Towards Better Graphics for Multivariate Analysis: the Interactive Factor Map

Jean Thioulouse

Laboratoire de Biométrie Génétique et Biologie des Populations, URA CNRS 2055, Université Lyon 1, 69622 Villeurbanne Cedex, France

Summary

The interpretation of the factors computed by multivariate analysis methods is often a difficult task. The main difficulties come from the cluttering of the factor map when the data set is large, and from the need to visualize the data when interpreting the factors. An interactive method based on the representation of the data on the factor maps is presented, with reference to a computer program available on the Internet network. Several techniques offered by this program are demonstrated with the help of a real-size example data set: searching, zooming, and plotting of rows and columns of the data table on the factor map.

Keywords: Multivariate analysis, exploratory data analysis, interactive graphics, factor map, principal component analysis

1 Introduction

Dynamic graphics are gaining more and more popularity with the availability of powerfull microcomputers and user-friendly graphical interfaces. Their help in analysing numerical data sets has been widely recognized, and the book by (Cleveland & McGill 1988) shows that there is a great variety of techniques in this field. Among these techniques, several deal more particularly with multivariate data. (Young, Kent & Kuhfeld 1988) present VEDA (Visual Exploratory Data Analysis) and MEDA (Multivariate Exploratory Data Analysis) as a way to explore multivariate data sets graphically with the VISUALS software. Their paper also provides a general overview of the methods available in this area, for example the grand tour (Asimov 1985), projection pursuit (Huber 1985), brushing and related methods (Becker & Cleveland 1987), 3D rotations (Donoho, Donoho & Gasko 1986). More recently, (Weihs & Schmidli 1990) have presented the OMEGA system (Online Multivariate Exploratory Graphical Analysis) as an alternative to VEDA and MEDA in the context of industrial practice.

However, as pointed out by (Thioulouse, Devillers, Chessel & Auda 1991), purely graphical methods cannot overcome the main difficulty of multivariate data analysis, namely the dimensionality reduction. Another problem, underlined by (Gower & Digby 1981) for graphics like Chernoff faces (Chernoff 1973), is the fact that graphical methods introduce subjective criteria that can be misleading. This is also true for dynamic graphics: data manipulation "by hand" on the computer screen is obviously subjective, but has the advantage of freeing the user from the constraints of classical data analysis methods. (Hurley & Buja 1990) show that user control is possible (and even required) in high dimensional rotation methods.

(Young, Faldowski & McFarlane 1993) proposed to use multivariate analysis (Principal Components Analysis, PCA and Multidimensional Scaling, MDS) in dynamic graphics. According to these authors, the problem in this case is "to present hD [high-dimensional] data information in a 2D plane, such that our 3D perception can understand the hD geometry". It may be argued however, that hD geometry is intrinsically out of reach of human understanding, and that 2D or 3D projections will only help in getting an insight into some particular features of data geometry. The principle of interactive graphical modeling in PCA proposed by (Young et al. 1993) also presents both an attractive side, because it "liberates an analysist from purely mathematical optimality criteria" ((Young et al. 1993), p. 985) and a drawback, because it introduces also a large part of subjectivity.

Multivariate analysis methods aim at summarizing high dimensional data sets with low dimensional subspaces. We do not intend to propose a new method to explore the multi-dimensional geometry of these data sets, but we want to show that the factors obtained with multivariate analyses can be interpreted more easily by using an interactive graphical display of factor maps. Moreover, these tools can be used without a thorough knowledge in statistics or computer science. Our proposal address mainly to those biologists who may have difficulties accessing more sophisticated methods like dynamic and motion graphics.

2 Presentation of the method

Factor maps, obtained by plotting two factors of a multivariate analysis, benefit from the optimality properties of these factors (e.g., maximization of the projected variance), but need to be interpreted to be really useful. The meaning of the factors is sometimes hard to find out, and several problems must be tackled: when the number of observations (or of variables) is high, the factor map is cluttered, and the user has difficulties in finding a particular observation. It is also frequently interesting to focus the investigations on a particular subset of observations instead of the whole set, or on a particular zone on the factor map. Moreover, a good interpretation of factors forces the user to go back to the data table to find out what features of the data set have been stressed by the factors. This way back to the data is uneasy if the user has to deal directly with the numerical data table, even with moderate size data sets (100 rows and 20 columns), and graphical displays reach their limits with large tables (hundreds of elements).

The ADEScatters program was written to make easier the interpretation of factor maps, by trying to overcome the above difficulties. It is part of the ADE (Analysis of Environmental Data) package for multivariate analysis (Thioulouse, Devillers, Chessel & Olivier 1995), and can be retrieved at the following URL, with a complete documentation:

http://biomserv.univ-lyon1.fr/ADE-4.html

All the methods available in this package cannot be described here, and only a short summary of the underlying mathematical model will be presented. This model is the duality diagram (Cailliez & Pages 1976); (Escoufier 1987), that is based on the concept of statistical triplet. A statistical triplet $(\mathbf{X}, \mathbf{D}_p, \mathbf{D}_n)$ is made of three matrices: the data matrix $\mathbf{X}$ (having n rows and p columns, with possibly an appropriate transformation, like centering or standardization), the matrix of row weights ($\mathbf{D}_n$), and the matrix of the metric used to measure the distances between rows ($\mathbf{D}_p$). $\mathbf{D}_p$ can also be seen as the matrix of column weights, and $\mathbf{D}_n$ as the matrix of the metric used to measure the distances between columns. The analysis of this statistical triplet is based on the eigenvector analysis of matrix $\mathbf{X}^t\mathbf{D}_n\mathbf{X}\mathbf{D}_p$. If $\mathbf{D}_p$ is not proportional to the identity, this matrix is not symmetric but the eigenequation can be written as $\mathbf{D}_p^{1/2}\mathbf{X}^t\mathbf{D}_n\mathbf{X}\mathbf{D}_p^{1/2}\mathbf{u}_\alpha = \lambda_\alpha\mathbf{u}_\alpha$, and solved without problem. Eigenvectors $\mathbf{u}_\alpha$ and $\mathbf{u}_\beta$ verify ${\mathbf{u}_\alpha}^t\mathbf{u}_\beta = \delta_{\alpha\beta}$, and the principal axes and row scores in the analysis of the triplet are respectively $\mathbf{a}_\alpha = \mathbf{D}_p^{-1/2}\mathbf{u}_\alpha$ and $\mathbf{c}_\alpha = \frac{1}{\sqrt{\lambda_\alpha}}\mathbf{X}\mathbf{D}_p^{1/2}\mathbf{u}_\alpha$.

This model allows a standardisation of the formal description of all the methods of multivariate analysis based on eigenvector analysis (e.g., principal components analysis, correspondence analysis, multiple correspondence analysis, discriminant analysis, canonical correlation analysis, canonical correspondence analysis, analyses on intrumental variables, etc.)

ADE is written in ANSI-C and is available for Apple Computer Macintosh microcomputers. Like the other graphical programs of the ADE package, ADEScatters offers the possibility to build automatically collections of graphics. These collections can be made by the columns of the input table, or by groups of rows selected by the user. ADEScatters presents a "Windows" menu that can be used to choose one of the three parameter windows that allows an interactive definition of the graphic parameters (Figure 1). The user can thus freely modify the values of all the graphic parameters and the resulting picture is displayed in the "Graphics" window. The main dialog window is used to choose the input files and related parameters.

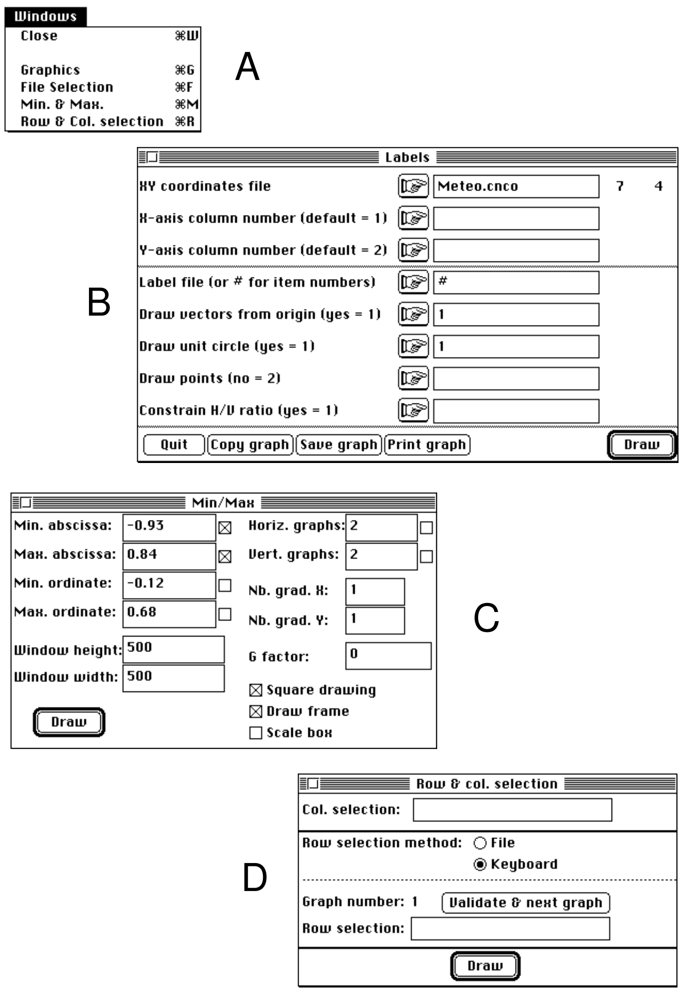

Figure 1: *Screen shots of ADEScatters user interface. The windows menu (A) allows the user to choose among the three parameter setting windows: the main dialog window (B), the minimum and maximum window (C), and the row and column selection window (D).*

In the "Min/Max" window, the user can set the values of the minimum and maximum of abscissae and ordinates, the number of horizontal and vertical graphics (in the case of collections of graphics), the graphical window width and height, legend scale options, and other details. In the "Row & Col. selection" window, he can choose the columns and the groups of rows of the table that will be used to make each elementary graphic of a collection.

The first special feature of ADEScatters is the “searchable factor map”; this means that the user can search for an element on the factor map. Two commands are available for this purpose in the Edit menu: Find and Find label. The Find option allows the user to enter an element number and displays the corresponding label on the factor map with blinking bold characters surrounded by a rectangular frame (Figure 2).

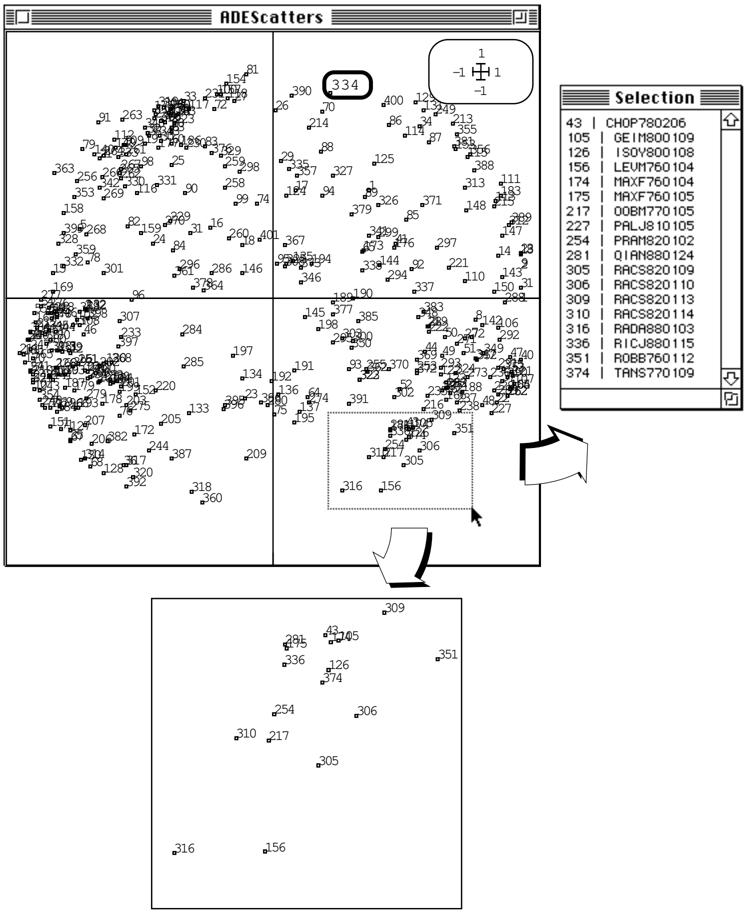

Figure 2: *Screen shots of ADEScatters showing how a searched element is displayed (element 334 , top of the window), and zooming or selection capabilities. Selecting a zone on the factor map zooms on this zone or can be used to obtain the list of elements inside the selection.*

The Find label option does the same thing, but the user enters the character string of the label instead of the element number (this spares the user the trouble of finding the rank of the element he is looking for). This feature is really useful only when there are many elements on the factor map, leading to a lot of superimposed points.

Another interactive characteristic of ADEScatters is zooming: using combinations of keystrokes and mouse clicks, the user can either zoom in a particular region of the factor map (by selecting this region with the mouse, see Figure 2, lower part), or globally enhance or reduce the scale of the whole map. This allows the user to focus his attention on a part of the graphic, and to analyse precisely the position of neighbouring points on the factor map. He can also obtain the list of the elements that belong to the selected region (Figure 2, right part).

The most important feature of ADEScatters is the possibility to represent the values of the data table on the factor map. For example, if the current graphic on the screen is a factor map of rows, then when the user clicks on a point the values of the corresponding row in the data table are automatically drawn in a small pop-up window. This window shows a bar chart, with the height of the bars proportional to the data (Figure 3).

In fact, the values represented are not the raw data, but the values transformed according to the multivariate analysis that was performed (principal component analysis, correspondence analysis, CA or multiple correspondence analysis, MCA). For PCA, these are the centered or standardized values, for CA the relative frequencies, and for MCA the centered and weighted disjunctive table associated to the qualitative variables.Thus, all the values are at the same scale. They are sorted according to the order of the columns of the data table, but it is also possible to sort them according to their factor score (the user must then choose which factor is to be used for sorting). This second possibility is very useful to show which columns mostly contribute to define the corresponding factor.

With a different combination of keystrokes and mouse click, it is possible to represent the factor map of the columns (instead of the bar chart), with circles and squares (Figure 3). The size of circles and squares is proportional to the data (with squares for negative values and circles for positive ones) and the proportionality coefficient can be changed. The factors for this map are the same as those of the main display (for example F1 and F2 in both cases).

The principle of this representation is analogous to the biplot representation (Gabriel 1971), but it is restricted to the representation of only one element, which prevents cluttering of the factor map. Indeed, the superimposition of all the rows and columns of the table on the factor map makes it unreadable for large data sets. The user has access to the number of the column corresponding to each circle or square by just moving the mouse cursor to its center: this number appears and stays displayed as long as the mouse is kept at the same place (see Figure 3).

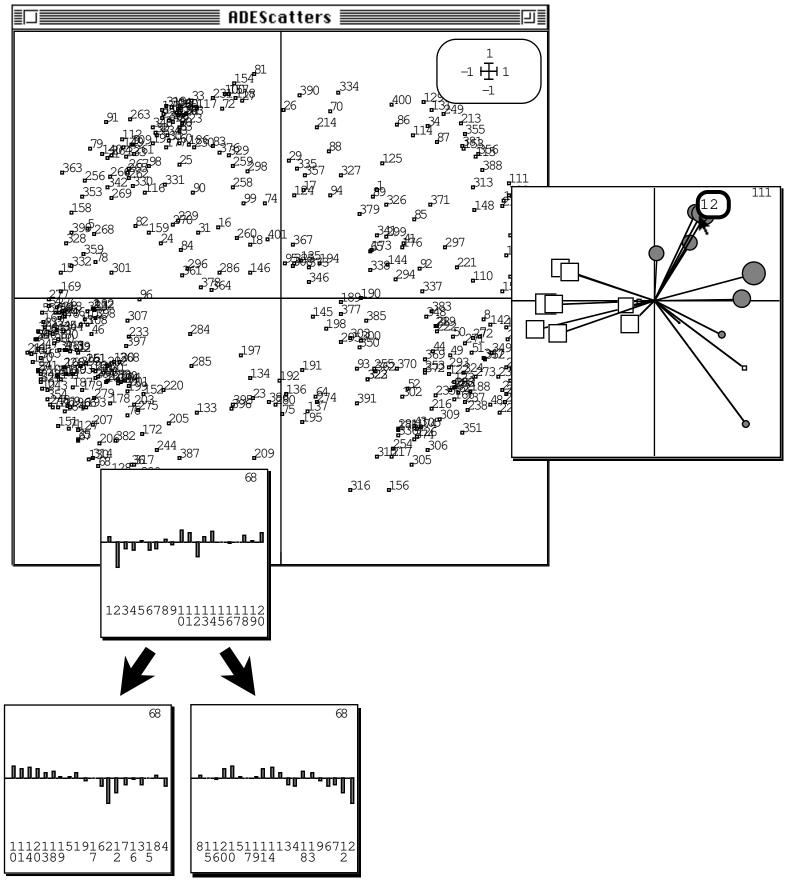

Figure 3: *Screen shots showing how data values can be displayed on the factor map. Lower-left: the (standardized) values of element 68 are displayed in a bar chart graphic; values can be sorted according to the first (left) or second (right) factor score. Upper-right: the values of element 111 are projected on the dual graphic with circles and squares (see text). When the user brings the mouse cursor on an element of this new graphic, the number of this element is displayed (number 12 here).*

Obviously, if the factor map on the screen is the map of columns, then the values drawn in the pop-up windows are the values of the column on which the user clicks. For the bar chart representation, these values can be sorted according to the factor scores of the rows, and for the biplot display, it is the factor map of the rows that is presented. The graphics that appear in the pop-up windows are automatically copied to the computer clipboard. The user can then paste them in a drawing software to keep track of the exploration process.

3 Examples of use

Figure 2 and 3 come from a molecular biology example by (Thioulouse & Lobry 1995). The data table (20 rows and 402 columns) contains the values of 402 physico-chemical and biological variables measured on the 20 natural amino-acids occurring in proteins (Nakai, Kidera & Kanehisa 1988). A PCA was performed on the standardized variables. ADEScatters was used to investigate the factor maps, and particularly the map of variables (figures show F1 x F2 maps). Zooming and mouse selection of regions of the map (Figure 2) helped us to analyse the first three principal components, totaling 65% of inertia. The first component was interpreted as a hydrophobicity scale, the second one was the amino-acid propensity to favour alpha helix or beta sheet structures in proteins, and the third one the weight and aromaticity of amino-acids.

Figure 3 gives an example of the possibility to have access to the values of the data set directly on the factor map. On the lower left part of the figure, the user has clicked on the point corresponding to variable 68 (the consensus normalized hydrophobicity scale, (Eisenberg 1984)). The resulting bar chart shows the values of this variable for the 20 amino-acids, ordered alphabetically as in the data table. By pressing simultaneously the option key, the 20 amino-acids can be sorted according to their factor score (the user is asked to enter the factor to be used). After sorting on the first or second factor (lower part), the user can see which amino-acids contribute to the definition of these factors. The right part of the figure shows what happens when the user also presses the shift key while clicking on an item: the dual factor map (i.e. that of the 20 amino-acids) is drawn in a small pop-up window, with circles and squares proportional to the values of variable 111 (amino-acid polarity, (Grantham 1974)). This variable appears to be closely linked to the first factor, since all the squares (except one) are in the left part of the map, while circles are in the right part. The user has brought the cursor on one of the circles, and the number of the amino-acid is displayed (number 12, which corresponds to Lysine).

4 Discussion and conclusion

Representing the values of a data set on the factor map has already been proposed as an aid to the interpretation of factors (see Thioulouse et al. 1991 for a review in the field of ecological data analysis). The method proposed here allows us to solve the problem of cluttering the factor map when the data table is large, and has the advantage of interactivity: the user can focus his attention on the elements his is interested in. The same technique could be used with three-dimensional factor representations, but the selection of objects in a three-dimensional space through the two-dimensional display of the computer screen is somewhat difficult, and it should be coupled with rotation methods. Other means of data representation could be investigated, for example representing the mean (or variance) of the values computed on groups of elements selected on the factor map. Brushing techniques on the scatterplot matrix of factor scores combined to data representation could also bring new insights into the structures of data tables.

References

Asimov, D. (1985), 'The grand tour: A tool for viewing multidimensional data', *SIAM Journal of Scientific and Statistical Computations* **6**(1), 128–143.

Becker, R.A. & Cleveland, W.S. (1987), 'Brushing scatter plots', *Technometrics* **29**, 127–142.

Cailliez, F. & Pages, J.P. (1976), 'Introduction à l'analyse des données'. SMASH, Paris.

Chernoff, H. (1973), 'Using faces to represent points in k-dimensional space graphically', *Journal of the American Statistical Association* **68**(1), 361–368.

Cleveland, W.S. & McGill, M.E. (1988), 'Dynamic graphics for statistics', Wadsworth & Brooks/Cole, Belmont.

Donoho, A.W., Donoho, D.L. & Gasko, M. (1986), 'MACSPIN: A tool for dynamic display of multivariate data', Wadsworth & Brooks/Cole, Monterey.

Escoufier, Y. (1987), 'The duality diagramm : a means of better practical applications', *in* Legendre, P. & Legendre, L. (eds.), 'Development in numerical ecology', NATO advanced Institute, Serie G, Springer Verlag, Berlin. pp. 139-156.

Eisenberg, D. (1984), 'Three dimensional structure of membrane and surface proteins', *Annual Review of Biochemistry* **53**, 595–623.

Gabriel, K.R. (1971), 'The biplot-graphic display of matrices with application to principal component analysis', *Biometrika* **58**, 453–467.

Gower,J.C. & Digby, P.G.N. (1981), 'Expressing complex relationships in two dimensions',*in* Barnett, V. (ed.), 'Interpreting multivariate data', John Wiley & Sons, New York.

Grantham, R. (1974), 'Amino-acid difference formula to help explain protein evolution', *Science* **185**(1), 862–864.

Huber, P.J. (1985), 'Projection pursuit', *Annals of Statistics* **13**(2), 435–475.

Hurley, C. & Buja, A. (1990), 'Analyzing high-dimensional data with motion graphics', *SIAM Journal of Scientific and Statistical Computing* **11**, 1193–1211.

Nakai, K., Kidera, A. & Kanehisa, M. (1988), 'Cluster analysis of amino acid indices for prediction of protein structure and function', *Protein Engineering* **2**, 93–100.

Thioulouse, J., Devillers, J., Chessel, D. & Auda, Y. (1991), 'Graphical techniques for multidimensional data analysis', *in* Devillers, J. & Karcher, W. (eds.), 'Applied multivariate analysis in SAR and Environmental Studies', Kluwer, Dordrecht, pp. 153-205.

Thioulouse, J., Dolédec, S., Chessel, D. & Olivier, J.M. (1995), 'ADE software: multivariate analysis and graphical display of environmental data', *in* Guariso, G. & Rizzoli, A. (Eds), 'Software per l'ambiente', Pàtron editore, Bologne, pp. 57-62

Thioulouse, J. & Lobry, J.R. (1995), 'Co-inertia analysis of amino-acid physico-chemical properties and protein composition with the ADE package', *Computer Applications in the Biosciences* **11**(3), 321–329.

Weihs, C. & Schmidli, H. (1990), 'OMEGA (Online Multivariate Exploratory Graphical Analysis): routine searching for structure', *Statistical Science* **2**(3), 175–208.

Young, F.W., Faldowski, R.A. & McFarlane, M.M. (1993), 'Multivariate Statistical Visualization', *in* Rao, C.R. (ed.), 'Handbook of statistics', Vol. 9, Elsevier, Amsterdam.

Young, F.W., Kent, D.P. & Kuhfeld, W.F. (1988), 'Dynamic graphics for exploring multivariate data', *in* Cleveland, W.S. & McGill, M.E. (eds.), 'Dynamic graphics for statistics', Wadsworth, Belmont. pp. 391-424.