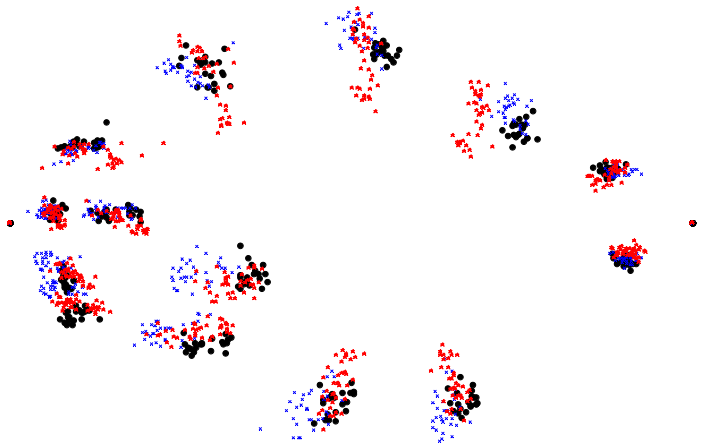
IMP: CVAGen6

H. David Sheets, Dept. of Physics, Canisius College, Buffalo, NY 14208, sheets@canisius.edu Last Altered July 25, 2006,.

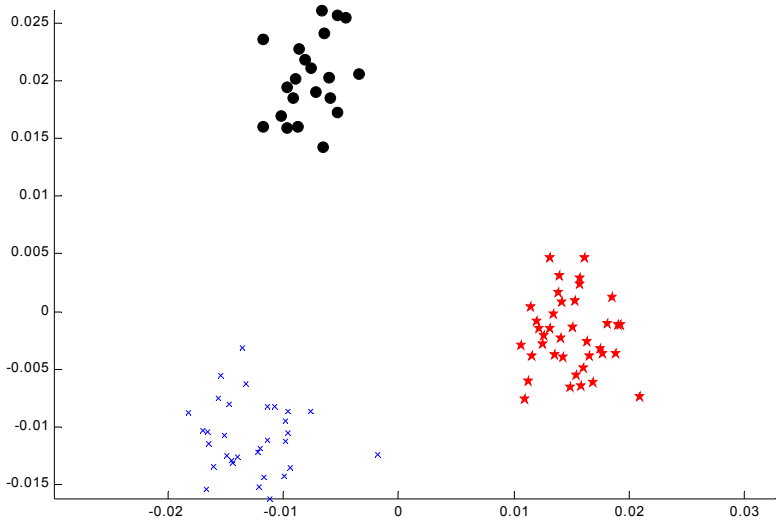
Introduction:

This is a Canonical Variates Analysis program for the analysis of shape, based on partial warp scores, and is part of the IMP software series. Canonical Variates Analysis is a method of finding the set of axes (or linear combination of variables) that allows for the greatest possible ability to discriminate between two or more groups. The program computes partial warp scores to a common reference and then does a Manova followed by a CVA. It determines how many distinct CVA axes there are in the data at a p=0.05 level of significance, and computes the canonical variates scores of all the specimens entered. It also uses Mahalanobis distances to assign all the loaded specimens to one of the groups loaded.

The image below shows the Bookstein coordinates of variables from the example file, threepir.txt, which contains three groups of piranha.

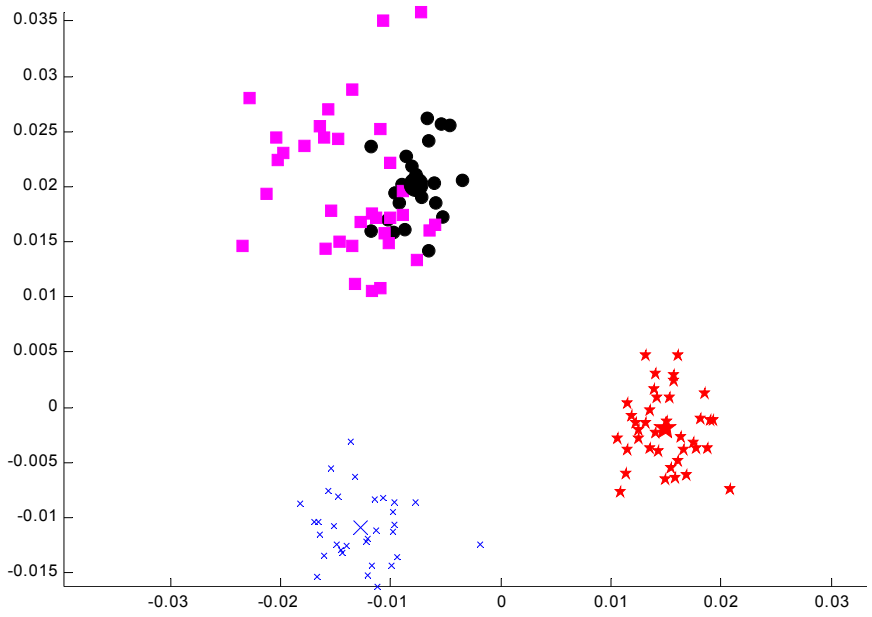


There are two distinct canonical variates axes for this set of three groups, and the canonical variate scores of these same specimens plotted along the canonical variates axes shown below illustrate the clear ability to separate these three groups.



There are a couple of recent additions to CVAGen (as of June 2002), including the ability to plot the mean of each groups CVA axes scores on a CVA axes plot, the ability to use the CVA axes determined by a number of known groups to assign unknown specimens to one of the known groups (using the Mahalanobis distance from the specimen to the mean of the nearest group). The validity of the assignments of specimens (from either a known group or an unknown) group can then be assessed using an assignment test that estimates the probability that a given specimen could be at a given observed Mahalanobis distance from the group mean by chance (based on the approach outlined by Cornuet et al. 1999).

The figure below shows the same set of three groups of piranha, with an "unknown" group added to the analysis, and the mean value of the CVA scores of each group also shown on the plot. The mean value of the CVA scores for each group are the large symbols in the rough center of each group (the black circle is admittedly hard to see). The magenta symbols in this plot are shown in black.



There is an option available which will number the specimens on the CVA axis plot, and there is a zoom option on the Axis Controls menu that will allow you to zoom in on the plot.

Using CVAGen6

I haven't had time to write a complete manual for CVAGen to date, just what is presented here. The operation of the program is very similar to that used by PCAGen, so obtain that program and it's manual if you have not done so already.

File Formats

CVAGen6 uses the same file formats as the PCAGen6 program, and essentially the same control format. If you haven't used PCAGen6, obtain it, read the manual and run it before proceeding with CVAGen6. The grouplist is required for CVAGen6 (unlike the optional group list in PCAGen6). If you will be working with unknown specimens (which is an option, not a requirement), these specimens must be in a separate IMP file, with a separate grouplist for the unknown specimens as well.

Sample Files

The files threepir.txt and threepirgrp.txt are a text file of three species of piranha and the associated group file necessary to run CVAGen6. Try these files to get a sense of how the program operates.

Tests of significance of the CVA axes used in CVAGen

The tests of significance of the canonical variate axes in CVAGen are all based on the Wilk's lambda (λ) value, which is the sum of squares within groups divided by the total sum of squares within and between groups

$$\lambda = |W| / |W+B| = |W| / |T|$$

Bartlett's test then uses the following test statistic

$$X^2 = -(w - (p-b+1)/2) \ln\lambda$$

where X^2 has an approximately chi-squared distribution, w is the degrees of freedom for the within group sum of squares, b is the degrees of freedom for the within group sum of squares and p is the number of variables, to determine if there are g = b+1 distinct groups. The degrees of freedom within is w=n-b, where n is the total number of samples and g is the number of groups.

The results of these tests will appear onscreen in a new window (the Auxillary Results Box) when you load the group list file. The contents of this window may be appended to a text file, by using the Append Results to File button on the bottom of the window. Note that if the filename you specify does not already exist, the Append Results to File Button will create a new text file to put the results in. You might sometimes get a warning message about overwriting a file. Don't worry about this message when using this option, it won't really overwrite the file, but rather will append it.

Output from Sample files

Below are the results obtained from the sample files. The first section lists the Wilk's Lambda value, chi-squared value and degrees of freedom as well as the p score for the various number of CVA axes (not axis, gotta fix that!).

```
Results from CVA/Manova
Axis 1 Lambda= 0.0031 chisq=430.9308 df=56 p<2.22045e-016
Axis 2 Lambda= 0.0684 chisq=199.7954 df=27 p<2.22045e-016
Groupings from CVA-Distance Based

 Original Groups along rows, CVA groups along columns
 -
 -   0   1   2   3
 -   1  21   0   0
 -   2   0  32   0
 -   3   0   0  38
```

The second section shows the assignments of specimens to various groups, based on the Mahalanobis distance in the space defined by the significant CVA axes. The original groups are along the rows of the matrix, the predicted group (from the CVA) is along the columns. In this case, all specimens were assigned to the correct groups. The Assignments test discussed later can be used to assess the reliability or validity of these assignments.

In the second example shown below, there were 9 groups and 8 significant axes. Membership assignment was always correct for groups 1 and 3, but specimens in Group two were also asigned to groups 7 (4 specimens) and 8 (1 specimen). One specimen from group 4 was assigned to group 5, one specimen from group 5 was incorrectly assigned to group 6, and so forth.

```
Groupings from CVA-Distance Based

 Original Groups along rows, CVA groups along columns
 -
 -   0   1   2   3   4   5   6   7   8   9
 -   1  31   0   0   0   0   0   0   0   0
```

-	2	0	16	0	0	0	0	4	1	0
-	3	0	0	32	0	0	0	0	0	0
-	4	0	0	0	45	1	0	0	0	0
-	5	0	0	0	0	37	1	0	0	0
-	6	0	0	0	0	7	33	0	0	0
-	7	0	9	0	0	0	0	77	3	0
-	8	0	1	0	0	0	0	1	32	0
-	9	0	0	0	0	0	0	0	0	37

Displaying Results

CVA Axis

The program can plot out the deformation implied by the CVA axes or the change in shape obtained by regressing the shape on the CVA axis scores. There is a pushbutton labeled *Regr?* on the screen in the Display CVA Axes section that control which you will see, if this button is pushed (black) you will see the regression deformation. The regression shows you all the changes in shape that are correlated with the CVA axis score, the CVA axis itself does not show all the correlated change in shape.

Show Deformation Implied by CVA

These options let you place two markers on the scatter plot of CVA scores, and then shows you the deformation implied by the markers, relative to the reference form. This allows you to see the effects of differences along two CVA axes at the same time.

Assignment of specimens to groups

A simple Mahalanobis distance-based approach is then used to determine which group each specimen belongs to, based on the canonical variate scores. The predicted group membership of each specimen based on the CVA scores is determined by assigning each specimen to the group whose mean is closest (under the Mahalanobis distance) to the specimen. Note that this is not a sophisticated approach to partitioning the space determined by the CVA axis, but it was algorithmically simple. I would like to find something better than this, but for now, it is better than nothing.

CVAGen produces a matrix showing group assignments, the original (user-assigned) groups are rows of the matrix, the new groups are along the columns of the matrix. The rows and columns are labeled with the group codes from the group list file input into the program. Study of this matrix allows the user to assess how effective the CVA is at separating groups, and which groups were not successfully separated. If you have a set of unknown specimens, their assignments will be listed on a separate row (labeled u) below the matrix.

Example of the group assignments matrix:

Groupings from CVA-Distance Based

Original Groups along rows, CVA groups along columns

-

- 0 1 2 3

- 1 21 0 0

- 2 0 32 0

- 3 0 0 38

-u 34 0 0

This example indicates that of the 21 specimens assigned by the user to group 1, all were assigned by the CVA process to group 1, of 32 specimens in the user's group 2, all were assigned to group 2 and of the 38 specimens in the user's group 3, all were assigned to the user's group 3. All of the thirty four unknowns (actually a different species), all of them were assigned to group 1.

Jack-knife test of group assignments-(*added March 2005)*

In the CVAGen6n version, I have added the ability to jackknife the assignment of specimens to groups. During this procedure, a single specimen is excluded from the CVA analysis at a time, and the entire analysis is repeated without that particular specimen. The excluded specimen is then assigned to one of the groups using the CVA axes derived from other specimens. This procedure is repeated sequentially for all specimens in the data set. The rate of correct assignments of excluded specimens is then calculated based on these results. This jackknife procedure may be a better estimate of the performance of the assignment procedure, since the assigned specimens are not used during the axes calculation, avoiding some level of circularity in the procedure. This test is available on the statistics menu.

PCA Reduction of the Number of Variables (*added March 2005)*

The CVA analysis requires a matrix inversion of the pooled variance-covariance matrix of within group variation in shape. If this matrix is not of full rank, it cannot be inverted and the program will produce an error. The matrix will not be of full rank if there are more measurements per specimen than total specimens in the analysis. This is particularly a problem when using semi-landmarks, which greatly increase the number of measurements per specimen, particularly since semi-landmarks have only one degree of freedom but two measurements.

One approach is to carry out a PCA analsis of the data, and discard some portion of the PC axes scores to reduce the number of variables. In fact, if we discard only the PCA

axes with zero eigenvalues (which explain no variance), we will not lose any of the variance in the original data. This means that if we use a number of PC axes equal to one less than the specimen count, no variance is lost and difficulties related to matrix inversions in the CVA are eliminated.

To use this option in CVAGen6n, load your specimens, then set the number of PCA axes to use and select the PCA reduction option on the lower left edge of the screen, before loading the group code file. Note that once the group code is loaded, the CVA analysis is executed, so you must select the PCA reduction prior to this point.

Working with a Set of Unknown Specimens

One possible use (to be regarded with some caution, despite it's appeal!) of the CVA is to use the CVA axes to assign an unknown specimen to one of the known groups. To carry out this procedure, the unknown specimen (or specimens) must be digitized, using the same set of landmarks as in the known specimens. The "unknown" specimens are then placed in an IMP format file, and a grouplist is constructed for the unknown specimens (this grouplist will be used only for color-coding, not in calculations).

Once you have the "unknowns" file and grouplist prepared, load your known specimens and their grouplist into CVAGen6 and carry out the usual CVA analyses. The known groups are used to determine the CVA axes, the unknowns are not used during the axis construction. After the axes are constructed successfully, use the *Load Unknowns File* option from the *File* Menu to load the unknowns. Next, use the *Load Unknown*

Group List option from the *File* Menu to load the group list for your unknown specimens. When you load this file, the program will plot the landmark configurations of all specimens (known and unknown) in Procrustes Superimposition on the reference form used for the known specimens. You may now plot the CVA axes, and the unknown specimens will also be shown on this plot.

To see the assignments of specimens (known and unknow) to groups, you can see a summary table of group assignments by using the *Show Groupings by CVA* on the *Statistics* menu, or you can use the *Run Assignments Test* on the *Statistics* menu.
If you are using the PCA reduction method, the unknown specimens will be analysed using their scores along the PC axes determined by the known specimens.

The Assignments Test

The assignment test discussed hereafter is modeled on an assignment test in the literature (Cornuet et al. 1999, Piry and Cornuet 1998), and appears in *Nolte and Sheets* (2005)

The specimens are assigned to groups by determining the Mahalanobis distance (along the significant CVA axes only) of each specimen from the mean value of the CVA scores for each group, and then assigning each specimen to the closest group. Note that the distanced used is not a variant of a Procrustes distance, but a Mahalanobis distance. This is one of the few instances in the IMP software where a distance other than a Procrustes distance is used.

The question that should then arise is how do we assess the validity or probability of these group assignments? How "good" are the assignments?

The approach used here follows the distance-based approach used by Cornuet et al. (1999). The distribution of distances produced by a Monte Carlo simulation is used to determine if the observed distance of a given specimen is consistent with the null model of random variation around the mean of the group to which the specimen is assigned to. The distance from a specimen to a group mean can then be assigned a p-score which describes how likely the specimen is to be a member of a group (under the null model used in the Monte-Carlo simulation). If the p score is smaller than 5%, then we can assert that there is a less than 5% chance that random variation could have produce a distance as large as that of the particular specimen from the group mean, and hence that the assignment of that specimen to the group is in doubt. A very low p score (1% or less) means that it is highly unlikely by chance that the specimen does belong the group it was assigned to. It should be noted that in a study with many specimens, a number of them will have low p-scores by chance (the Bonferroni problem), and so to assess the validity of the assignments of the set as a whole, the researcher should assess the number of specimens expected to have p values less than 5%. It will then be possible to determine if the observed number of low p values exceeds that expected by chance.

Details of the Monte-Carlo simulation of the distribution of Mahalanobis Distances around each group mean.

The model used in the Monte Carlo simulation of the distribution of Mahalanobis distances of specimens within a group around the group mean is based on a normal model of the distribution of the CVA scores of each group about the mean of that group. For a given group, it appears possible that the CVA scores along each CVA axes for the specimens within the group are correlated, that there exists within each group a covariance structure to the CVA scores of specimens within the group. In carrying out the Monte-Carlo simulation of the distribution of distances of specimens within the group from the mean, we need to preserve this covariance structure to produce a valid model of the distribution. To do this, we form the variance-covariance matrix of the CVA scores within a group, and then do an eigenvalue decomposition of this variance-covariance matrix to find the principal component axis of the within group variance-covariance matrix of CVA scores. This yields the same number of variables as the CVA scores, but now with uncorrelated axis (the eigenvectors) each of which has a variance given by the corresponding eigenvalue. So we may now form a model of the distribution that assumes the population has an independent random normal distribution along each of these eigenvectors (principle component of variation axes), with amplitudes given by the square roots of the eigenvalues (remember the eigenvalues are the variances of the group along the corresponding eigenvectors, so that the square root of the eigenvalue is the standard deviation of the population along that eigenvector).

So an independent, normal distribution with a known amplitude is assumed along each eigenvector. This allows us to generate a Monte Carlo population of specimens, assuming the independent normal distribution along these principle component axis. Each specimen is generated using a random number generator to compute locations along the eigenvectors, which are then translated back into CVA axes scores. *The Monte Carlo generated CVA axis scores will have the same mean and variance-covariance structure as the original population did.* We can now calculate the distance from the group mean of each of the specimens generated by the Monte Carlo procedure and from these generated the distribution of distances under the null model of random variation that we which to test the Mahalanobis distances used for assignments against.

If the observed Mahalanobis distance exceeds a percentage 100(1-p) of the Monte Carlo distances, then the observed Mahalanobis distance has a probability p of being produced by the null model inherent in the Monte Carlo simulation. A low p value is grounds to assert that the observed Mahalanobis distance is not consistent with the null model of random variation, and would lead us to doubt the validity of the assignment.

Running the Assignments Test

Load all your specimens and grouplists (known and unknown data, as desired). Then use the *Run Assignments Test* on the *Statistics* menu. Be a bit patient, this is a substantial simulation. The program will then prompt you for a file name for the output file produced by the analysis.

The assignment of all known specimens will be shown first in the output file, an example is shown below:

```
For the specimens with user group assignments
Specimen GroupSymbol OrdinalGroup AssignedGroup AxesDistance Signif.
 1          1           1           1           0.382469 p>0.05
 2          1           1           1           1.793519 p>0.05
 3          1           1           1           0.725210 p>0.05
 4          1           1           1           1.880006 p>0.05
 5          1           1           1           1.877013 p>0.05
 6          1           1           1           0.513559 p>0.05
```

The first column is the specimen number in the input data file, the second value is the group code assigned by the user to that specimen (taken from the Group Code file). The third column is the *ordinal number* of the group to which the user has assigned the specimen. The software keeps track of groups by ordinal number, rather than by using the assigned group codes (don't ask why…thanks…). I will fix this eventually, but for the moment, you will have to look at the file to determine which of your group codes corresponds to ordinal number 1, which group code is two etc. I really should fix this…

The fourth column is the ordinal number of the group which the CVA analysis has assigned the specimen. The fifth column is the axis distance, which is to say the distance of that specimen from the group mean of the group which the CVA analysis has assigned that specimen. The p-score in the 6^{th} column is the probability under the Monte Carlo model described earlier that the observed distance (in the 5^{th} column) could have arisen by chance. For the specimens in the above example, all the p values are above 5%, meaning we could not reject the null of the Monte Carlo model, so we have no statistical reason to doubt these assignments.

Following the assignment of all known specimens (which had a user assigned group initially) is a listing of the assignments of all unknown specimens, if unknown

specimens have been loaded into CVAGen6. An example of this section of the output file is shown below:

```
For the specimens with unknown group assignments
Specimen GroupSymbol OrdinalGroup AssignedGroup AxesDistance Signif.
 1          u           u           1           3.122955 p<0.001
 2          u           u           1           2.674542 p<0.001
 3          u           u           1           1.376086 p>0.05
 4          u           u           1           1.580605 p>0.05
```

The first column is the specimen number in the unknown file, the group symbol and Ordinal number are both unknown (u) since the group affinities are not known to the user. The fourth column is the group (by ordinal number, see discussion above) of the group the CVA procedure has assigned the specimen. The 5th column is the Mahalanobis distance of the specimen to the mean of the group to which it was assigned, followed by the probability that this distance could have arisen by chance under the null model used in the Monte Carlo simulation. In addition to examining the p-scores in this example (which would allow us to reject the null model for several specimens in this case), it may perhaps be useful to compare the distribution of distances in for the known specimens versus the unknown specimens, as shown in the output file. The unknown specimens have typically larger distances than the known in this example, leading us to suspect that the assignments of these unknown specimens to group 1 is probably not consistent with the null model. See the CVA axis plot showing the distribution of the magenta symbols shown earlier in this document.

Additions to the Assignment Test, *Sept. 2004*

Lorenz Hauser (U. of Washington) suggested adding the p-values for each group to the output file, rather than just the p-value for the closest group. Having this information in the output file for the assignment test allows you to determine if an

assignment is unique or not. If we can reject the hypothesis that the specimen is within the typical range (95%) of all the groups except one, then the specimen may be uniquely assigned to that one group.

The option "Run Detailed Assignments Test" on the statistics menu will run the assignment test and output a file with the specimen assignment information, the Mahalanobis distance to each of the group means and the p-value associated with that particular distance for each group, as well as a code indicating whether or not the assignment was unique. An assignment is said to be unique if the specimen falls within the 95% confidence interval ($p>0.05$) of one and only one group. If the assignment is unique, the code will indicate which group it belongs to, a unique code of zero means there is no unique grouping.

Further Changes to support use of the assignment test, *November 18, 2004.*

After some suggestions from Arne Nolte, I have added two features to make the assignment test in CVAGen more useful.

The first is an output file option to save the CVA scores of the unknown specimens to a file. This option is accessed on the file menu as "Save Unknown Spccimens CVA Scores". The scores of the unknowns are saved in the same order they appear in the input file of unknowns. You can not use this option until you have loaded the unknows and the unknown group list.

The second is a jackknife test of the effectiveness of the assignment test. In this jackknife procedure, a percentage of the known input data is randomly chosen as "unknowns". The CVA analysis is then carried out on the remaining known data, and the

"unknowns" are assigned to one of the known groups. The assignment test is used to determine whether each assignment is significant at 5% or not. This is carried out for a large number of jackknife test sets. The distributions of the 4 possible outcomes:

-correctly assigned and judged significant

-correctly assigned and judged non-significant

-incorrectly assigned and judged non-significant

-incorrectly assigned and judged significant

If the discrimination and assignment test were working perfectly, we would expect all assignments to be correct and 5% of them to be judged non-significant (due to the fact we are working with estimated 95% confidence intervals in the assignment test). Substantial numbers of incorrect assignments would be very worrisome, particularly if they are judged as significant. The jackknife procedure gives us some sense of how effective we can expect the discrimination and assignment to be, given a specific data set.

To run the jackknife test of the functioning of the assignment test, load your data and group list for your known specimens as usual. Then select the "Jack-knife Assignments" option on the statistics menu. A pop-up window will then ask you to enter the number of jackknife resamplings to carry out (the default is 100) and the fraction of the data to use as unknowns in the jackknife (the default is 0.10 or 10%). Of hand, I would suggest using 100 to 1000 jackknife sets, using 1% to 20% of your data as "unknowns". When you hit the "Okay" button, the program will run the jackknife test, which may take up to 10 minutes to run, depending on your computer speed. I would suggest running it using 10 jackknife sets the first time, just to get a sense of how long it will take to run on your computer given your data. The output will appear in the

Auxillary Output window (AuxBox). The example below shows the results of 100 jackknifes of the threepir.txt data set:

Results of the Jack-knife test of the CVA performance
During this test using 91 total specimens
10.0 percent (10) were used as "unknowns"
A total of 100 "unknowns" were tested in 10 trials

96 (96.0 percent) were correct and significant

4 (4.0 percent) were correct and non-significant

0 (0.0 percent) were incorrect and significant

0 (0.0 percent) were incorrect and non-significant

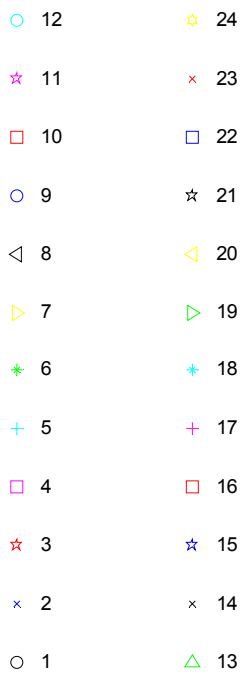
Note that the output indicates the number of specimens in the entire data set, the percentage and number used as “unknowns” in each trial, and the overall percentages of each of the possible outcomes. This particular trial looks very good, with no incorrect assignments and only 4% judged non-significant. We would expect roughly 5% to be non-significant by chance.

Alterations 7/25/06 (CVAGen6o)

1.) I altered the behavior of the copy image to eps function, which had been cutting the edges off of some images, I think that it is working correctly now.

2.) Fixed the plot density=30 setting so it works now.

3.) Added some more group codes (up to 24 groups now). The symbols are shown below:



Literature Cited:

Cornuet, J-M., S.Piry, G. Luikart, A. Estoup and M. Solignac. 1999. New Methods Employing Multilocus Genotype to Select or Exclude Populations as Origins of Individuals. Genetics 153:1989-2000.

Nolte, A. W. and Sheets, H. D. 2005. Shape based assignment tes reveal transgressive phenotypes in natural sculpin hybrids (Teleostei, Scorpaeniformes, Cottidae). Frontiers in Zoology, 2:11, published June 29, 2005 www.frontiersinzoology.com

Piry, S. and J-M. Cornuet, 1998. Gene-Class Users Manual. http:\\WWW.ensam.inra.fr/CBGP.