that vary from case to case and do not allow simple predictions on the magnitude of the impact and whether more or less points is better. Thus, we can only stress that:

- great caution is required in interpretations of patterns of variance–covariance using both landmarks and semilandmarks;
- “high-density” GMM cannot ignore potential issues with inaccurate TEMI;
- there is definitely a need for more studies on the impact of the Procrustes superimposition, plus sliding (the extra layer of complexity), on GMM analyses of modularity and integration.

Therefore, as I emphasized both here and in my 2019 paper, the issue is complicated, and predicting the impact of the modification of the variance–covariance in TEMI using Procrustes shape data is not straightforward. The simulations of G19 using landmarks (but no sliding semilandmarks) suggest that, with a realistic scenario of variable directions in covariance, the results of GMM studies of modularity and

integration may be reliable despite inaccuracies. Indeed, as I wrote in the previous article (p. 98): “large samples [in my simulations] ... might have overemphasized the importance of small covariances introduced by the superimposition” and thus that “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”. Nevertheless, *the modification of the pattern of covariance by the superimposition is supported by all studies and well known since the early days of GMM* (Rohlf and Slice 1990; Rohlf 1998). This inevitably introduces a degree of inaccuracy in both statistical testing and estimates of parameters and effect sizes, with the magnitude of the errors specific to the dataset. *Thus, I doubt one can argue that this is “Not-Really-a-Problem” and that we are confident that “these potential issues are unlikely to obscure genuine biological signal”* (p. 669 G19). The first claim makes a good title, but denies the presence of a problem that we still know little about, except that it is very real. The second one is our hope, but seems rather premature as a conclusion and cannot be said based on a single study, neither G19 nor mine. I suspect that, besides the potential of GMM for studies of modularity and integration, as well as its limits and possible inaccuracies, stronger results in this field will come from an integrative evo-devo approach, mixing multiple lines of evidence (from genetics and embryology to morphometrics), as it is becoming customary in taxonomy (Schlick-Steiner et al. 2010; Padial et al. 2010). For now, however, I urge morphometricians to be open but cautious, and, on “high-density” GMM, I do suggest less tautology and more biology.

Acknowledgements I am very much in debt to David Polly for the stimulating discussions which we always have had and also for his wonderful review of this paper. A few other scientists could have been as balanced and positive as he was: his supportive comments are the best acknowledgement I could hope for. If David’s task, as an author of G19, was particularly challenging, I did not give for granted that also the second referee was equally positive. For this, I sincerely thank Paul O’Higgins, who also made a great number of useful comments and who, most importantly, first taught me to pay a lot of attention to the implication of the mathematical treatment of landmarks and semilandmarks. We may disagree on whether content is more important than form, but I am in debt to Ulrike Muller for her most helpful replies to my informal initial inquires on a possible response to the paper of G19. In *Integrative and Comparative Biology*, where G19 is published, was the obvious place to first explore the possibility of a comment, I am very happy that my paper will be out in *Zoomorphology*, a journal I have strong ties with and one that over the years has become a main venue for morphometric studies. Thus, for the excellent (as usual!) editorial work, I thank a lot both Andreas Schmidt-Rhaesa and Carmelo Fruciano. Finally, I am most grateful to a number of morphometricians and evolutionary biologists who, over the years, have provided feedback (and a good deal of arguing!) on some of the problems I discuss in this paper, and especially to Sandro Minelli, Charles Oxnard, Chris Klingenberg, Mike Collyer, Philipp Mitteroecker, Carmelo Fruciano (again!), and Sarah Elton: with some of them, we are definitely on

the same wavelength; with others, I fear we will have to agree that we may disagree!

Funding The study has not received any funding.

Compliance with ethical standards

Conflict of interest I declare that I have no conflict of interest.

References

- Adams DC, Cardini A, Monteiro LR et al (2011) Morphometrics and phylogenetics: Principal components of shape from cranial modules are neither appropriate nor effective cladistic characters. *J Hum Evol* 60:240–243
- Adams DC, Collyer ML, Kaliontzopoulou A, Sherratt E (2017) Geomorph: Software for geometric morphometric analyses. R package version 3.0.5. <https://cran.r-project.org/package=geomorph>
- Adams DC, Rohlf FJ, Slice DE (2004) Geometric morphometrics: ten years of progress following the ‘revolution’. *Italian J Zool* 71:5–16
- Adams DC, Rohlf FJ, Slice DE (2013) A field comes of age: geometric morphometrics in the 21st century. *Hystrix Italian J Mammals* 24:7–14
- Bookstein FL (2017) A newly noticed formula enforces fundamental limits on geometric morphometric analyses. *Evol Biol* 44:522–541. <https://doi.org/10.1007/s11692-017-9424-9>
- Bookstein FL (2015) Integration, disintegration, and self-similarity: characterizing the scales of shape variation in landmark data. *Evol Biol* 42:395–426. <https://doi.org/10.1007/s11692-015-9317-8>
- Cardini A (2019) Integration and modularity in procrustes shape data: is there a risk of spurious results? *Evol Biol*. <https://doi.org/10.1007/s11692-018-9463-x>
- Cardini A, Loy A (2013) On growth and form in the computer era: from geometric to biological morphometrics. *Hystrix Italian J Mammal* 24:1–5. <https://doi.org/10.4404/hystrix-24.1-8749>
- Cardini A, O’Higgins P, Rohlf FJ (2019) Seeing distinct groups where there are none: spurious patterns from between-group PCA. *Evol Biol* 46:303–316. <https://doi.org/10.1007/s11692-019-09487-5>
- DeQuardo J, Bookstein FL, Green WDK et al (1996) Spatial relationships of neuroanatomic landmarks in schizophrenia. *Psychiatry Res Neuroimag* 67:81–95. [https://doi.org/10.1016/0925-4927\(96\)02733-3](https://doi.org/10.1016/0925-4927(96)02733-3)
- Felsenstein J (2004) *Inferring Phylogenies*. Sunderland, Massachusetts, Sinauer Associates, Incorporated
- Ferretti A, Cardini A, Crampton JS et al (2013) Rings without a lord? Enigmatic fossils from the lower Palaeozoic of Bohemia and the Carnic Alps. *Lethaia* 46:211–222. <https://doi.org/10.1111/let.12004>
- Goswami A, Watanabe A, Felice RN et al (2019) High-density morphometric analysis of shape and integration: the good, the bad, and the not-really-a-problem. *Integr Comp Biol* 59:669–683. <https://doi.org/10.1093/icb/icz120>
- Gunz P, Mitteroecker P (2013) Semilandmarks: a method for quantifying curves and surfaces. *Hystrix Italian J Mammal* 24:103–109
- Gunz P, Mitteroecker P, Bookstein FL (2005) Semilandmarks in Three Dimensions. In: Slice DE (ed) *Modern morphometrics in physical anthropology*. Kluwer Academic Publishers-Plenum Publishers, New York, pp 73–98

- Gunz P, Mitteroecker P, Neubauer S et al (2009) Principles for the virtual reconstruction of hominin crania. *J Hum Evol* 57:48–62. <https://doi.org/10.1016/j.jhevol.2009.04.004>
- Kendall MG, Buckland WR, Institute IS (1957) A dictionary of statistical terms. Published for the International Statistical Institute by Oliver and Boyd
- Klingenberg CP (2008) Novelty and “Homology-free” Morphometrics: What’s in a Name? *Evol Biol* 35:186–190. <https://doi.org/10.1007/s11692-008-9029-4>
- Klingenberg CP (2013) Visualizations in geometric morphometrics: how to read and how to make graphs showing shape changes. *Hystrix Italian J Mammal* 24:15–24
- Klingenberg CP (2016) Size, shape, and form: concepts of allometry in geometric morphometrics. *Dev Genes Evol* 226:113–137. <https://doi.org/10.1007/s00427-016-0539-2>
- Kovarovic K, Aiello LC, Cardini A, Lockwood CA (2011) Discriminant function analyses in archaeology: are classification rates too good to be true? *J Archaeol Sci* 38:3006–3018. <https://doi.org/10.1016/j.jas.2011.06.028>
- de León MSP, Zollikofer CPE (2001) Neanderthal cranial ontogeny and its implications for late hominid diversity. *Nature* 412:534–538. <https://doi.org/10.1038/35087573>
- McCane B (2013) Shape variation in outline shapes. *Syst Biol* 62:134–146. <https://doi.org/10.1093/sysbio/sys080>
- Navarro N, Maga AM (2016) Does 3D phenotyping yield substantial insights in the genetics of the mouse mandible shape? *G3 Genes Genomes Genetics* 6:1153–1163. <https://doi.org/10.1534/g3.115.024372>
- O’Higgins P (1997) Methodological issues in the description of forms. Fourier descriptors and their applications in biology. In: Lestrel P (ed) *Fourier Descriptors and their Applications in Biology*, Cambridge University Press, Cambridge, pp. 74–105 <https://doi.org/10.1017/CBO9780511529870.005>
- O’Higgins P, Cobb SN, Fitton LC et al (2011) Combining geometric morphometrics and functional simulation: an emerging toolkit for virtual functional analyses. *J Anat* 218:3–15. <https://doi.org/10.1111/j.1469-7580.2010.01301.x>
- Paul OP (2000) The study of morphological variation in the hominid fossil record: biology, landmarks and geometry. *J Anat* 197:103–120
- Oxnard C, O’Higgins P (2009) Biology clearly needs morphometrics. Does morphometrics need biology? *Biol Theory* 4:84–97. <https://doi.org/10.1162/biot.2009.4.1.84>
- Padial JM, Miralles A, De la Riva I, Vences M (2010) The integrative future of taxonomy. *Front Zool* 7:16. <https://doi.org/10.1186/1742-9994-7-16>
- Perez SI, Bernal V, Gonzalez PN (2006) Differences between sliding semilandmark methods in geometric morphometrics, with an application to human craniofacial and dental variation. *J Anat* 208:769–784. <https://doi.org/10.1111/j.1469-7580.2006.00576.x>
- Polly P (2017) Morphometrics and evolution: the challenge of crossing rugged phenotypic landscapes with straight paths. *Vavilov J Genetics Breeding* 21:452–461
- Polly PD (2008a) Developmental dynamics and G-Matrices: can morphometric spaces be used to model phenotypic evolution? *Evol Biol* 35:83. <https://doi.org/10.1007/s11692-008-9020-0>
- Polly PD (2008b) Adaptive zones and the pinniped ankle: a three-dimensional quantitative analysis of carnivoran tarsal evolution. In: Sargis EJ, Dagosto M (eds) *Mammalian evolutionary morphology*. Springer, Netherlands, pp 167–196
- Rohlf FJ (2000a) On the use of shape spaces to compare morphometric methods. *Hystrix Italian J Mammal* 11:1–17. <https://doi.org/10.4404/hystrix-11.1-4134>
- Rohlf FJ (2000b) Statistical power comparisons among alternative morphometric methods. *Am J Phys Anthropol* 111:463–478
- Rohlf FJ (1998) On applications of geometric morphometrics to studies of ontogeny and phylogeny. *Syst Biol* 47:147–158
- Rohlf FJ, Marcus LF (1993) A revolution morphometrics. *Trends Ecol Evol* 8:129–132
- Rohlf FJ, Slice D (1990) Extensions of the procrustes method for the optimal superimposition of landmarks. *Syst Zool* 39:40–59. <https://doi.org/10.2307/2992207>
- Schlager S, Rüdell A (2015) Analysis of the human osseous nasal shape—population differences and sexual dimorphism. *Am J Phys Anthropol* 157:571–581. <https://doi.org/10.1002/ajpa.22749>
- Schlick-Steiner BC, Steiner FM, Seifert B et al (2010) Integrative taxonomy: a multisource approach to exploring biodiversity. *Annu Rev Entomol* 55:421–438
- Slice DE (1994) GRF-ND: Generalized rotational fitting of n-dimensional landmark data. Free software. Department of ecology and evolution, State University of New York, Stony Brook, New York, USA
- Smith GR (1990) Homology in morphometrics and phylogenetics. In: *Proceedings of the Michigan morphometrics workshop*. University of Michigan Museum of Zoology, Ann Arbor, pp 325–338
- Sokal RR, Sneath PHA (1963) *Numerical taxonomy. The principles and practice of numerical classification*. Freeman WH, San Francisco
- Viscosi V, Cardini A (2011) Leaf morphology, taxonomy and geometric morphometrics: A simplified protocol for beginners. e25630
- Watanabe A (2018) How many landmarks are enough to characterize shape and size variation? *PLoS ONE* 13:e0198341. <https://doi.org/10.1371/journal.pone.0198341>

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.