Palaeo-Math 101

A Picture May Be Worth 1,000 Landmarks

I began this series on mathematical methods in palaeobiology in 2004 with a consideration of regression methods that describe the relationship between two simple linear distance variables, illustrating these with data collected from a series of digital images of trilobites. Obviously our discussion has ranged rather broadly across the mathematical, data analysis, and taxonomic landscapes since then. For this essay — the last in the series — I'd like to return to the question of how we can extract data from the morphological information presented to us by organismal forms, usually (these days) through the medium of digital images.

Perhaps the most basic concept in the field of quantitative data analysis is that of the variable. Variables are the observations we make, usually in the form of measurements, on the set of specimens that comprise a sample of some population of interest. Variables come in many types and forms. If you take nothing else away from these *Palaeo-Math* essays please let it be that, for the purposes of describing and comparing specimens across a sample (not to mention drawing inferences about the population[s] from which the sample was drawn), the variables *are* the specimen; the only information any data analysis procedure has access to. As such, it is of the utmost importance that the variables we choose to represent our specimens be appropriate, both to the specimens in question and the hypotheses under consideration. If our variables do not meet these criteria in some reasonable and defensible manner, it is likely that any results we generate from their analyses will be compromised and/or (ultimately) questioned. For example, if we are interested in phenomena pertaining to the arrangement of component parts of the specimen relative to each other, and if the majority of those component parts are not located on the specimen outline, it makes little sense to restrict data collection to the geometry of specimen outlines. Similarly, if we are interested in questions pertaining to the general form of the specimens, and if this general form is best represented by the specimen's outline, it makes little sense to restrict data collection to a small set of landmarks located within the outline.

But what if we're not sure what parts of the morphology are important in terms of assessing similarity and/or difference relations among the specimens comprizing the sample? What if the variability in your specimens is such that the identification of corresponding point locations (landmarks, starting points for outline digitization) across the sample is simply not possible? Perhaps even more importantly, suppose we want to include as much morphological information about our specimens as possible and/or are uncomfortable with the idea of abstracting the specimens down to a (relatively) small set of variables a priori? Is there a way of handing this generalized problem within the context of a morphometric analysis?

Recently I ran into a situation of just this sort in the form of an student's MSc project. The student wanted to quantify the pattern of wing ornamentation of mimetic butterfly species morphs (Fig. 1) in order to compare wing colour variation to gene sequence variation. On the face of it this seemed simple enough, just determine which aspects of the ornamentation pattern were common to all specimens in the sample and base the measurement system on those using landmarks or outlines or landmarks + outlines where appropriate. After explaining these principles to the student I sent her away to design her measurement system and collect the data. Much to my surprise, she reappeared at my office door a few days later in a somewhat disturbed state having tried repeatedly to follow my instructions but failing at each attempt. After a bit of discussion it became apparent why. As can be seen in Figure 1, there are very few features of the wing ornamentation pattern present across Batesian mimic species and those that are present tend to be of quite a generalized character (e.g., orange region in the centre, black peripheral areas with light coloured spots and/or blotches). In essence, these mimics are somewhat approximate — not close — copies.

Figure 1. Morphological variation in a set of Batesian mimic butterfly species. A. *Danaus chrysippus* (model species). B. *Danaus chrysippus f. trophonius* (mimic). C. *Danaus chrysippus f. lamborni* (mimic). Note the similarity of all three colour morphs in terms of the general distribution of colours across the wings, but also the level of fine-scale difference in the number, sizes, and shapes of spot patterns. These inconsistencies make it difficult to use standard morphometric variables to characterize patterns of wing colour morph similarity and difference.

The standard response to a situation like this would be (sadly) to either abandon a morphometric approach entirely or collect a small set of landmarks or outline semilandmarks and analyse these rather than the colour blotch patterns. After all, if there are no form-related features common to the blotch patterns in all specimens there would appear to be no basis on which to compare them. At least that's how the logic would typically run. But it's obviously a false logic. The fact is entomologists, non-specialist collectors, and even butterfly predator species are able to make comparisons between the patterns of these butterfly wings. Indeed, that's the whole point of Batesian mimicry! The former two groups have been making such comparisons for (literally) centuries and it's been going on for millions of years in the case of the predators. If entomologists and butterfly predators can make comparisons between morphologies like these, quantitative morphologists and morphometricians should be able at least to make a stab at treating the same problem quantitatively. But how?

So as not to compromise the student's ability to publish her mimetic butterfly results, I'll shift at this point to an analogous image dataset I often use to introduce the concept of morphological variation to my students (Fig. 2). While this small collection of ladybird beetle images are decidedly not fossils, they serve to illustrate the full range of morphological variation in biological datasets better than a typical fossil dataset might. Regardless, the methods I'll develop below will apply equally well to images of fossil specimens where colour blotches on the specimen are often irrelevant.

Figure 2. Drawings of 24 ladybird beetle species (Family Coccinellidae) illustrating a range of body form and colour morphs. As with the Batesian mimic butterflies (see Fig. 1) the complexity of variation in body form, colour, the distribution, numbers, sizes and shapes of spots makes this sample difficult to characterize using standard linear distance, landmark, or semilandmark variables.

Among these drawings of ladybird beetles we see copious variation in body shape, leg & antennae pose, and both colouration and colour texture of the elytra, thoracic and head shields. Despite being all too typical of morphological datasets — open just about any well-curated museum drawer; this is what you'll see — none of the many tools of numerical data analysis and/or morphometrics I've discussed in this column can handle the problem of characterizing similarity/dissimilarity relations within this sample — or at least not obviously so.

The patterns we see in this figure seem too complexly structured to be characterized adequately by sets of linear distances, landmarks or semilandmarks. Yet, in the absence of taking some sort of measurement we cannot answer even the most basic questions about this sample. What is the mean form and colour pattern of this sample? Is morphological variation distributed continuously or discontinuously? Does the distribution of forms and colour blotches have a single or multiple modes? If the latter, how are those modes arranged relative to one another? More depressing still, these are the easy, descriptive questions. If we want to provide answers to more complex biological and/or evolutionary questions such as how these patterns of morphological variation covary with environment, geography, ecology, behaviour, genotype, phylogeny, etc, an ability to compare each of these forms to one another quantitatively is crucial. But if none of the tools, concepts, or methods we've discussed to date are up to this task, have we simply been wasting our time learning methods that might pertain to some small subset of morphological data-analysis problems, but are unsuited to the majority of routine morphological data-analysis situations with which we are confronted? Is it really true that the best we can do in this case is shrug our shoulders and go back to the qualitative inspection and appeals to 'authority' in deciding these issues?

Of course we can do better than this. Like so many problems in science — and especially in mathematics — this seeming intractable problem yields with surprising ease to a simple shift in the conceptual frame of reference. The standard way of approaching the morphology sampling problem is to find a series of topologically homologous, relocatable points across the forms of interest and record their coordinate positions. From these coordinate positions either linear distances or shape coordinates can be calculated.

However, if the image of a specimen has been digitized it has already been subdivided into a series of topologically homologous coordinate points — the pixel grid (Fig. 3).

Figure 3. Three example representation of ladybird beetle morphology and ornamentation pattern with a superposed 14 x 14 pixel grid.

These pixel locations represent a set of semilandmarks that exhibit a consistent spatial structure across the entire set of forms. So long as the dimensions and resolution of this grid remain constant for all images in the sample, and so long as the specimens are oriented in the grid in some reasonably consistent manner, this grid can be used to extract comparable descriptions of their form geometries. Quantitative descriptions of specimen variation extracted in this manner can be organized into many formats. If only the outline of the form is needed all pixels whose colour or grey level differs from that of the background can be assigned the same colour (usually black or white depending on the background).[1] If colour is not a parameter of interest the grid can be set to sample only the grey-level values of the individual pixels comprizing the image. If colour is of interest the sampling grid can be used to extract the red, blue, and green (RGB) values of each pixel. Figure 4 illustrates the effect of various grid sampling decisions on the representation of the first of the three ladybird beetle forms illustrated in Figure 3.

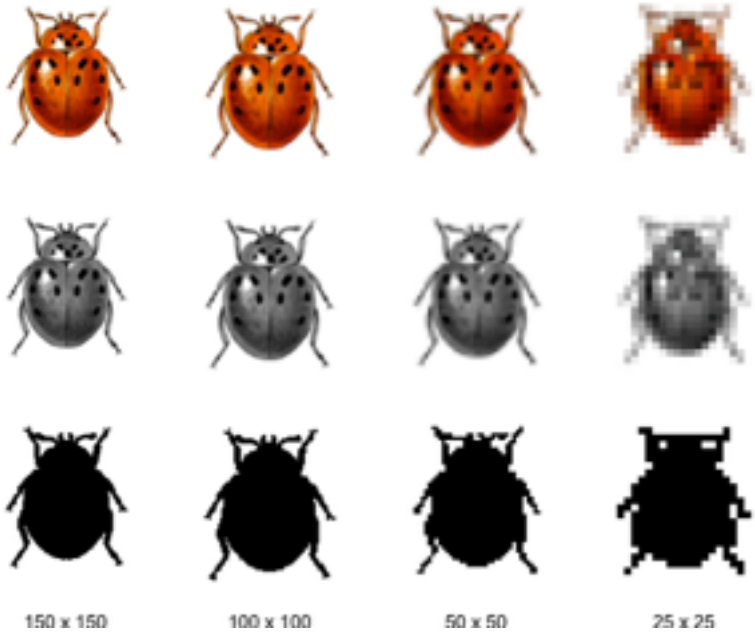

Figure 4. Ladybird beetle morphology represented as four different pixel grid sampling resolutions and three different colour formats: 8-bit RGB values (upper row), 8-bit greyscale values (middle row), and binary (1-bit) values (bottom row). Note fidelity with which detail is retained even at low spatial resolutions.

What is remarkable about this comparison of sampling resolutions and colour formats is the level of form and texture information content that's retained even at relatively low sampling resolutions. This suggests that, for the purpose of form and texture characterization, many pixels in a normal-resolution image (e.g., 72 pixels per inch) are redundant: the value of any particular pixel is much the same as the values of the pixels

[1] In this case the data would be more efficiently represented as a set of boundary outline semilandmark coordinates rather than a coordinate sampling grid.

immediately adjacent. In mathematical jargon this self-similarity is termed spatial autocorrelation. If possible the spatial autocorrelation of our raw morphological data should be reduced prior to during data analysis so that the effective dimensionality of the data-analysis problem can be optimized. But can we accomplish this reduction in the context of a digital image?

As an initial step we can decrease the image resolution to the point where the number of self-similar pixels is minimized relative to the overall information content of the pixel grid. While there are algorithmic ways of accomplishing this minimization, for biological images I recommend adoption of an experimental approach to determining the spatial resolution and colour format required for each analysis on a case-by-case basis. This recommendation reflects my belief that the analyst (or taxonomic specialist), rather than an algorithm, is usually best placed to determine which morphological features need to be included in an image to ensure the down-sampled image set remains appropriate for the hypothesis test(s) under consideration.

Once the spatial and colour resolution necessary to represent the set of images have been established it is a simple matter to assemble the resultant pixel brightness or colour values into a mathematical description of the specimen. This is done by rearranging the matrix of pixel values into a single row of values in a standard data matrix. Each of the pixels, then, becomes a variable and the colour or greyscale data the values of those variables. Since each of the images processed in this manner is composed of the same number of variables (pixels), since each variable has a constant spatial relation to every other variable, and since the values each of these variables can adopt are of the same type and range of magnitudes, collectively these variables (pixels) form a mathematical space within which each of the specimens comprizing the sample can be located. Specimens whose form, colour, and blotch patterns are similar will lie close to each other in this space while those whose morphological attributes are distinct will lie at some remove from one another. This is precisely the same sort of inter-specimen representation we achieve in a linear distance, landmark, or semilandmark-based analysis. Nonetheless, by using pixel values as our variables rather than coordinate point locations we are preserving as much of the total morphological content of each image (and so each specimen) as possible as well as avoiding having to make any decisions about what aspect(s) of the morphology may, or may not, be important for resolving the problem at hand at the outset of an analysis.

This is quite a flexible approach to the analysis of biological specimen morphology as well as one that incorporates many of the data-analysis features we've been discussing in the context of different morphometric data types. The conceptual link to outline and 3D surface analysis seems straight forward. But even in the context of traditional forms of landmark analysis all we're really doing is changing our focus from the location of a small number of points (implicitly) embedded in a coordinate system to that of the total information content of the coordinate system itself. The price we're paying for this change of focus involves having to deal with a much larger number of variables than would be the case in traditional landmark and semilandmark analyses. But the benefit is that we are able to include much more morphological information that might be relevant to the questions we are asking than would be the case otherwise.

To illustrate this procedure let's take the set of ladybird beetle images in Figure 2 and ask whether this sample represents a continuously variable set of colour morphs (the null hypothesis) or whether we have two basic types of beetles here: orange beetles with black spots and black beetles with orange or red spots (the alternative hypothesis). Using a strictly Gestalt assessment of these morphologies I'll posit that the sample can be subdivided into orange and black colour morphs as listed in Table 1. Within this classification specimens 5 and 6 in row 2 of Figure 2 appear close to the intermediate condition between the two groups with the former being slightly more orange and the latter slightly more black. However, whether the boundary between these putative groups is gradational or disjunct within this sample I have no idea.

Table 1. Putative ladybird beetle colour morph assignments (by row and specimen no.).

Orange Morphs		Black Morphs	
1-1	2-3	1-5	2-8
1-2	2-4	1-6	3-4
1-3	2-5	1-7	3-5
1-4	3-1	1-8	3-6
2-1	3-2	2-6	3-7
2-2	3-3	2-7	3-8

Before reading any further you might like to consider Figure 2 yourself and come to your own preliminary conclusion as to whether the null or alternative hypothesis is more likely to be correct. It's not as easy at it you might think.

Figure 5. Resampled of 24 ladybird beetle (Family Coccinellidae) drawings (see Fig. 2) using a 32 x 32 colour (RGB) pixel grid. This resampling protocol reduced each specimen's pixel number by c. 80% will little loss in the spatial resolution of features necessary for characterizing patterns of morphological variation.

In preparation for our analysis we can reduce the spatial resolutions of the images that comprise Figure 2 from 150 x 150 pixels to 32 x 32 pixels (Fig. 5). From inspection of the plate of beetle images at this reduced level of resolution you can see that, despite the fact that we've decreased the total number of pixels representing each specimen by 95 percent, these lower resolution images preserve virtually all features of the specimens observable in the original image set. Thus, this operation has reduced the number of variables required to represent our specimens (= reduced their dimensionality) by preferentially eliminating redundant pixels (= reduced spatial autocorrelation) while suffering only a very minor loss of information content. So far, so good.

Since these are colour images, and since, in this case we do want to include an assessment of colour variation in our dataset, each specimen's image can be described completely by converting the pixel format to a 32 x 32 matrix of red, green, and blue (RGB) colour intensity values; an operation that results in the specification of 32 x 32 x 3, or 3,702 variables. This is a large and somewhat awkward dataset insofar as we have many more variables than specimens. However, it's a dataset we can work with and such skewed data matrices are by no means uncommon either in data analysis generally or morphometric data analysis in particular.

A covariance-based principal component analysis (PCA) of these images shows that 95 percent of the observed form, colour, and texture variation can be represented on 18 orthogonal components or axes of variation. This operation further reduces the effective dimensionality of our measurement system from 3,702 variables to 18 (a reduction of 99.5%). This operation also drops the number of variables down below the number of specimens in our dataset. The number of specimens in a morphometric dataset represents an important "curse of dimensionality" threshold (see Bellman 1957; MacLeod, 2007, Mitteröcker and Bookstein 2011) of which I'll have more to say below.

We can have a look at this principal component space to see what it tells us about the major features of variation within our sample (Fig. 6). The first two principal components (Fig. 6A) account for the largest single share of observed form, colour, and texture variation (37.44%). But variation in this sample is such that these axes portray a minor share of the sample's variation overall. The distribution of the images in this subspace suggests that there's more going on within this sample than just the simple black-orange distinction that seemed 'obvious' to me after a cursory inspection of Figure 2. Broad-scale clustering of the two putative colour morphs is evident in this plot, but the distinction between them involves both PC axes. Moreover, the character of the inter-morph boundary is not well defined. While the ordination of the orange morphs within this plane appears relatively gradational, the ordination of the black morphs appears highly structured.

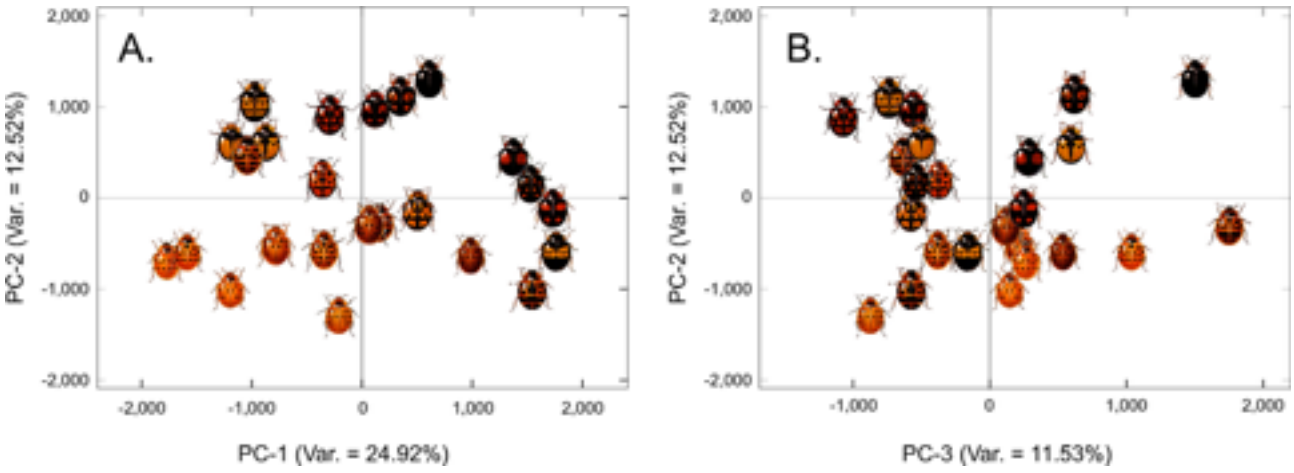

Figure 6. Distribution of ladybird beetle form and colour texture similarity and difference relations on the first three principal component axes of the image covariance matrix. Black icons = putative black beetle morphs. A. The subspace formed by PC-1 (*x*-axis) and PC-2 (*y*-axis). B. The subspace formed by PC-3 (*x*-axis) and PC-2 (*y*-axis). Note the comparative lack of specimens that project to positions along the principal component axes. Lack of of the empirical definition of axis form trends that might be afforded by such specimens makes it quite difficult to arrive at detailed geometric interpretations of the overall form space. See text for discussion.

Casual, qualitative inspection of this ordination pattern suggests that the black morph comes in two varieties: a black body with orange spots (the subgroup that plots toward the low end of PC-1) and a black body with red spots (the subgroup that plots toward the high end of PC-1). Thus, the main distinction within this sample appears to be that between orange beetles and red spotted beetles rather than orange and black beetles. The orange-black distinction that appeared so striking to me is reflected predominantly in the ordination of specimens along PC-2 along which the former exhibit low scores and the latter high scores. Also, along this axis the putative black morphs appear to exhibit a more-or-less unified distribution whereas the putative orange morphs comes in two varieties, a large group of bright orange and dusky orange morphs with much smaller black spots and a smaller group of morphs in which the black spots have coalesced to the extent that the total number of black and bright/dusky orange pixels is subequal (specimens 2-5 and 3-2). If we add PC3 into our assessment (Fig. 6B) this interpretation does not change greatly. Along the PC-3 axis there appears to a distinction between bright orange (low PC-3 scores) and dusky orange (high PC-3 scores) morphs, at least along the lower reaches of PC-2.

Given the complex character of image variation within this space detailed interpretations of these axes are surprisingly difficult to devise. We've seen this before with different types of data. But for these images the situation is decidedly more complex owing, no doubt, to the large number of original variables we're dealing with. Using qualitative 'eyeball' methods to interpret these axes the best we can do is compare and contrast images that lie at the extremes of variation along each axis and hope that these represent all, or at least the majority, of the contrasts controlling each specimen's placement. This is a particularly hazardous interpretive strategy to follow, especially when the distribution of forms in this space does not include many that lie close to the actual traces of the PC axes, as it is here for PC-1, PC-2 and PC-3. In such cases it's far better to calculate models of the image variation at specific positions along any axis you might want to interpret and base your interpretations on those models. Such models can be calculated in precisely the same manner as we have been calculating them for other data types (see MacLeod 2009, 2012, 2013) Along-axis image models for the first three principal components of the ladybird analysis are shown in Figure 7.

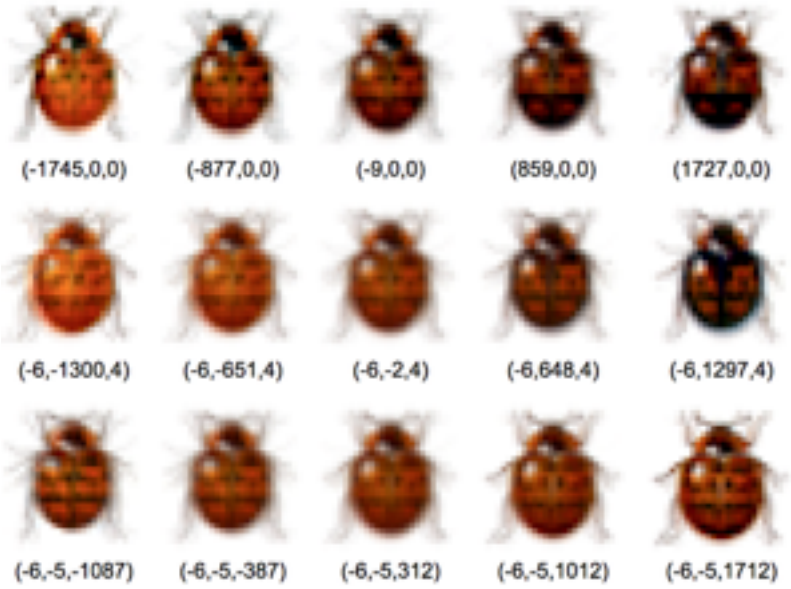

Figure 7. Image models for the ladybird beetle dataset calculated along at equally spaced coordinates along the first three principal component axes. See text for discussion. Numbers on parentheses refer to the coordinate positions at which each image model was calculated (see Fig. 6).

Inspection of these models clarify interpretation of the major modes of variation in our dataset immensely. Here we can see that PC-1 actually captures the distinction between orange morphs with numerous small black spots (extreme negative scores), through dusky orange morphs with larger black spots, to a median region in which forms composed of subequal black and dusky orange regions occur and on to morphs characterized by a predominantly black body with a single pair of large red spots on the elytra and margins of the thoracic shield (extreme positive scores). The second PC axis is subtly different in that it contrasts bright orange morphs with a smaller number of larger black spots (extreme negative scores), passes through a median region composed of dusky orange morphs with small black spots and on to forms characterized by black bodies with four symmetrically placed red spots on the elytra, the anterior of which have black spots in their centres, and elongate red spots at the margins of their thoracic shields (extreme positive scores). Finally, PC-3 captures the distinction between forms characterized by black bodies with large, elongate orange spots whose axes are at right angles to the beetles' anterio-posterior axis of symmetry (extreme negative scores), through a median region characterized by dusky orange morphs with black spots, and on to a region characterized by forms with black bodies and dusky orange stripes formed from the progressive amalgamation of orange spots (extreme positive scores). Close inspection of the PC-3 model set also reveals that, during the course of the transition from high negative to high positive scores a black band along the medial elytral margin changes colour and becomes orange. If we had not had access to these hypothetical image models to use as precisely located points of reference and comparison it is very doubtful that such a detailed interpretation of the major dimensions of form variation could be deduced from a simple inspection of the actual specimen ordinations alone.

Although calculation of along-axis shape models allows more detailed and useful interpretations of the PC ordination space to be made, the interpretations offered above remain only semi-quantitative insofar as comparisons between the models calculated along each axis were done by eye. An even more complete and nuanced picture of similarities of differences among these models can be assembled by calculating difference maps of comparisons between these models (Fig. 7).

Difference maps compare the (in this case) RGB values of corresponding pixel locations across two or more images and assign a colour to the mapped pixel whose hue is based on the amount of change recorded at the individual pixel locations. Typically these maps employ a temperature metaphor to express these results visually. Under this convention regions of the image characterized by little of no change are signified by blue (= cool) and regions characterized by high levels of change by (progressively) yellow, orange, and red (= hot). This type of representation has the advantage of being more objective, quantified, and detailed assessment of observed changes than the qualitative inspection of image differences. However, it should be remembered that there is no simple relation between the absolute amount of pixel value change at any particular location and the biological significance of that change. The strength of the difference map approach is that the maps draw our attention on particular aspects of form comparisons, or regions of form change, that might be overlooked as our eyes try to process and pick out consistent patterns or trends in the image model results. Difference maps can be used either in an exploratory sense (e.g., to identify regions of high or low that we might want to pay special attention to in subsequent analyses) or as a means of testing

specific hypotheses (e.g., if a generative hypothesis predicts that some sort of change will be localized in particular regions of the form), but should (almost) never be used on their own to interpret the biological significance of the form changes they record.

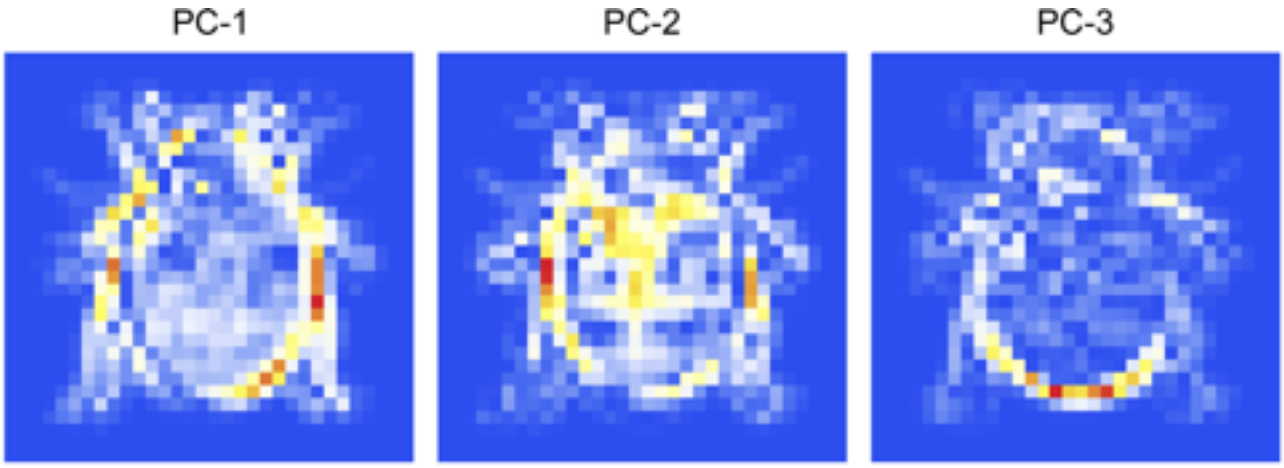

Figure 8. Along-axis image model difference maps for the first three sets of principal component axis models (see Fig. 7). This plots summarize regions of the image at which the different proportions of change in pixel colour values are taking place as one moves along each of the principal component axes. In these plots the individual pixels have been colour coded as follows: blue = no change, white = moderate change, yellow = moderately high change, orange = high change, red = very high change. Note how different PC axes summarize changes in different regions of the form.

When the difference map approach is used to compare differences between extreme low and high score ladybird beetle models in each of the along-axis sequences (Fig. 8) they do pick out unique aspects of variation that were not readily apparent in the qualitative inspection of the models shown in Figure 7. For example, there appears to be a general broadening of the elytra and thoracic shields along PC-1 with elytral colour pattern variation differentially occurring in the posterior portion of the form. Along PC-2 this broadening is confined to the middle part of the elytra with substantial colour pattern variation occurring in the anterior elytral region. along PC-3 moderate modes of colour pattern variation occur across the elytra accompanied by a slight concentration in the posterior portion of the thoracic shield. However, the predominant mode of form variation expressed along this axis is a slight lengthening of the body that appears to be on the order of 2-3 pixels in aspect.

And what of our original question regarding the distinctness of orange and black morphs? Figure 6C suggests (weakly) that the character of this transition might be gradational. However, just looking at a few planes through the original 18-dimensional PC space is insufficient for determining the actual degree of separation between these morphs. If a separation between groups can be located along any of these axes, either singly or when employed in combination, the groups are separable. Since we have neither the time, patience, nor software to visually inspect all possible geometries of points in an 18-dimensional space, some other procedure must be used to settle this question.

Here though we run into a bit of a minor, but current, controversy among morphometricians. The classic way to handle this problem would be to subject either the raw data or a reduced set of PC scores to a discriminant analysis procedure such a canonical variates analysis (CVA). As I've pointed out before (MacLeod 2009), CVA takes a multivariate dataset and performs a series of standardized axis rotations and transformations that result in the data for two or more groups being projected into a multivariate space in which each group centroid is maximally separated from every other group centroid relative to within-group dispersions. This method cannot usually be applied to raw or Procrustes-aligned morphometric data because of computational difficulties that arise when attempting to invert the within-groups covariance matrix. Nevertheless, these difficulties can be solved (usually) by taking the raw data through an initial PCA and discarding the component axes that represent a statistically or biologically insignificant proportion of the observed form variation.

This procedure has recently been criticised by Mitteröcker and Bookstein (2011) as unsuitable for morphometric data analyses for the following reasons.

1. Orientations of the CV axes are not orthogonal to each other in the space of the original variables used to calculate the CVA; as a result true Euclidean and/or Procrustes distances between specimens are not represented faithfully in the CVA ordination space.

2. If the number of variables greatly exceeds the number of specimens comprising a dataset CVA will always be able to find a linear combinations of variables that separates groups even in cases of artificially generated data drawn from the same multivariate normal distribution.[2]
3. The scores of objects projected into the CV space do not represent a linear transformation of the original variables, but a linear transformation of these variables after they have been transformed into a distorted space formed by an operation that renders the within-group covariance matrices spherical. Thus, modes of shape variation that optimally separate groups are difficult to interpret and difficult to model properly.
4. Canonical variate results are sensitive to the number and scaling of the variables used in their calculation.

As an alternative, Mitteröcker and Bookstein (2011) advocate use of a 'between groups PCA' in which the group means are used to determine the orientations of a set of eigenvector axes and the data comprising the sample projected into this group mean-determined PCA ordination space.

Certainly the between-groups PCA (BG-PCA) is a simple procedure that does have the advantage of preserving true Euclidean and Procrustes distances more faithfully — though not exactly — in the ordination space than a typical CVA would. On the one hand, for datasets in which the number of objects exceeds the number of variables, BG-PCA will, in most cases, produce a result that is suboptimal in terms of group centroid separation within its discriminant space relative to the discriminant space calculated as the result of a normal CVA. On the other, for datasets in which the number of variables greatly exceeds the number of objects a CVA may paint a misleading picture of the true degree of between-groups difference. Both approaches are compromised by the curse of dimensionality, but in different ways. Conducting a preliminary PCA in either case will reduce the influence of this curse and allow more of the original data collected to participate in the analysis. As for the ability of CVA results to be interpreted and modelled, MacLeod (2009) reviewed procedures for projecting CVA axes (or any group-separation trajectories specified in the CVA space) back into the space of the original variables and MacLeod (2011) presented procedures for modelling the results of this projection in the PCA space. These projection and modelling procedures are (slightly) more complex for a CVA than for a BG-PCA, but not seriously so.

From my personal point of view both procedures have their merits and their limitations. If the purpose is to achieve an optimal group assignment decision or to test hypotheses concerning a morphological distinctions referenced to an optimal between-groups discrimination space, and if the size of the sample is large relative to the number of variables being used to define the discriminant analysis, a CVA analysis of PCA scores should return an appropriately robust result. Alternatively, if the purpose is to achieve between-groups discrimination while preserving a high degree of correspondence in the ordination result to the fundamental Euclidean or Procrustes character of the data, and/or if the size of the sample is small relative to the number of variables being used to define the discriminant analysis, a BG-PCA should return an acceptable result. Since both methods are reasonably easy to calculate, prudence suggests that both approaches be utilized and their results compared for agreement.

As my original orange morph vs. black morph hypothesis involves only two groups a single discriminant axis will suffice for testing our hypothesis. Figure 9 presents results of both a BG-PCA and CVA in the form of frequency histograms of scores for the ladybird beetle image set projected onto their respective discriminant axes.

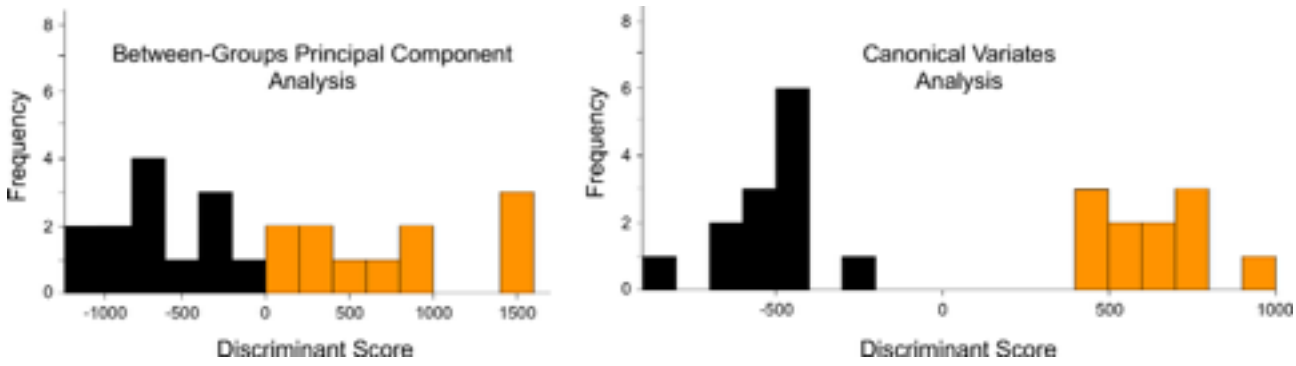

Figure 9. Discriminant analysis results for the distinction between orange and black ladybird beetle morphs. See text for discussion.

[2] This is an aspect of what I've referred to as the 'curse of dimensionality' and discussed in several columns in this series.

As expected, a standard CVA of the PCA scores produced the more definitive result with clear separation between these two putative beetle morphs. The BG-PCA result is consistent with that of the CVA result, but much less resolved. Indeed, if the BG-PCA result had been the only test used a distinction between orange and black morphs would be confirmed, but the character of the transition between them would have remained ambiguous. Perhaps the best way of conceptualizing the relation between these two results is that the BG-PCA result is what we would see if we could view the 18-dimensional PCA space in an orientation that best separated these two putative morphs whereas the CVA result (in this case) allows us to perform a similar operation and, at the same time, hold a mathematical magnifying glass up to the critical transition interval. A parametric Monte Carlo simulation test using hypothetical random datasets of the same size but identical means indicates that the curse of dimensionality has not played a significant role in the determination of either of the ordination results shown in Figure 9 (see MacLeod in press).

Although these results do not 'prove' that ladybird beetles came in two varieties, they do show that, for the dataset illustrated in Figure 2, we can use the data encoded in digital images to test specific hypotheses regarding the character of variation within a morphologically complex sample of organisms even if we cannot reasonably specify the positions of relocatable landmarks or boundary outlines. This 'image analysis' technique is probably best used in an exploratory context, as a way of generating specific morphological hypotheses that can be tested using more normal landmark, outline, or landmark + outline approaches. Nevertheless, it represents a conceptual link between morphometrics and approaches to computer vision that may have widespread application across the biological sciences, and perhaps even further (see MacLeod et al. 2013).

To give credit where credit is due I must admit to not being the first to come up with the idea of using digital images directly as input to a morphometric analysis. In researching this article I ran across a similar application, called 'eigenfaces', which is currently being used in biometric face recognition and cognitive neuroscience studies (see Sirovich and Kirby 1987, Turk and Pentland 1991a,b). The approach I outline above also shares certain historical similarities with the Digital Image Analysis System (DAISY) design for generalized semi-automated specimen identification systems (ONeill 2007). Owing to the obvious implications such technology has for the security and surveillance business sectors much primary research is being done on the general problem of automated object identification at the moment. While this technology has yet to make a substantial impact in the mainstream taxonomic and biological sciences, tantalizing glimpses of what these approaches might be capable of in the near future are available (see MacLeod 2007). In addition, the need for the introduction of such systems into research programmes that rely on rapid, accurate, and consistent taxonomic identifications is becoming more widely recognised with each passing year (e.g., MacLeod et al. 2010, Culverhouse et al. 2013).

How can you experiment with this more direct form of image-based morphological analysis? It's easier than you might think. Most of us are already used to working with digital images these days. Many commercial and public-domain image-processing software packages (e.g., *Photoshop*, *Graphic Converter*, *Gimp*) have the ability to change the spatial resolutions of digital images such that you can control the number rows and columns in the pixel grid. These same packages will also allow you to convert a colour image into its greyscale or binary equivalent and give you control over image exposure settings. Once your image set is in the correct format you'll need a way of converting your images into ASCII datafiles. Older versions of Photoshop will do this as will *Graphic Converter* and *Mathematica*. Once you've converted your image files into greyscale or colour (RGB) brightness values all you need to do is reformat your image data matrices into a single row of values and assemble these into a standard data matrix in which the rows represent the specimens and the columns the pixel variables. This will typically be a large data matrix with many more columns than rows. Once your data are in this format any reputable PCA programme should be able to carry out the preliminary PCA analysis, though run times for very large datasets may be long. Secondary CVA or BG-PCA analysis can then be carried out on the preliminary PCA results as outlined above. All datasets and procedures used in this essay are available as *Mathematica* notebooks from myself. But all steps in the analysis can be performed by software available to anyone either free of charge of for a small licensing fee (depending on the type of computer you have access to).

As this will be the last essay in the *Palaeo-Math 101* series all that's left to do is to thank the past three editors of the Palaeontological Association Newsletter — Phil Donoghue, Richard Twitchett, Al McGowran for their indulgence over the years of this column's publication, to send a special thanks to Nick Stroud (the unseen force behind the newsletter) for his kind attention in laying out the column articles sympathetically and instituting the numerous last-minute edits to which I am prone, to acknowledge Mark Sutton's support with regard to the pdf and web site versions of the articles, to thank the Palaeontological Association generally for creating a forum where a series on such an unusual subject might exist all, to thank all the people who have contacted me over the years with questions, corrections, extensions, queries, requests for advice, etc. about quantitative data analysis in the palaeontological and biological sciences as a result of these essays. Last but not least, I must thank you dear reader for taking the time to consider how quantitative forms of data analysis might be useful in your data-to-day research. Although this series has only

covered two aspects of palaeontological data analysis in any depth, I hope it's demonstrated the power and utility of such approaches generally.

When I started this series in 2004 my goal was to write a small series of essays that would provide students post-docs, and young researchers with sort of practical, easy-to-follow discussions of the the ins and outs of palaeontological data analysis that I had often wished I had when I was just starting out in this field. I've enjoyed writing each of these columns immensely and, based on the many appreciative comments I've received over the years from readers like you, am content that I've reached that goal. It's been time well spent.

Norman MacLeod
The Natural History Museum
N.MacLeod@nhm.ac.uk

REFERENCES

BELLMAN, R. E. 1957. *Dynamic programming*. Princeton University Press, Princeton 340 pp.

CULVERHOUSE, P. F., MacLEOD, N., WILLIAMS, R., BENFIELD, M. C., LOPES, R. M. and PICHERAL, M. 2013. An empirical assessment of the consistency of taxonomic identifications. *Marine Biology Research*, **10**, 73-84.

MacLEOD, N. 2004. Prospectus & Regressions 1. *Palaeontological Association Newsletter*, **55**, 28–36.

MacLEOD, N. 2007a. Groups II. *Palaeontological Association Newsletter*, **65**, 36–49.

MacLEOD, N. 2007b. *Automated taxon identification in systematics: theory, approaches, and applications*. CRC Press, Taylor & Francis Group, London 339 pp.

MacLEOD, N. 2009. Form & shape models. *Palaeontological Association Newsletter*, **72**, 14–27.

MacLEOD, N. 2012. Going round the bend: eigenshape analysis I. *Palaeontological Association Newsletter*, **80**, 32–48.

MacLEOD, N. 2013. Semilandmarks and surfaces. *Palaeontological Association Newsletter*, **83**, 37–51.

MacLEOD, N. in press. The direct analysis of digital images (eigenimage) with a comment on the use of discriminant analysis in morphometrics. *In* LESTREL, P. E. (ed.) *Proceedings of the Third International Symposium on Biological Shape Analysis*. World Scientific, Singapore, pp. Custom 7.

MacLEOD, N., BENFIELD, M. and CULVERHOUSE, P. F. 2010. Time to automate identification. *Nature*, **467**, 154–155.

MITTERÖCKER, P. and BOOKSTEIN, F. L. 2011. Linear discrimination, ordination, and the visualization of selection gradients in modern morphometrics. *Evolutionary Biology*, **38**, 100-114.

ONEILL, M. 2007. *DAISY*: a practical tool for semi-automated species identification. *In* N. MacLeod (ed). *Automated taxon recognition in systematics: theory, approaches and applications*. Taylor & Francis, London, 101–114 pp.

SIROVICH, L. and KIRBY, M. 1987. Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America*, **A4**, 519–524.

TURK, M. and PENTLAND, A. 1991a. Face recognition using eigenfaces. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, **June 1991**, 586–591.

TURK, M. and PENTLAND, A. 1991b. Eigenfaces for recognition. *Journal of Cognitive Neurosciences*, **3**, 71-86.