PalaeoMath 101

Groups II

Last time out we began to confront the problems presented by datasets that include group-level structure. We also developed some statistical tools we could use to determine whether that structure was reflected in statistically significant differences in group means and to assign unknown specimens to the closest group mean. So far, so good. But what we really want is some way of defining a space—like a PCA space or a PCoord space—in which the groups are maximally separated.

You'll recall this plot of the *Iris* data from the previous essay (Fig. 1, see Palaeontological Association Newsletter, 64:35–45, also see that essay, or the *PalaeoMath101* Excel spreadsheet at http://www.palass.org/modules.php?name=palaeo_math&page=1 for a listing of these data). This captures the problem nicely. Given just four variables there are effectively six different ways of looking at the problem if we ignore plots of the four variables against themselves and the plots in which the same variables are plotted on different axes. Each plot yields some information about both within-group variation and between-group separation. Some plots seem more informative than others. But no single plot tells the whole story.

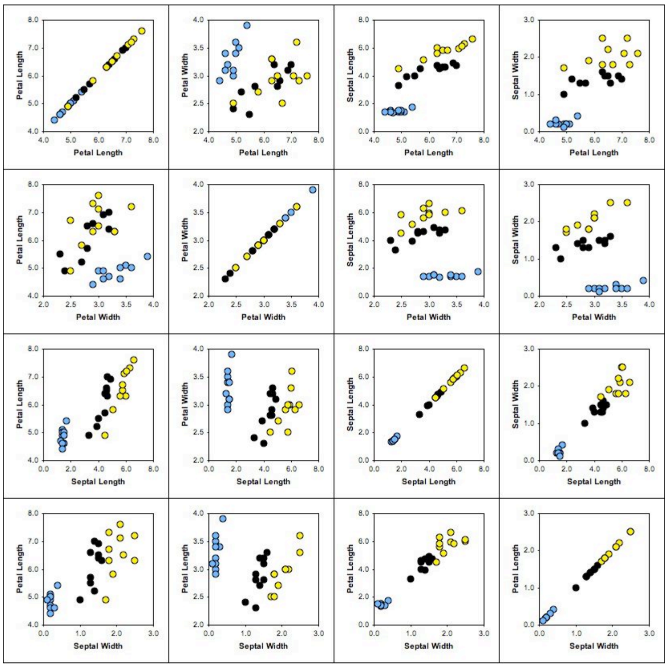


Figure 1. Crosstabulation diagram for Fisher *Iris* data. *I. setosa* (cyan), *I. versicolor* (black), *I. virginica* (yellow).

Ideally we'd like to see our data transformed into a low-dimensional space, such that the majority of the between-group separation is summarized in just a few axes. Also, if the equations of the axes could give us some indication of which single or combination of the original variables was most important for achieving group discrimination (which is another way of saying 'most important for group characterization'), that would be nice too. It's another tall order, but our colleagues over in the maths department have some ideas along these lines. Let's have a look.

Before we can begin our discussion make sure we have some basic concepts straight, specifically the difference between discrimination and classification. Both involve groups but there is a world of difference between them—mathematically speaking—that we need to understand before the mathematical operations will make much sense. Understanding these concepts will also help us understand the difference between the material I presented in the Groups I essay and what I'll be presenting below.

Discrimination is the act of determining a mathematical expression that distinguishes between groups of measurements or observations. In order to perform a discrimination or 'discriminant' analysis the groups need to be specified at the outset of an investigation. *Classification* is the act of determining how many groups are present in a collection of measurements or observations. This procedure does not require knowledge of the number of groups beforehand. Rather, that information is the purpose or result of the classification analysis. One group of techniques tells you how best to separate groups (discriminant analysis) the other tells you how many groups a sample contains (classification analysis). Of course, in the real world palaeontologists want to know both. The problem is you can't get at both questions in any single analysis. The mathematics that optimize the discrimination of groups of data require specification of the number of groups to be discriminated and the mathematics of classification analysis require that the characteristic differences between groups be known. What to do?

Inevitably, we fall back on using combinations of analyses. Principal components analysis, factor analysis, principal coordinates analysis, correspondence analysis and the rest of the 'single group' methods can be used to obtain a sense of how many groups there might be in a dataset. They can do other things too, but practically speaking this is one of their primary uses. Once some hypotheses about possible classification schemes have been developed based on results of a single-group analysis, those can be checked for statistical significance using the methods of mean-difference analysis (e.g., likelihood-ratio test, Hotelling's T^2-test). These results will allow decisions to be made regarding a viable classification scheme, after which consideration of the discriminant problem can begin. Mahalanobis distances can be used to affect identification by assigning individuals to groups based on their proximity to the group centroid (after scaling the variables by the inverse of the pooled covariance matrix). However, the space in which the Mahalanobis distance operates has not been optimized for maximal group separation. Nonetheless, it is possible to create a space that optimizes the difference between classification groups—at least the distances between their centroids. It is to this missing piece of the puzzle we now turn.

Most discussions of discriminant analysis begin with a discussion of the two groups case—where the point is to find a linear discriminant function that separates two groups. This is obviously the simplest case of discrimination and, because of this the mathematics involved can be simplified. Nevertheless, the two-sample case hardly ever comes up in real situations. For the most part we need to distinguish between three or more groups and so need an approach to determining discriminant functions that is powerful enough to handle any number of groups. Since the simplified mathematics of two-group linear discriminant analysis cannot be extended to the multiple-groups case, we'll proceed directly to the multiple-groups problem, the most popular solution to which is called canonical variates analysis (CVA). Should we ever need to discriminate between just two groups, CVA works fine for those data too.

In our example analysis we'll stick with the Fisher *Iris* data from the previous essay, but bump up the number of individuals in each group in order to get a better estimate of group variation and to illustrate some features of the technique. The following table lists these example data.

Table 1. First twenty-five specimens from each species included in Fisher (1936) *Iris* data.

	Iris setosa				*Iris versicolor*			
	Petal		Sepal		Petal		Sepal	
	Length	Width	Length	Width	Length	Width	Length	Width
1	5.1	3.5	1.4	0.2	7.0	3.2	4.7	1.4
2	4.9	3.0	1.4	0.2	6.4	3.2	4.5	1.5
3	4.7	3.2	1.3	0.2	6.9	3.1	4.9	1.5
4	4.6	3.1	1.5	0.2	5.5	2.3	4.0	1.3
5	5.0	3.6	1.4	0.2	6.5	2.8	4.6	1.5
6	5.4	3.9	1.7	0.4	5.7	2.8	4.5	1.3
7	4.6	3.4	1.4	0.3	6.3	3.3	4.7	1.6
8	5.0	3.4	1.5	0.2	4.9	2.4	3.3	1.0
9	4.4	2.9	1.4	0.2	6.6	2.9	4.6	1.3
10	4.9	3.1	1.5	0.1	5.2	2.7	3.9	1.4
11	5.4	3.7	1.5	0.2	5.0	2.0	3.5	1.0
12	4.8	3.4	1.6	0.2	5.9	3.0	4.2	1.5
13	4.8	3.0	1.4	0.1	6.0	2.2	4.0	1.0
14	4.3	3.0	1.1	0.1	6.1	2.9	4.7	1.4
15	5.8	4.0	1.2	0.2	5.6	2.9	3.6	1.3
16	5.7	4.4	1.5	0.4	6.7	3.1	4.4	1.4
17	5.4	3.9	1.3	0.4	5.6	3.0	4.5	1.5
18	5.1	3.5	1.4	0.3	5.8	2.7	4.1	1.0

19	5.7	3.8	1.7	0.3	6.2	2.2	4.5	1.5
20	5.1	3.8	1.5	0.3	5.6	2.5	3.9	1.1
21	5.4	3.4	1.7	0.2	5.9	3.2	4.8	1.8
22	5.1	3.7	1.5	0.4	6.1	2.8	4.0	1.3
23	4.6	3.6	1.0	0.2	6.3	2.5	4.9	1.5
24	5.1	3.3	1.7	0.5	6.1	2.8	4.7	1.2
25	4.8	3.4	1.9	0.2	6.4	2.9	4.3	1.3
Σ	125.7	87.0	36.5	6.2	150.3	69.4	107.8	33.6
Min.	4.3	2.9	1.0	0.1	4.9	2.0	3.3	1.0
Max.	5.8	4.4	1.9	0.5	7.0	3.3	4.9	1.8
Mean	5.0	3.5	1.5	0.2	6.0	2.8	4.3	1.3
Median	5.0	3.4	1.5	0.2	6.1	2.8	4.5	1.4
Variance	0.2	0.1	0.0	0.0	0.3	0.1	0.2	0.0
S. Dev.	0.4	0.4	0.2	0.1	0.5	0.4	0.4	0.2

	Iris virginica			
	Petal		Sepal	
	Length	Width	Length	Width
1	6.3	3.3	6.0	2.5
2	5.8	2.7	5.1	1.9
3	7.1	3.0	5.9	2.1
4	6.3	2.9	5.6	1.8
5	6.5	3.0	5.8	2.2
6	7.6	3.0	6.6	2.1
7	4.9	2.5	4.5	1.7
8	7.3	2.9	6.3	1.8
9	6.7	2.5	5.8	1.8
10	7.2	3.6	6.1	2.5
11	6.5	3.2	5.1	2.0
12	6.4	2.7	5.3	1.9
13	6.8	3.0	5.5	2.1
14	5.7	2.5	5.0	2.0
15	5.8	2.8	5.1	2.4
16	6.4	3.2	5.3	2.3
17	6.5	3.0	5.5	1.8
18	7.7	3.8	6.7	2.2
19	7.7	2.6	6.9	2.3
20	6.0	2.2	5.0	1.5
21	6.9	3.2	5.7	2.3
22	5.6	2.8	4.9	2.0
23	7.7	2.8	6.7	2.0
24	6.3	2.7	4.9	1.8
25	6.7	3.3	5.7	2.1
Σ	164.4	73,2	141.0	51.1
Min.	4.9	2.2	4.5	1.5
Max.	7.7	3.6	6.9	2.5
Mean	6.6	2.9	5.6	2.0
Median	6.5	2.9	5.6	2.0
Variance	0.5	0.1	0.4	0.1
S. Dev.	0.7	0.4	0.6	0.3

Canonical variates analysis was invented by R. A. Fisher (1936) with important contributions by Bartlett (1951, regarding how to calculate the inverse of a matrix), Mahalanobis (1936, regarding use of Mahalanobis distances in discriminant analysis), and Rao (1952, in synthesizing Fisher's and Mahalanobis' concepts into the modern procedure). The basic idea behind modern approaches to CVA is reasonably simple. It is in essence a two-stage rotation of a data matrix that has been subdivided into groups, hence the name *canonical* variates.

Campbell and Atchley (1981) provide an excellent discussion of the geometric transformations implicit in CVA. Their presentation has served as a model for the geometric explanation presented below. In the actual algorithm (which we'll discuss after the geometric presentation) several of these steps are performed simultaneously. Most textbook descriptions of CVA only focus on presenting a recipe of equations and plots such that comparatively few practitioners gain much understanding of the geometry inherent in the methods. In my presentation we'll review of few basic equations (which readers of this column have seen before) and then let the pictures do most of the talking.

First, recall that in our previous discussion of the likelihood-ratio test we developed the idea that total similarity relations (*T*) within grouped data matrices could be subdivided into 'within-groups' (*W*) and between-groups (*B*) partitions.

$$T = B + W \tag{11.1}$$

There are different ways to operationalize this concept, but in the case of CVA the *T* matrix usually represents the total sums of squares and cross products (SSQCP) for all variables and has the following form.

$$t_{r,c} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{i,r,j} - \bar{x}_r)(x_{i,c,j} - \bar{x}_c) \tag{11.2}$$

In this expression *r* and *c* refer to the rows and columns of the *T* matrix (any cell of which is occupied by a value *t*) with $\bar{x}_r$ and $\bar{x}_c$ representing the grand means for the entire, combined dataset. The grand mean is the centre of the pooled sample of all measurements. Matrix *T* then summarizes the dispersion of the total dataset about this group-independent, fixed reference.

The *W* matrix represents the within-groups SSQCP matrix and has a corresponding form.

$$w_{r,c} = \sum_{j=1}^{k} \sum_{i=1}^{n_j} (x_{i,r,j} - \bar{x}_{jr})(x_{i,c,j} - \bar{x}_{jc}) \tag{11.3}$$

Once again, *r* and *c* refer to the rows and columns of the *W* matrix (any cell of which is occupied by a value, *w*). Now the variables $\bar{x}_{jr}$ and $\bar{x}_{jc}$ refer to the analogous group-specific means. Here, the group mean is the centre of the cloud of points representing each group in Figure 1. Matrix *W*, then, summarizes the dispersion of each dataset relative to its own group-specific reference.

Once *T* and *W* have been found the most intuitively way of determining the *B* matrix is to simply subtract each element of the *W* matrix from the corresponding element of the *T* matrix (*B* = *T* - *W*). Conceptually though, the between-groups matrix summarizes the dispersion of the group means from the grand mean.[1]

[1] Confusingly (in my view) a number of programmes currently available for implementing CVA operate on matrices that violate the basic *T* = *W* + *B* relation. In such cases the authors of those algorithms are usually trying to take account of differences between the number of specimens representing each group. Unfortunately, they rarely specify exactly how their programmes undertake this correction, often resulting in slight non-comparabilities between the results reported by different programmes.

In our geometric example analysis we'll reduce the Table 1 data to just two variables: petal width and petal length.[2] Canonical variates analysis begins (conceptually) with a standard PCA analysis of the within-groups dispersion matrix (Fig. 2).

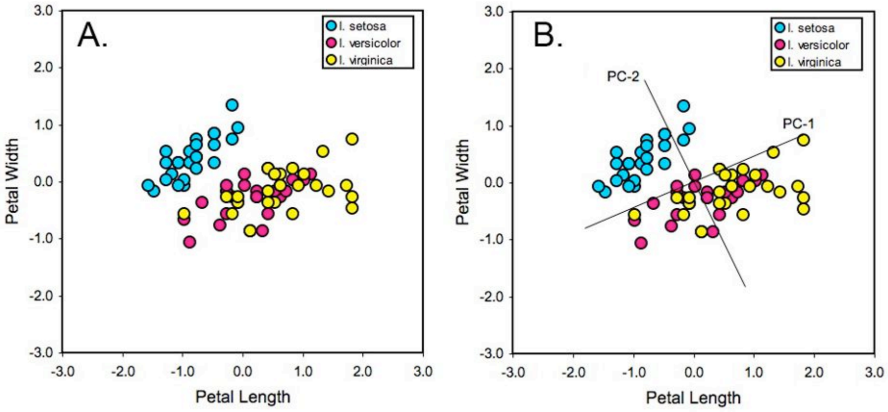


Figure 2. Stage 1 CVA implicit rotation. A. Scatterplot of first two *Iris* variables for example dataset. B. Orientation of the two pooled-sample principal components of the within-groups SSQCP matrix (*W*).

The purpose of this step is to re-describe the dispersion of the entire dataset in terms of a set of uncorrelated variables. Although the *W* matrix calculates dispersion from the group means, this operation involves a rigid rotation of the data about the grand mean. In order to facilitate plotting it is often convenient to mean-centre the entire dataset about the grand mean prior to analysis, in which case the grand mean will be the origin of the data's coordinate system. This convention has been followed in Figure 2 and throughout all subsequent analyses.

Next, CVA performs a somewhat counter-intuitive scaling operation. As you can see from Figure 2B, the scatters of the original groups are elongated with much more variance along PC-1 than PC-2. This reflects the greater variation of the petal length relative to petal width data, which in turn reflects the fact that *Iris* petals are much longer than they are wide. In order to achieve maximum separation between the group centroids the principal components are scaled by the square root of the associated eigenvalue. This operation involves multiplying each individual's PC score by the reciprocal of that square root. The result of this intermediate scaling operation is illustrated in Figure 3.

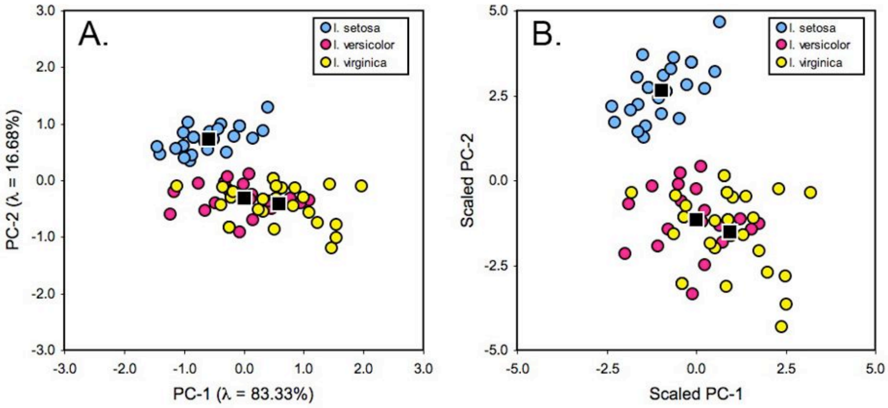


Figure 3. Intermediate scaling operation of a CVA. A. Scatterplot of *Iris* PC scores for the Stage 1 rotation (see Fig. 2). B. Result of scaling the two within-groups principal components by the square roots of their associated eigenvalues. Note difference in separation of the group centroids (black squares) after scaling.

[2] A listing of all calculations is provided in the *PalaeoMath101* Excel spreadsheet at http://www.palass.org/modules.php?name=palaeo_math&page=1.

The effect of this scaling is subtle, but important. Note how the range of variation for each group has been adjusted so that it is subequal along both axes. This is a form of data standardization. The scaling operation forces each eigenvector (= principal component) to have the same length. Thus, the data have been relatively stretched along PC-2 (the shorter eigenvector) and compressed along PC-1 (the longer eigenvector). This transforms the formerly elongate distributions of the group-based point clouds into forms that are more spherical. Note also how this operation has greatly increased the separation of group centroids or means from one another, especially in terms of the separation of *I. setosa* from the other two species. That looks like a big advantage in terms of accomplishing discrimination, which is what CVA is all about. But the significance of this operation is actually both more and less profound than it might appear at first.

What we're doing by scaling the PC space in this way is reminding ourselves of what we mean by 'distance' in a multivariate space. As we discussed last time, correlations between variables matter when it comes to assessing the separation between any two points in a space defined by multiple variables. We apply a similar scaling operation to the Mahalanobis distance calculation specifically to correct for distortions caused by inter-variable correlations. The scaling operation we've just performed in the intermediate stage of our CVA analysis distorts the PC space such that the geometric reality of the distribution of points in that space matches our 'common sense' notion of distance (recall we performed the original PC rotation on *W*, not *T*). This scaling operation shows us that the notion of distance between points in the standard PC multivariate space can be just as distorted as it is in ordinary scatterplots. By using the eigenvalues to scale the eigenvectors we can construct a 'true' picture of the separations between points in this group-defined space, one that conforms to the world of spatial relations in which we live. Thus, our three *Iris* species are actually more distinct than figures 2A, 2B, or 3A would have us believe. That's the profound bit. The trivial bit is that all this complexity is taken into consideration by the Mahalanobis distance. Thus we've had a way of taking the distortions inherent in the spaces represented by 2A, 2B, and 3A into account all along.

The second and final stage of a CVA focuses on the group centroids. While the first rotation summarized within-groups dispersion patterns, a second rotation is required to summarize between-groups dispersion patterns. This is accomplished by conducting a second PCA, this time using only data from the positions of the group means in the orthogonal *and* variance standardized—or orthonormal—space (Fig. 3B). Figure 4 illustrates the result of this operation.

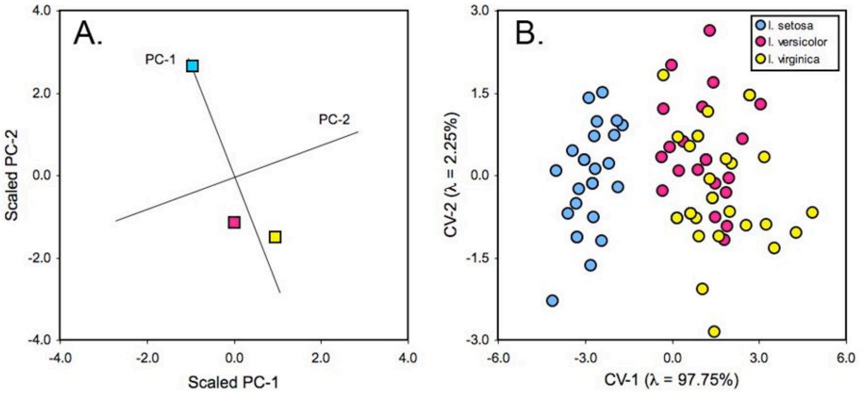


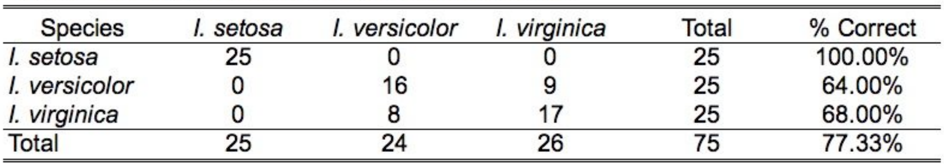
Figure 4. Stage 2 CVA implicit rotation. A. *Iris* group centroids plotted in the within-groups orthogonal-orthonormal space (see Fig. 3B) with between groups PC (= CVA) axes. B. Reduced *Iris* dataset plotted in the space defined by the CVA axes.

The scatterplot shown in Figure 4B is typically presented as the CVA ordination. Generally speaking there are one fewer CVA axes with positive between-groups eigenvalues than the number of groups present in the analysis. Once these results have been obtained most routines will also report statistical tests for group distinctiveness (e.g., Hotelling's T^2) and a

Mahalanobis distance-based cross-tabulation analysis of the data used to define the CVA space. The former are used to confirm group distinctiveness (see previous column for examples and details of these calculations) while the purpose of the latter is to determine the degree to which these particular CVA results can provide a reliable basis for achieving discrimination between the groups.

Note these are very different questions. It is quite possible for group means to be distinct relative to their within-groups dispersion yet contain so much overlap between their respective point clouds that effective discrimination is more-or-less impossible. Results of this cross-tabulation analysis are usually presented in the form of a ‘confusion matrix’ that summarizes the extent to which specimens assigned *a priori* to a given group are placed in the appropriate group by a Mahalanobis distance analysis (see previous column for details of this calculation). The confusion matrix for the two-variable *Iris* dataset is provided below.

Table 2. Confusion matrix for the reduced *Iris* dataset.



Species	*I. setosa*	*I. versicolor*	*I. virginica*	Total	% Correct
I. setosa	25	0	0	25	100.00%
I. versicolor	0	16	9	25	64.00%
I. virginica	0	8	17	25	68.00%
Total	25	24	26	75	77.33%

As can be seen from both this matrix and Figure 4B, *I. setosa* is perfectly discriminated from *I. versicolor* and *I. virginica* by the first CVA axis. However, approximately one-third of the specimens assigned to the latter two species are mis-assigned to these other groups. Is this a good result? The answer depends on the question you’re asking along with your ability to collect other information and/or access additional specimens of each group. If it is of the utmost importance to identify all specimens perfectly using only these variables, the fact that this analysis produced something like 35 percent incorrect identifications for two of the three groups *for the sample used to define the discriminant space* is a matter of concern. Still, for many applications—including most replication-based studies of systematic identifications—a consistent identification accuracy of 65 percent is competitive with most human experts (see MacLeod 1998; Culverhouse in press). Of course, this question is moot for the *Iris* dataset as we have ready access to measurements from additional specimens (which would improve our estimates of *W* and *B*) and additional variables (see below).

There is one additional issue we need to discuss before we leave this simple example. As with all the single-group data analysis methods we’ve discussed to date, we would like to use the equations of the CVA axes to tell us something about the geometric meaning of the space portrayed in Figure 4B, especially the identities of the variables most useful for group characterization/discrimination. For CVA this is more complex than for the previous ordination methods we’ve discussed.

The first interpretational complication arises because of the nature of the mathematical operations implicit in CVA. In Figure 4B the CVA axes are portrayed (correctly) as being orthogonal to one another. But recall the PCA that produced those axes was undertaken on a series of group centroid locations that had already been transformed from their original positions through rotation (Fig. 2) and differential scaling (Fig. 3). In order to determine how the CVA axes relate to the original variables it’s not enough to simply inspect the CVA axis loadings because those refer to the rotated and scaled variables. Rather, we must undo these prior transformations in order to understand how the CVA loadings relate to the original data. Figure 5 illustrates the results of these back-transformation operations.

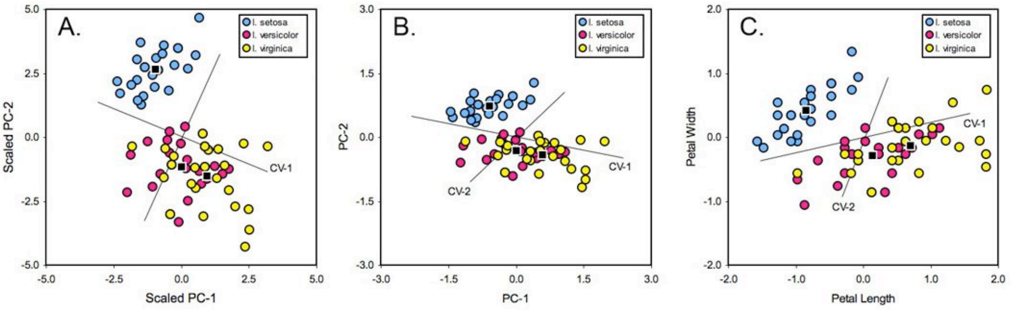


Figure 5. Back-calculation of final CVA axis orientation through the intermediate stages of the canonical rotations and scalings. A. Orientation of final CVA axes in the space of the scaled within-groups principal components (compare to Fig. 3A). B. Orientation of final CVA axes in the space of the raw within-groups principal components (compare to Fig. 3B). C. Orientation of final CVA axes in the space of the original variables (compare to Fig. 2).

The really important thing to note in these diagrams is that, unlike PCA axes which are orthogonal in the space of the original variables (see Fig. 2A), CVA axes are usually non-orthogonal (i.e. not oriented at right angles to one another) in the space of the original variables. This makes CVA axes more difficult to interpret because the same original variable(s) may have a dominant influence on the projection of specimens onto more than a single CVA axis. In this particular *Iris* analysis petal length is the dominant variable involved in separation of *I. versicolor* and *I. virginica*, but petal width has a strong influence on group separation as well. Regardless, these variables are not very efficient discriminators of those groups. Both petal length and width variables are also involved in the discrimination between *I. setosa* and the other two species. Because the traces of both CVA axes exhibit positive slopes in the space of the original variables, *their relative proportions of influence are similar*. But in the latter case the discrimination efficiency is much better. By comparing the sequence of Figure 5 plots we can also trace the alignment of the final CVA axes with the dominant modes of within-groups and between-groups variation.

There is yet another problem with the assignment of importance to the canonical variables, though. Campbell and Atchley (1981) note that many authors assess importance of the variables to between-groups discrimination by scaling the canonical variate loadings by the standard deviations of the pooled within-groups variables. This operation produces a crude and *ad hoc* measure of the correspondence between high levels of variation in aspects of the sample and alignment of the between-groups discriminant axes. The fly in the ointment here is covariation. If two variables covary to a substantial degree both could be identified as having either a large or small importance with respect to group discrimination, whereas one may be the real driver of this relation and the other a more passive passenger. Campbell and Reyment (1978, see also Campbell 1980) advocate analysis of the stability of the CV loadings as a method of approaching this problem and have developed the method of ‘shrunken [CVA] estimators’ to be used in this context.

Now that we understand exactly what CVA is doing to our data we can briefly review the mathematics used in contemporary approaches to implementing this method (in which several of the steps outlined above are combined) and undertake an expanded example analysis using the full 3 group, 25 specimen, and 4 variable *Iris* dataset.

The modern algorithm is based on the parallel between CVA and the statistical procedure known as analysis of variance (ANOVA). We begin with the *T*, *W*, and *B* matrices calculated in precisely the manner given by equations 11.1, 11.2, and 11.3 (above). Rather than undertaking the separate rotation and scaling operations outlined in our previous geometric dissection of the method, these steps are combined by noticing that the quantity we are after is a set of axes that are aligned with the maximum differences between the *B* and *W* matrices. In effect we need to maximize the *B/W* ratio. Without going into the precise matrix equation derivation, suffice it to say that this ratio will be maximized by calculating the first eigenvector (principal component) of the $W^{-1}B$ matrix[3]. Subsequent eigenvectors of this matrix represent subdominant modes variation that contribute most (in a major-axis sense) to maximizing the distinction between *B* and *W*. Together, this set of eigenvectors will represent the best single set of discriminant axes that can be calculated for the sample. Of course, since discrimination between groups is the focus of this analysis there will be one fewer eigenvectors than the number of groups in the dataset that are assigned positive eigenvalues.

A minor complication arises owing to the fact that the $W^{-1}B$ matrix will not be symmetric. This is a direct reflection of the implicit differential scaling of *B* by the within-groups eigenvalues. Fortunately, this situation is easily handled by employing singular value decomposition (SVD) as the basis for decomposition of the $W^{-1}B$ matrix. Recall that the eigenanalysis of a non-symmetric matrix produces non-orthogonal eigenvectors in the context of the original variables, which we have also seen is the case for CVA axes (see Fig. 5).

Applying these relations to the full *Iris* dataset (Table 1), the total, within-groups, and between-groups matrices are given below.

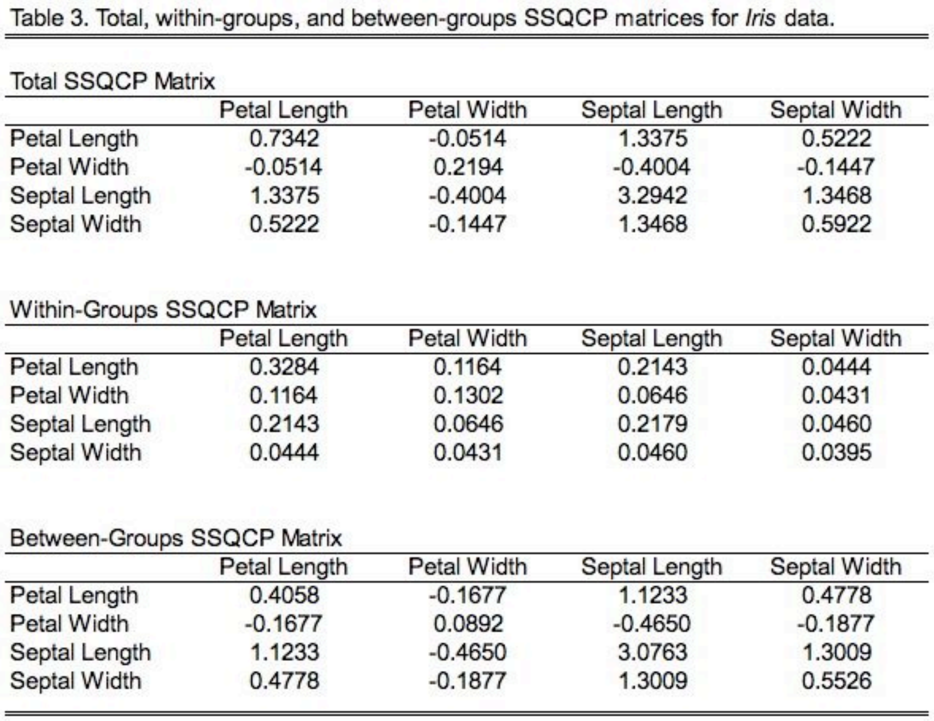


Table 3. Total, within-groups, and between-groups SSQCP matrices for *Iris* data.

Total SSQCP Matrix

	Petal Length	Petal Width	Septal Length	Septal Width
Petal Length	0.7342	-0.0514	1.3375	0.5222
Petal Width	-0.0514	0.2194	-0.4004	-0.1447
Septal Length	1.3375	-0.4004	3.2942	1.3468
Septal Width	0.5222	-0.1447	1.3468	0.5922

Within-Groups SSQCP Matrix

	Petal Length	Petal Width	Septal Length	Septal Width
Petal Length	0.3284	0.1164	0.2143	0.0444
Petal Width	0.1164	0.1302	0.0646	0.0431
Septal Length	0.2143	0.0646	0.2179	0.0460
Septal Width	0.0444	0.0431	0.0460	0.0395

Between-Groups SSQCP Matrix

	Petal Length	Petal Width	Septal Length	Septal Width
Petal Length	0.4058	-0.1677	1.1233	0.4778
Petal Width	-0.1677	0.0892	-0.4650	-0.1877
Septal Length	1.1233	-0.4650	3.0763	1.3009
Septal Width	0.4778	-0.1877	1.3009	0.5526

The basis matrix for the CVA analysis, then, is as follows.

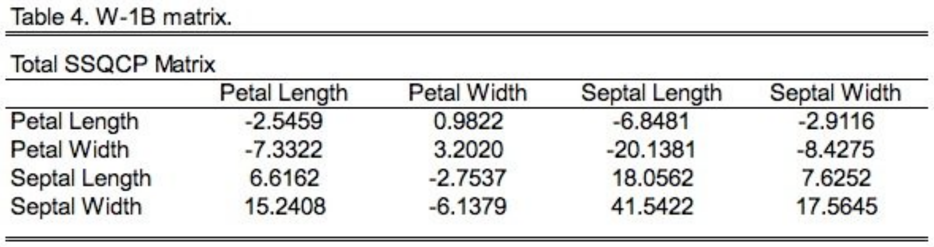


Table 4. W-1B matrix.

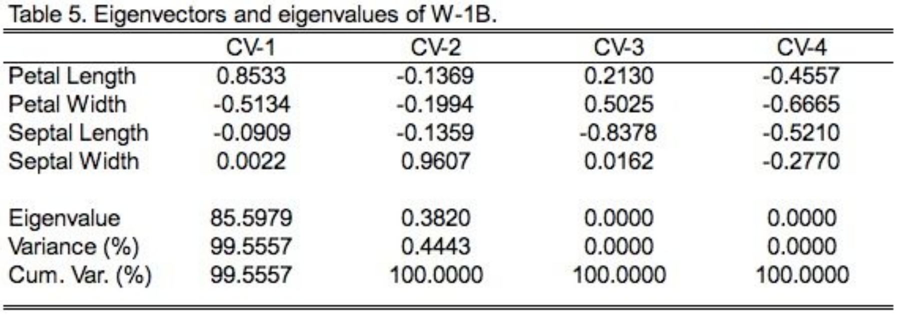
Total SSQCP Matrix

	Petal Length	Petal Width	Septal Length	Septal Width
Petal Length	-2.5459	0.9822	-6.8481	-2.9116
Petal Width	-7.3322	3.2020	-20.1381	-8.4275
Septal Length	6.6162	-2.7537	18.0562	7.6252
Septal Width	15.2408	-6.1379	41.5422	17.5645

[3] Recall the -1 superscript refers to the inverse of a matrix. The $W^{-1}B$ matrix then is the matrix formed by the between-groups SSQCP matrix being pre-multiplied by the inverse of the within-groups SSQCP matrix (see example calculations in the *PalaeoMath101* Excel spreadsheet at http://www.palass.org/modules.php?name=palaeo_math&page=1.

Note the non-symmetrical form of this matrix. Decomposition using SVD yields the following eigenvectors and eigenvalues.

Table 5. Eigenvectors and eigenvalues of W-1B.



	CV-1	CV-2	CV-3	CV-4
Petal Length	0.8533	-0.1369	0.2130	-0.4557
Petal Width	-0.5134	-0.1994	0.5025	-0.6665
Septal Length	-0.0909	-0.1359	-0.8378	-0.5210
Septal Width	0.0022	0.9607	0.0162	-0.2770
Eigenvalue	85.5979	0.3820	0.0000	0.0000
Variance (%)	99.5557	0.4443	0.0000	0.0000
Cum. Var. (%)	99.5557	100.0000	100.0000	100.0000

Observe there are only two eigenvectors with non-zero eigenvalues. These are the canonical variate (= discriminant) axes. Plotting the original data in the space of these two axes produces the following ordination.

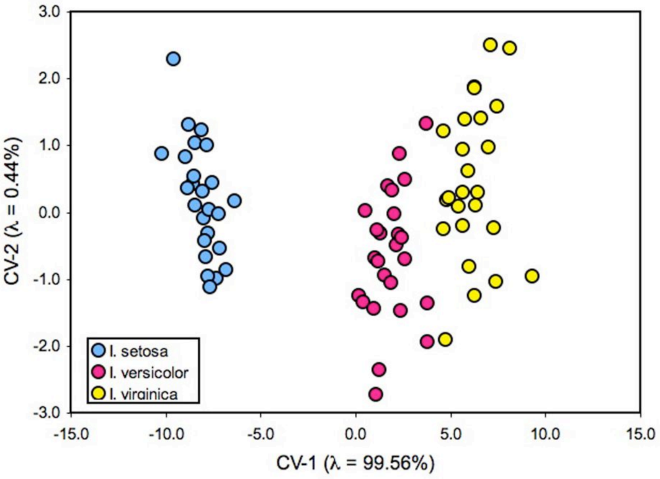
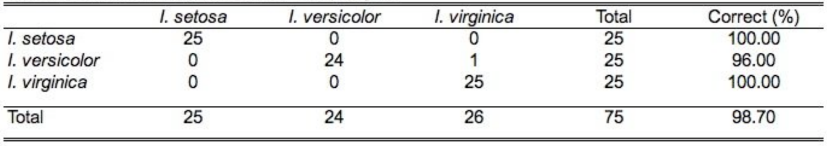


Figure 6. Scatterplot of *Iris* data in the space of the two CVA axes.

Once again, *I. setosa* is well separated from *I. versicolor* and *I. virginica*. Unlike the previous two-variable result, however (see Fig. 4B), the presence of the additional septal variables allow a much better discrimination between these latter two species, albeit along the same CV axis. The fact that CV-2 plays such a small role in between-group discrimination is reflected in its small eigenvalue. This far superior discriminant result is reflected in the confusion matrix for the analysis which measure the ability of the variable set to characterize-discriminate between the different groups.

Table 6. Confusion matrix for the *Iris* CVA analysis



	I. setosa	*I. versicolor*	*I. virginica*	Total	Correct (%)
I. setosa	25	0	0	25	100.00
I. versicolor	0	24	1	25	96.00
I. virginica	0	0	25	25	100.00
Total	25	24	26	75	98.70

There are many variants to this generalized procedure, as there are with all the methods we've covered. The important thing, as always, is to understand the basic concepts so you can make appropriate interpretations of the results reported by any programme.

As I hope you can appreciate now, CVA is very different from PCA, principal coordinates (PCoord), factor analysis (FA), correspondence analysis (CA), and the other data analysis procedures we've discussed to date. Whereas it wouldn't make much sense to (say) perform a PCA analysis and then submit the result to a correspondence analysis, there is an inherent logic to submitting the results of a PCA to a CVA. For example, PCA could be used to gather the principle sources of variation in the raw data into a small number of composite variables. Then these could be used as the basis for optimal discriminant functions.

A final word on the 'supernatural' aspects of CVA (and other multivariate methods). As should seem obvious to you by now, multivariate procedures are absolutely dependent on using sets of specimens to estimate the within-groups and between-groups geometry of their variables or measurements. Single-group methods (e.g., PCA, PCoord) focus only on within-groups variation while multiple-group methods (e.g., PLS, CVA) focus on the within-groups and between-groups distinctions. In the *Iris* dataset we saw dramatic improvement in the between-groups discrimination resulting from addition of two variables: septal length and septal width. Generally speaking the more sources of information you have about a system of measurements the better. But this improvement comes at a cost.

Consider a square space containing 100 evenly spaced points. If the square is 10 units on a side the inter-point distance is 0.010. That's pretty good characterization of the space. However, if we turn the square into a cube by adding another variable the same number of points only achieves an inter-point spacing of 0.1. That's an order of magnitude reduction in our information about the space in which our data reside. In order to achieve the same resolution in the cube space I'd need to increase sample size to 1000. If we added additional variables we'd need to increase sample size to … you get the picture.

Adding variables to a multivariate problem almost always results in a substantial drop in resolution. This is called the 'curse of dimensionality' (Belman 1957). The effects of the curse are especially noticeable in discriminant analyses because we're trying to estimate the character of within-groups variation *and* between-groups variation. For the *Iris* dataset, because the number of variables was small and the number of specimens relatively large our CVA analysis was able to pick up major differences in *W* and *B* despite the fact that addition of the septal variables resulted in an overall loss of resolution. In other words, we beat the curse of dimensionality, this time. If you undertake multivariate procedures be mindful of the curse and don't expect to beat it all the time.

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

REFERENCES

BARTLETT, M.S., 1951. An inverse matrix adjustment arising in discriminant analysis. Annals of Mathematical Statistics, **22**, 107–111.

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

CAMPBELL, N., 1980. Shrunken estimators in discriminant and canonical variate analysis. Applied Statistics, **29**, 5–14.

CAMPBELL, N.A. and ATCHLEY, W.R., 1981. The geometry of canonical variate analysis. Systematic Zoology, **30**, 268–280.

CAMPBELL, N. and REYMENT, R.A., 1978. Discriminant analysis of a Cretaceous foraminifer using shrunken estimators. Mathematical Geology, **10**, 347–359.

CULVERHOUSE, P., in press. Natural object categorization: man vs. machine. In: N. MacLeod (Editor), *Automated taxon recognition in systematics: theory, approaches and applications*. The Systematics Association, Boca Raton, Florida.

FISHER, R. A. 1936. The utilization of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**, 179–188.

MacLEOD, N., 1998. Impacts and marine invertebrate extinctions. In: M.M. Grady, R. Hutchinson, G.J.H. McCall and D.A. Rotherby (Editors), *Meteorites: flux with time and impact effects*. Geological Society of London, London, 217–246.

MAHALANOBIS, P., C. 1936. On the generalized distance in statistics. *Proceedings of the National Academy of Science, India*, **12**, 49–55.

RAO, C.R., 1952. Advanced statistical methods in biometric research. John Wiley and Sons, New York.

Don't forget the *Palaeo-math 101* web page, now at a new home at:
http://www.palass.org/modules.php?name=palaeo_math&page=1