BioNumerics Tutorial:

Calculating a PCA and a MDS on a character data set

1 Aim

Principal Components Analysis (PCA) and Multi Dimensional Scaling (MDS) are two alternative grouping techniques that can both be classified as dimensioning techniques. In contrast to dendrogram inferring methods, they do not produce hierarchical structures like dendrograms. Instead, these techniques produce two−dimensional or three−dimensional plots in which the entries are spread according to their relatedness. Unlike a dendrogram, a PCA or MDS plot does not provide "clusters". The interpretation of the obtained comparison is, more than in cluster analysis, left to the user. In this tutorial you will learn how to create a PCA and MDS and how to change the layout of the obtained plots.

2 Preparing the database

The **DemoBase Connected** will be used in this tutorial and can be downloaded directly from the *BioNumerics Startup* window or restored from the back-up file available on our website:

1. To download the database directly from the *BioNumerics Startup* window, click the ***Download example databases*** link, located in the lower right corner of the *BioNumerics Startup* window. Select **DemoBase Connected** from the list and select ***Database*** > ***Download***. Confirm the download action.

2. To restore the database from the back-up file, first download the file `DemoBase_Connected.bnbk` from `http://www.applied-maths.com/download/sample-data`, under 'DemoBase Connected'.

   In the *BioNumerics Startup* window, press the button, select ***Restore database***, browse for the downloaded file and select ***Create copy***. Specify a name and click <***OK***>.

In contrast to other browsers, some versions of Internet Explorer rename the `DemoBase_Connected.bnbk` database backup file into `DemoBase_Connected.zip`. If this happens, you should manually remove the `.zip` file extension and replace with `.bnbk`. A warning will appear ("If you change a file name extension, the file might become unusable."), but you can safely confirm this action. Keep in mind that Windows might not display the `.zip` file extension if the option "Hide extensions for known file types" is checked in your Windows folder options.

3 Principal components analysis (PCA)

Principal components analysis (PCA) is another way to visualize relationships among entries. Unlike a Multidimensional scaling (MDS), PCA uses the data set itself instead of the similarity matrix to measure relatedness. PCA maximizes the variation among entries along the first two or three dimensions, which can then be displayed. These are the *principal components*.

PCA does not work on sequence types and fingerprints can only be analyzed by PCA after a band matching table is generated.

1. Open the **DemoBase Connected** by double-clicking on this database in the *BioNumerics Startup* window.

2. In the *Database entries* panel of the *Main* window, select all entries using the **Ctrl+A** keyboard shortcut. Unselect the three entries defined as STANDARD using the space bar.

3. Highlight the *Comparisons* panel in the *Main* window and select ***Edit*** > ***Create new object...*** (✚) to create a new comparison for the selected entries.

4. Click on the 👁 next to the experiment name **FAME** in the *Experiments* panel to display the data in the *Experiment data* panel.

Next, we will create groups based on the content of the ”Genus” database field.

5. Click on the database field ”Genus” in the *Information fields* panel.

6. Select ***Groups*** > ***Create groups from database field*** and press <***OK***>.

Group colors are now assigned to the different ”Genus” groups.

7. Make sure **FAME** is selected in the *Experiments* panel and select ***Statistics*** > ***Principal Components Analysis...*** (▦).

8. Press <***OK***> to start the calculation of the PCA.

The *Principal Components Analysis* window is divided in different panels (see Figure 1):

- The *Entry coordinates panel* shows the entries plotted in an X-Y diagram corresponding to the first two components. In the *Components panel*, the first 20 components are shown, with their relative contribution and the cumulative contribution displayed. Also, the components used as X, Y and Z axes are indicated.

- The *Character coordinates panel* shows the characters plotted in the same X-Y diagram showing the contribution that each character has to the two displayed components, and hence, what contribution it has to the separation of the groups along the same components.

- The *Components panel* lists the principal components in the order of their contribution to overall variance. The components used as X, Y and Z axes are also indicated.

A character that appears near the edge of the plot is a *strong* discriminator, while a character near the center is a *weak* discriminator. Furthermore, a character that appears near the position of an entry is an *indicator* for that entry.

9. Switch from color indication for the groups to symbol indication with ***Layout*** > ***Show color coding*** (▮).

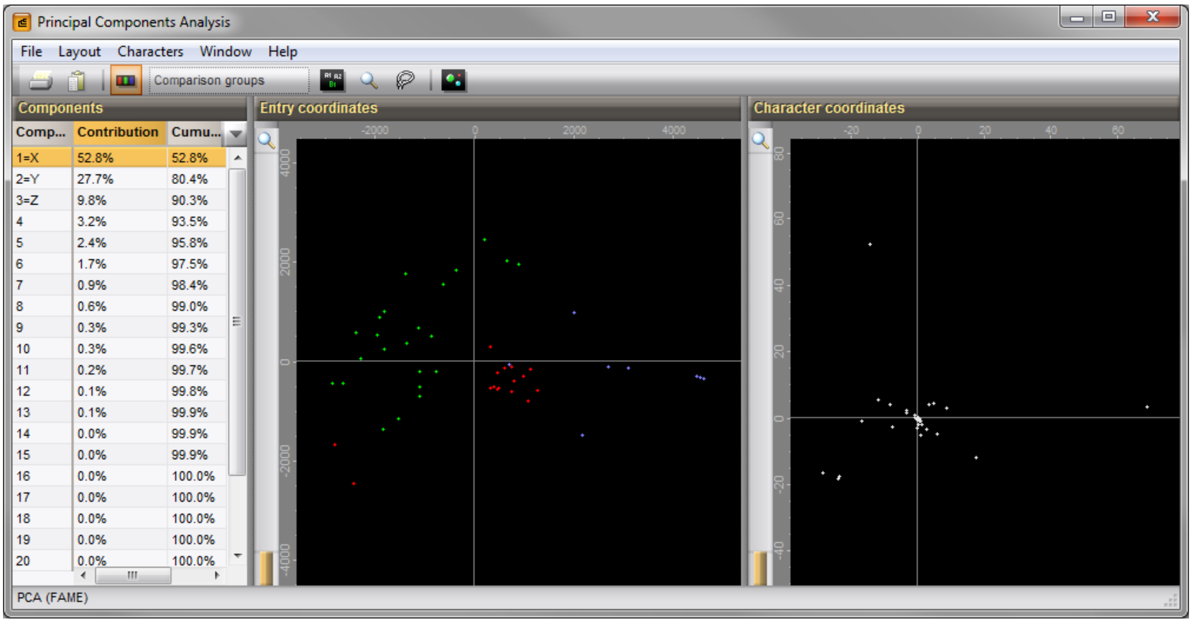
10. Show the keys or a unique label based upon the groups for the entries with ***Layout*** > ***Show keys*** (▦).

11. With ***Layout*** > ***Zoom in / zoom out*** (🔍), you can zoom in on any part of the entries or characters panel of the PCA plot: drag the mouse pointer to create a rectangle; the area within the rectangle will be zoomed to cover the whole panel.

12. In order to restore the original size of the image, simply left-click within the panel. Disable the zoom mode afterwards.

13. Entries can be selected in an *Entry coordinates panel* by holding the **Shift**-key down and selecting the entries in a rectangle using the left mouse button. Selected entries are encircled in blue. Press **F4** to unselect entries.

14. If you move the mouse pointer over the characters in the *Characters coordinates panel*, the name of the pointed character is shown.



Comp...	Contribution	Cumu...
1=X	52.8%	52.8%
2=Y	27.7%	80.4%
3=Z	9.8%	90.3%
4	3.2%	93.5%
5	2.4%	95.8%
6	1.7%	97.5%
7	0.9%	98.4%
8	0.6%	99.0%
9	0.3%	99.3%
10	0.3%	99.6%
11	0.2%	99.7%
12	0.1%	99.8%
13	0.1%	99.9%
14	0.0%	99.9%
15	0.0%	99.9%
16	0.0%	100.0%
17	0.0%	100.0%
18	0.0%	100.0%
19	0.0%	100.0%
20	0.0%	100.0%

PCA (FAME)

Figure 1: The *Principal Components Analysis* window.

15. The entry plot can be printed with ***File*** > ***Print image (entries)...*** (🖨) and the character plot can be printed with ***File*** > ***Print image (characters)...***.

16. Alternatively, the entry plot can be copied to the clipboard with ***File*** > ***Copy image to clipboard (entries)...*** (📋) and the character plot can be copied to the clipboard with ***File*** > ***Copy image to clipboard (characters)***.

17. To create a three-dimensional plot from the PCA, select ***Layout*** > ***Show 3D plot*** (●).

See 5 for more information about this 3-D representation.

18. Close the windows with ***File*** > ***Exit***.

4 Multi Dimensional scaling (MDS)

Multi Dimensional scaling (MDS) is an optimized three-dimensional representation of the similarity matrix. The Euclidean distance between two points (entries) reflects the similarity between them as well as possible, while providing a convenient visual interpretation. A similarity matrix must be present before an MDS can be calculated.

1. Select **FAME** in the *Experiments* panel of the *Comparison* window and calculate a dendrogram based on the ***Euclidean distance*** with ***Clustering*** > ***Calculate*** > ***Cluster analysis (similarity matrix)...***: select ***Euclidean distance*** in the first step, and ***UPGMA*** in the last step.

2. If group colors are no longer available, proceed as described in 3 to create groups based on the ”Genus” field.

3. Select ***Statistics*** > ***Multi-dimensional scaling...*** (▨).

4. Check ***Optimize positions*** and press <***OK***> to start the calculations.

The MDS is calculated and the *Coordinate space* window is shown (see Figure 2). The *Coordinate space* window shows the entries as dots in a cubic coordinate system.

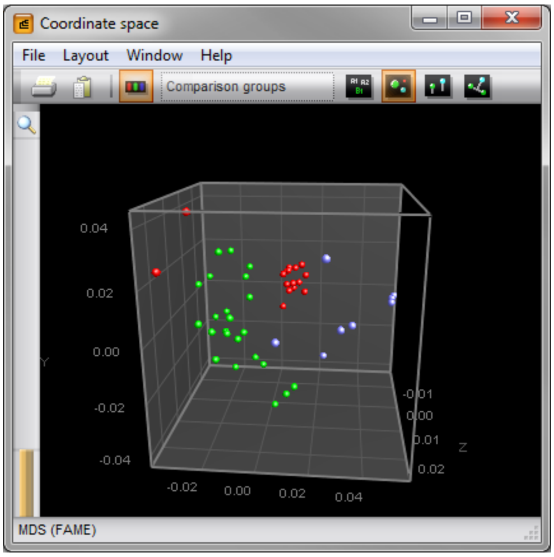


Figure 2: The *Coordinate space* window.

5 Changing the coordinate space layout

1. To zoom in or zoom out on the image, use ***Layout*** > ***Zoom in*** (**Pge Down**) or ***Layout*** > ***Zoom out*** (**Pge Up**) , respectively.
2. The image can be rotated in real time by dragging the mouse pointer in any direction.

The entries are represented as small dots, in the colors as defined in the *Comparison* window.

3. With ***Layout*** > ***Show keys*** (), you can display the database keys of the entries instead of the dots.

Entry keys may be long and uninformative for the user. The entry keys can be replaced by a group code: an alphabetical numbering of groups with an index per entry in a group. The group codes are shown as follows:

4. In the *Comparison* window, select ***Layout*** > ***Use group numbers as key***.
5. A list of entry indices as used in the PCA and the corresponding entry names can be obtained by selecting ***File*** > ***Export*** > ***Export database fields...*** in the *Comparison* window.
6. Alternatively, you can select a field in the *Comparison* window, for example the 'Strain number' field, and select ***Layout*** > ***Use field as key***.
7. With ***Layout*** > ***Show group colors*** (), you can toggle between the color representation and the non-color representation.
8. With ***Layout*** > ***Show construction lines*** (), the entries are displayed on vertical lines starting from the bottom of the cube.
9. ***Layout*** > ***Show rendered image*** () displays the coordinate system in realistic three-dimensional perspective.
10. Another very interesting option is ***Layout*** > ***Show dendrogram*** ().

The entries in the coordinate system are connected by the branches of the dendrogram that was calculated in the *Comparison* window (see Figure 3). This is an ideal combinatation to co-evaluate a dendrogram and a coordinate system.

11. The image can be printed with ***File*** > ***Print image...*** (). The image will print in color if the colors are shown on the screen.

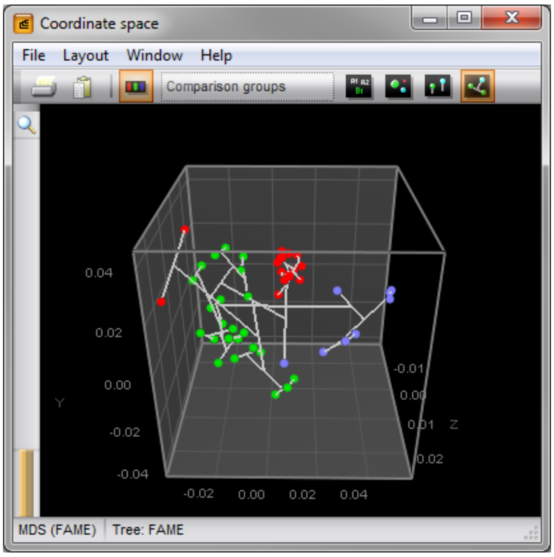


Figure 3: Co-evaluation of a dendrogram and a coordinate system.