Analysis of modularity in configurations of landmarks

Christian Peter Klingenberg

Faculty of Life Sciences, The University of Manchester

1 Introduction

Most living systems are organized in a modular manner: there are groups of parts, called modules, which are tightly integrated by many or strong interactions among the parts, and which are relatively independent from other modules because there are only few or weak interactions between parts of different modules. This type of organization applies to networks of gene regulation, metabolic networks, neural circuitry as well as to morphological variation. Modularity is therefore relevant for the analysis of biological shapes. This paper outlines a methodology for assessing a-priori hypotheses of modularity in configurations of landmarks.

In the context of shape analysis, a definition of modularity is required that is applicable to configurations of landmarks. Interactions of biological processes are not directly observable in morphometric data, but the covariation of the relative positions of landmarks can be interpreted as reflecting interactions of the developmental processes that produce the biological structure under study. A module can be defined as a subset of landmarks that covary strongly with each other, but are relatively independent of the landmarks in other subsets of the overall configuration. It is important to notice that this is a question of relative independence, because some degree of overall integration throughout organs and even the whole organism can usually be expected.

A hypothesis of what may be the modules in the configuration corresponds to a partition of its landmarks into two or more subsets. If the subdivision into subsets corresponds to the true boundaries between modules, covariation between sets will be due to the relatively weak interactions among modules, and is therefore expected to be weak. In contrast, subdivisions that cut through the true modules will show covariation that reflects the strong interactions within modules, and is therefore expected to be stronger. A measure of covariation for the true partition into modules should therefore yield a measure of covariation that is lower than that for most or all alternative partitions.

Accordingly, the analysis of modularity in a configuration can be viewed as a combinatorial problem of enumerating and comparing the possible partitions of the landmarks into subsets. This approach requires a number of considerations, which will be discussed in this paper. The first is the choice of a scalar measure of association between two or more sets of variables, which is needed for the comparison of alternative partitions. The second is the potentially large computational effort required for enumerating and comparing alternative partitions; introducing suitable constraints is therefore desirable. I discuss one constraint that is based on the fact that biological interactions are mediated by the tissue between landmarks, and that modules therefore should be spatially contiguous. Finally, I demonstrate a practical application of the method and give an outlook on further developments.
2 A measure of association

The measure of association between subsets of landmarks is the RV coefficient (Escoufier, 1973), which was proposed as an analogue to bivariate correlation for quantifying the association between two sets of variables $X_1$ and $X_2$. The joint covariance matrix of both sets, $S$, can be written as

$$S = \left( \begin{array}{cc} S_1 & S_{12} \\ S_{21} & S_2 \end{array} \right).$$

In this matrix, the blocks $S_1$ and $S_2$ are the covariance matrices of the two sets of variables and the blocks $S_{12}$ and $S_{21}$ are matrices of covariances between the two sets ($S_{12} = S_{21}^T$). The RV coefficient can then be computed as

$$RV = \frac{\text{trace}(S_{12}S_{21})}{\sqrt{\text{trace}(S_1S_1) \text{trace}(S_2S_2)}}.$$

The numerator can be interpreted as a measure of the magnitude of covariation between the two sets, and the division provides a scaling by the amount of variation within blocks. If both blocks of variables contain only one variable each, this coefficient reduces to the bivariate $R^2$, the squared correlation coefficient. The RV coefficient can take values from 0 (if all covariances in $S_{12}$ are 0) to 1 (if $X_2 = BX_1 + C$, where $B$ is a $q \times p$ matrix with $q \leq p$ so that $B^TB = lI_q$ for some $l \neq 0$ and some constant vector $C$). The RV coefficient is invariant under orthogonal transformations and scaling of $X_1$ and $X_2$.

As a measure of association among $k$ sets of variables ($k \geq 2$), I use the average of Escoufier’s RV coefficient over all pairs of sets. This definition of a multi-set measure of association treats all subsets equally, regardless of the number of variables included or the amount of variance or covariance for which the subsets account. Because the computation of the RV coefficient does not involve operations such as matrix inverse, it is computationally fast, which is an advantage particularly in the context of the combinatorial analysis.

3 Spatial contiguity of modules

Biological interactions between precursors of parts are mediated by the intervening tissues, because any communication between parts proceeds by diffusion of substances or similar processes, but not ‘at a distance’. Therefore, only subsets of landmarks that are spatially contiguous are biologically reasonable candidates as modules. In addition, a requirement of spatial contiguity is useful as a constraint to limit the number of alternative partitions to be compared.

A subset of landmarks is said to be spatially contiguous if they are connected by the edges of a adjacency graph, and a partition of the configuration is said to be contiguous if each of the resulting subsets is itself contiguous. The adjacency graph can be obtained from a Delaunay triangulation of the mean landmark configuration (Procrustes consensus), but usually some editing is required to ensure that the adjacency relationships between landmarks make biological sense (e.g., eliminating edges linking opposite ends of a concave portion of the configuration).

Imposing the constraint of spatial contiguity is effective at reducing the number of configurations to be considered. For the configuration of landmarks in the fly wing shown in Fig. 1, there are 6,435 partitions of the landmarks into two subsets of 7 and 8 landmarks, but only 655 of these are contiguous. This difference is even more pronounced if there are more subsets: for the subdivision into three subsets of 6, 5 and 4 landmarks, there are 630,630 partitions, of which 2,946 are contiguous.
4 Application

The method is applied to a data set of fly wings (109 female \textit{Drosophila melanogaster}). The a-priori hypothesis is the division of the wing into anterior and posterior compartments (dashed line in the inset of Fig. 2). The observed RV coefficient for this partition is 0.406, which is slightly above the median for the distribution of RV coefficients across the 655 contiguous partitions. This value is not at or near the lower extreme, as it would be expected if the anterior and posterior compartments were modules.

![Figure 2: Analysis of modularity in the Drosophila wing.](image)

The analysis with all 6,435 possible partitions of the 15 landmarks into subsets of 7 and 8 landmarks, including those that are not contiguous, provided a similar result.

5 Outlook

All the discussion in this paper has considered the situation where an a-priori hypothesis of modularity is available. This is the case in many biological applications, but an extension of the approach for the exploratory search for modules would be highly desirable. Because the number of modules and the number of landmarks per module needs to be determined by the analysis in this case, the number of possible partitions is substantially greater and the use of suitable heuristics needs to be considered.

For symmetric configurations, both in analyses with an a-priori hypothesis and in exploratory searches, the size of the problem can be reduced by taking into account the symmetry (including only the median landmarks and paired landmarks from one side only).
Phenomena such as allometry, the dependence of shape on size, provide overall integration of the entire configuration and can possibly interfere with the analysis of modularity. Allometry can be taken into account by performing the analysis on the covariance matrix of the residuals from a regression of shape on size.

The RV coefficient only takes into account the linear associations between subsets of landmarks. This is sufficient in most biological data sets (particularly if studies are limited to adults or a single growth stage). While measures of association suitable for nonlinear relationships exist (e.g. R Iaci et al., unpublished MS), they are computationally demanding and thus will not be suitable in the combinatorial context discussed here.

References