Exploring Data Through Archetypes

Chapter · May 2010

DOI: 10.1007/978-3-642-10745-0_31

CITATIONS
4

READS
329

3 authors:

maria rosaria d'esposito
Suor Orsola Benincasa University of Naples
25 PUBLICATIONS **159** CITATIONS
SEE PROFILE

Giancarlo Ragozini
University of Naples Federico II
87 PUBLICATIONS **620** CITATIONS
SEE PROFILE

Domenico Vistocco
University of Naples Federico II
68 PUBLICATIONS **507** CITATIONS
SEE PROFILE

Exploring Data Through Archetypes

Maria Rosaria D'Esposito, Giancarlo Ragozini, and Domenico Vistocco

Abstract In this paper we propose a mixed analytical and graphical exploratory strategy based on data archetypes for the exploratory analysis of multivariate data. Our approach is of considerable help in exploring the periphery of the data scatter, exploiting an outward-inward perspective, to highlight small peripheral groups as well as anomalies, outliers and irregularities in the data cloud shape. The strategy is carried out in a comprehensive quantitative programming environment provided by the joint use of the software system R and of the visualization system GGobi. It provides a visualization system involving both static and dynamic graphics based on the so-called multiple views paradigm. The views are organized in a spreadplot and heavily exploit dynamics and interactive statistical graphics.

1 Introduction

Exploratory data analysis (EDA) is, in the words of Tukey (1977) p. 21, "a detective work, finding and revealing the clues", i.e. uncovering unanticipated structures in the data. EDA uses numerical as well as visual and graphical techniques to accomplish its aims. Graphical and visual tools (such as stars, glyphs, parallel coordinates) become particularly necessary for the exploration of multivariate data as they make it possible to visualize multidimensional data in 2D (for a review see, among others, Chambers et al. 1983; Wegman and Carr 1993).

However, visualization systems that focus on the graphical representation of information can run into several problems. Mainly, there can be loss of valuable information – when too much data are visualized on the screen – and needs to organize discoveries off line (Yang et al. 2007). Interactive and dynamic statistical graphics which incorporate motion and user interaction with graphical display – such as brushing and slicing, coloring and rendering, empirical and algebraic linking

G. Ragozini (✉)
Department of Sociology, Federico II University of Naples, Italy
e-mail: giragoz@unina.it

H. Locarek-Junge and C. Weihs (eds.), *Classification as a Tool for Research*,
Studies in Classification, Data Analysis, and Knowledge Organization,
DOI 10.1007/978-3-642-10745-0_31,

(Young et al. 1993) – can become part of the visualization techniques to enhance the results of the analysis and to overcome some of the problems of the visualization systems. A major advance can also be obtained by introducing integrated analytical devices into the visualization systems, which could aid users in the knowledge discovery task.

In this paper we propose a mixed analytical and graphical exploratory strategy based on data archetypes for the exploratory analysis of multivariate data. Archetypes are few pure types given by weighted average of the data. They are useful in summarizing the data, but, in our opinion, may also be profitably employed to explore the periphery of the data scatter.

It is known that small peripheral groups, anomalies, outliers and irregularities in the data cloud shape in higher dimensions can easily hide in marginal projection or in usual graphical representations. On the contrary, by adopting an outward-inward perspective, i.e. by analyzing the archetypes' surroundings, all the previous data structures can be more easily detected.

With this purpose, we propose an integrated strategy: first, based on the aims of the analysis, extract the archetypes from the data, then represent the data in the space spanned by the archetypes adopting dynamic and interactive visualization tools. The strategy, carried out in the comprehensive quantitative programming environment provided by the joint use of R (R Development Core Team 2008) and GGobi (Cook and Swayne 2007; Lang et al. 2008), will result in a visualization involving both static and dynamic graphics based on the so-called multiple views paradigm (Wilhelm 2005), which allows interaction through dynamic graphics and provides multiple different and simultaneous plots of the same data. The views will be organized in a spreadplot, a spreadsheet-like arrangement of linked, dynamic, interactive plots (Young et al. 1992, 1993).

The paper is organized as follows: Sects. 2 and 3, respectively, present the basic elements of archetypal analysis and spreadplot design; the proposed procedure is illustrated in Sect. 4, both theoretically and practically, exploiting a real dataset; concluding remarks are in Sect. 5.

2 Elements of Archetypal Analysis

Archetypal analysis is a quite recent statistical method for multivariate data analysis (Cutler and Breiman 1994). It aims at finding archetypes that represent a sort of "pure individual types", i.e. few points lying on the boundary of the data scatter that are intended as a synthesis of the observed points. At the same time, as they are not necessarily observed points, they represent ideal objects on which the observed data may be patterned.

Formally, the archetypes $\mathbf{a}'_j$ are a convex combination of the observed data:

$$\mathbf{a}'_j = \boldsymbol{\beta}'_j \mathbf{X} \tag{1}$$

where $\mathbf{X}$ is the observed data matrix, $\boldsymbol{\beta}_{ji} \geq 0 \quad \forall j, i$ and $\boldsymbol{\beta}'_j \mathbf{1} = 1 \quad \forall j$.

On the other hand, all the data points can be expressed in terms of the archetypes:

$$\mathbf{x}_i' = \gamma_i' \mathbf{A} \tag{2}$$

with $\gamma_{ij} \geq 0 \ \ \forall i, j$ and $\gamma_i' \mathbf{1} = 1 \ \ \forall i$. In (2) $\mathbf{A}$ is the archetype matrix and γ_i' are weights of the archetypes for each data point.

Equation (1) and the related constraints on β's coefficients imply that archetypes belong to the convex hull boundary of the data, while (2) and the related constraints on γ's coefficients imply that all the data belong to the convex hull boundary of the archetypes. Hence, the archetypes must coincide with the v vertices of the data convex hull to fulfill the previous conditions (Porzio et al. 2008).

However, in practice, the number of the data convex hull vertices is generally too large to properly synthesize the data. For this reason, looking for a smaller number of pure types, and wishing to preserve their closeness to the data, Cutler and Breiman (1994) defined the archetypes as those m, with $m \leq v$, points that fulfill (2) as far as possible, satisfying all the other conditions. Hence, given m, the archetypes can be defined as the points $\mathbf{A}(m) = (\mathbf{a}_1, \ldots, \mathbf{a}_m)$ minimizing the distances between the observed data points $\mathbf{x}_i'$ and the reconstructed points $\tilde{\mathbf{x}}_i'(m)$, with $\tilde{\mathbf{x}}_i'(m) = \gamma_i'(m) \cdot \mathbf{A}(m)$.

Formally, the archetypes are those points that minimize the quantity:

$$RSS(m) = \left\| \mathbf{X} - \tilde{\mathbf{X}}(m) \right\|_F = \left\| \mathbf{X} - \boldsymbol{\Gamma}(m)\mathbf{A}(m) \right\|_F = \left\| \mathbf{X} - \boldsymbol{\Gamma}(m)\mathbf{B}'(m)\mathbf{X} \right\|_F \tag{3}$$

holding all the other conditions, and where $\|\mathbf{Y}\|_F = \sqrt{Tr\,(\mathbf{Y}\mathbf{Y}')}$ is the Frobenius norm for a generic matrix $\mathbf{Y}$, with $\mathbf{B}(m) = [\beta_{ji}]$ and $\boldsymbol{\Gamma}(m) = [\gamma_{ij}]$. The solution to this minimization equation depends on m, and solutions are not nested as m varies. That is, the archetypal points that solve (3) for $m = m^*$ are not necessarily a subset of the points that solve (3) for $m = m^* + 1$. For this reason, we denote with $\mathbf{a}_j'(m)$ the j-th archetype for a given m, and we will generally have $\mathbf{a}_j'(m) \neq \mathbf{a}_j'(l)$, for $m \neq l$. Theoretically, the $RSS(m)$ in (3) is a decreasing function of m that has the maximum for $m = 1$ and goes to zero for m approaching the number of the convex hull vertices. For a given m, it highlights the synthesizing power of the archetypes since it shows how well archetypes reconstruct data.

Figure 1 exhibits a set of 50 simulated data points with seven convex hull vertices. For this dataset, the $RSS(m)$ function shows that three archetypes are sufficient to synthesize the data. Indeed, for $m = 3$ the $RSS(m)$ is close to zero as the majority of the data belongs to the archetypes' convex hull (the triangle highlighted in Fig. 1), and it is well reconstructed by the archetypes.

Up to now archetypal analysis has been applied in many fields. In the field of physics, it has been used to detect clusters of cellular flames (Stone and Cutler 1996; Stone 2002) and of galaxy spectra (Chan et al. 2003). It has found application as a tool for image decomposition (Marinetti et al. 2006, 2007), where archetypal analysis seems to provide results which are easily interpretable in terms of physical meaning.

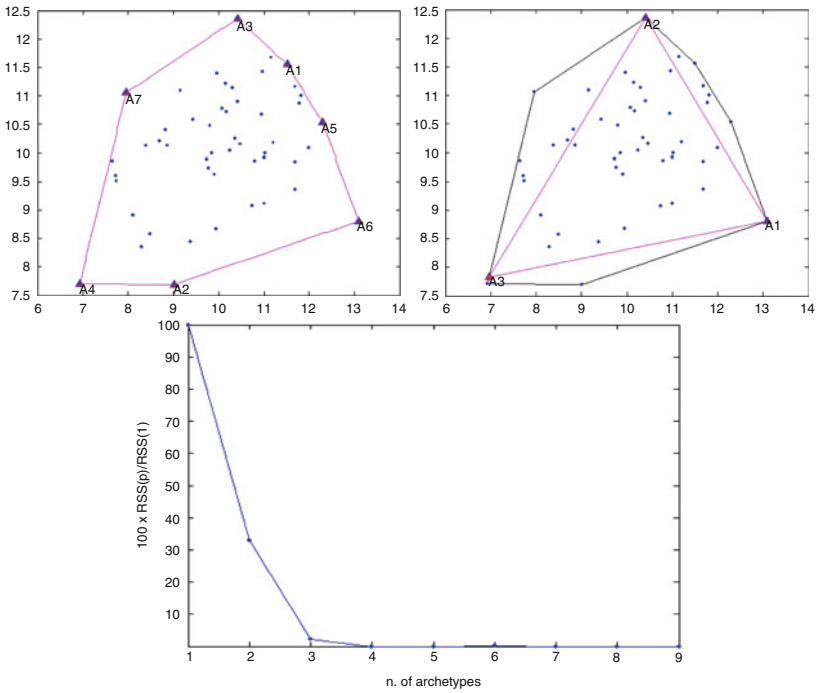

Fig. 1 A set of 50 simulated data points with seven convex hull vertices. *Clockwise*: the dataset with the convex hull boundary and $m = 7$ archetypes highlighted; the dataset with three archetypes highlighted, data convex hull boundary and archetype convex hull boundary; the $RSS(m)$ function

Marketing research is also a relevant field of applications. The idea of archetypes has been associated to the idea of archetypal consumers; they have been exploited for market segmentation and consumer fuzzy clustering (Elder and Pinnel 2003; Li et al. 2003). In the same field, the extension of archetypal analysis to interval coded data has been recently proposed (D'Esposito et al. 2006).

In performance analysis, archetypes have been exploited to construct data driven benchmarks (Porzio et al. 2008), to analyze CPU performance (Heavlin 2007), and to obtain a multivariate ordering procedure based on the idea of the *"worst-best" direction* selected through the archetypes (D'Esposito and Ragozini 2008).

In all the above mentioned applications the archetypes have been mainly adopted as summarizing observations. In this paper we propose to use them as a tool to analyze the structure of the data in a genuine exploratory fashion by merging the analytics of archetypes with dynamic and interactive visualization tools according to a spreadplot scheme (Young et al. 1992).

3 Elements of Spreadplot Design

Enhanced man-machine interaction makes it possible to design tools for an interactive visual exploration of data, and for visually querying the data. Strictly related to interactivity is the paradigm of the linked views (Wilhelm 2005). It consists of linking, empirically or algebraically, several views of the same dataset and in propagating information through different plots that display different aspects (dimensions) of the same dataset (Stuetzle 1987; Young et al. 1993). Brushing, slicing and coloring allow one to visually explore data and to investigate if a particular pattern or position is confirmed on different views of the same dataset.

In this framework, different data views can be arranged as a spreadplot (Young et al. 1992), a spreadsheet-like arrangement of dynamic and interactive plots empirically and algebraically linked together by equations. Each individual view can be either dynamic or static, and can support high-interaction direct manipulation. Moreover, different views can be linked and when the user changes the information shown in one view, the changes can be processed by a set of equations and instantly represented in the others. The user can interactively modify analysis parameter by acting on user interface tools of the spreadplot, such as moving points, sliding cursors.

In this paper, we propose an exploratory strategy based on the archetypes and visually translated into a spreadplot. This includes a set of ad hoc *R* routines implementing the method and is based on open source software whose core is the *R* software system integrated with the visualization system GGobi (Buja et al. 2003) through the *rggobi* package (Lang et al. 2008). We choose the GGobi system since it contains several dynamic and interactive graphics such as tourplot, scatterplot, bar chart and parallel coordinate plot. The whole software architecture exploits the Model-View-Controller (MVC) design pattern and the Observer design pattern (Buschmann et al. 1996; Gamma et al. 1995).

4 The Proposed Exploratory Data Analysis Strategy

It is well known that for high dimensions data structures anomalies and outliers, irregularities in the data cloud shape, and small peripheral groups can easily hide in marginal projections or in usual graphical representations. At the same time, archetypes which are located on the boundary of the data convex hull, can provide an outward-inward point of view on the data scatter that will allow to explore the data cloud peripheries and highlight many data patterns more easily.

The strategy we propose for the exploration of a multivariate dataset is based on the outward-inward perspective given by the archetypes, combined with the geometric properties of the γ coefficients in (2), and the dynamic and interactive visualization tools conveyed in a spreadplot.

The proposed exploratory strategy consists of the following steps:

- Derive the archetypes by increasing their number m one unit at time and look at the $RSS(m)$ function in order to understand the synthesizing power of each additional archetype;
- For each m analyze the archetypes in the data space through some graphical representation (e.g. percentile profile plots, star plot, parallel coordinate plot) to interpret and compare the archetypes;
- On the basis of the previous steps, choose the first interesting m and for this:
  - Represent, through a parallel coordinate plot, the data within the space spanned by the archetypes according to the γ coefficients derived from (2);
  - Use interactive and dynamic tools like brushing, coloring and linking to highlight peripheries selecting the data with γ_{ij} coefficients close to 1 on each archetype, i.e. select the outer data looking for gaps and peripheral structures and for isolated data points;
- Iterate by increasing m, i.e. selecting another subsequent interesting m;
- Stop when increasing m does not provide additional information.

The previous steps are detailed in the following subsections using a real dataset adopted as an illustrative example. The data refer to a study on the performance of central processing units (CPUs) and consist of a set of 209 CPUs to be compared by considering seven performance indicators (Cycle time – ns, Minimum memory – kb, Maximum memory – kb, Cache size – kb, Minimum channels, Maximum channels, Relative performance); low values of the indicators stand for poor performances (Ein-Dor and Feldmesser 1987). Apart from the usual goals (discovering patterns, groups, outliers, ...), the analysis will aim also at finding CPUs with good or bad performance in terms of the seven indicators, and in comparing all the others with them.

4.1 Deriving and Analyzing Archetypes by Varying m

The first step consists of looking at the $RSS(m)$ function defined in (3) and deriving the archetypes as m increases. When the aim is data synthesis, it is necessary to choose one appropriate m. This will corresponds to the one that does not yield a significant decrease in the $RSS(m)$. If the aim is to explore data, we suggest to look at archetypes for different values of m, evaluating also their interpretability by visualizing them through percentile profile plot, stars or other graphical representations. In particular, the percentile profile plot displays for each archetype a sequence of vertical bars (one for each variable) with heights equal to the value of the cumulative empirical distribution evaluated at the archetypal point. It visualizes the archetype's relative standing with respect to the others points.

Figures 2 and 3 show two spreadplots, respectively for $m = 4$ and $m = 6$, portraying the $RSS(m)$ function, percentile profile plots/star plots along with the

parallel coordinate plot for the computer data. Note that in the spreadplots the sliding cursor makes it possible to change interactively the number of archetypes in order to see how percentile profiles and stars change. The values of $m = 4$ and $m = 6$ have been chosen because, by looking at the $RSS(m)$ function in Fig. 3, it appears that $RSS(m)$ decreases sharply up to $m = 6$ with a quite flat behavior between $m = 2$ and $m = 4$.

An inspection of the percentile profile plot in Fig. 2 highlights that archetypes $a_2(4)$ and $a_3(4)$ correspond to bad performances except for the first indicator in $a_3(4)$, as all the percentile bars are low in values, i.e. the archetypes are close to the low values of the indicators. On the other hand $a_4(4)$ represents the best CPUs and $a_1(4)$ the average CPUs. The coordinate parallel plot shows that the majority of CPU data are close together with very few of them having better performances on different indicators. The parallel coordinate plot also highlights skewed marginal distributions.

With $m = 6$ (Fig. 3) we note that one archetype represents the best CPUs ($a_3(6)$), and two archetypes represent the worst CPUs ($a_4(6)$ and $a_5(6)$). The remaining three archetypes represent CPUs with high values on only some indicators. Even if $RSS(6)$ is much lower than $RSS(4)$, the two additional archetypes give no further information and there does not appear to be much gain in interpretability.

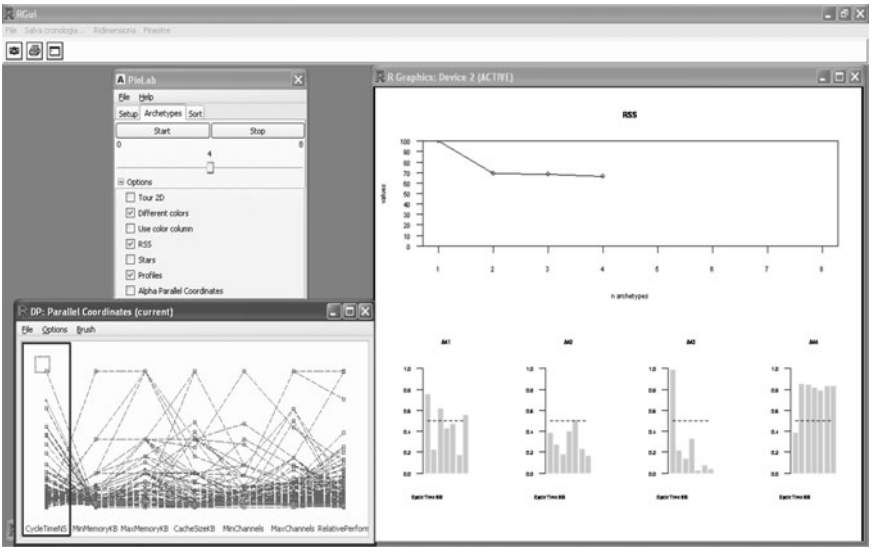

Fig. 2 A spreadplot for the CPUs dataset for $m = 4$ archetypes. *Clockwise*: the control panel which clearly shows the sliding cursor for interactively choosing the number of archetypes, and the list of possible graphical representations; the $RSS(m)$ function and the percentile profile plots for the chosen four archetypes; the parallel coordinate plot of the CPUs data along with the archetypes

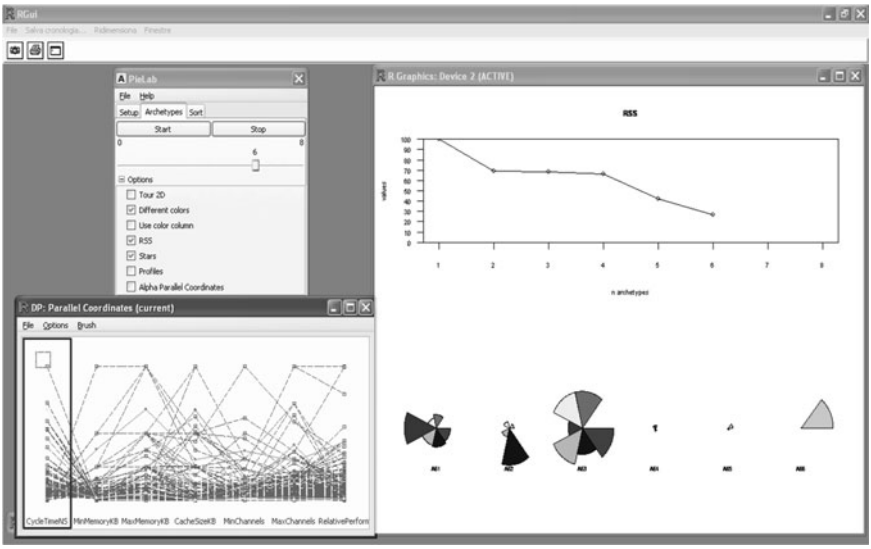

Fig. 3 A spreadplot for the CPUs dataset for $m = 6$ archetypes. *Clockwise*: the control panel which clearly shows the sliding cursor for interactively choosing the number of archetypes, and the list of possible graphical representations; the $RSS(m)$ function and the star plots for the chosen six archetypes; the parallel coordinate plot of the CPUs data along with the archetypes

4.2 Representing Data in the Spaces Spanned by the Archetypes

The next step of the proposed procedure relies on the representation of the data in the m-dimensional spaces spanned by the archetypes.

The archetypes are vertices of a simplex in the data space $\Re^p$, and for each data point $\mathbf{x}'_i$ new coordinates with respect to the archetypes can be obtained by solving the equation $(\lambda_{i1} + \cdots + \lambda_{im})\,\mathbf{x}'_i = \lambda_{i1}\,\mathbf{a}'_1 + \cdots + \lambda_{im}\,\mathbf{a}'_m$.

The coefficients $(\lambda_{i1}, \ldots, \lambda_{im})$ are the new coordinates of $\mathbf{x}'_i$ in an associated space, and they are called barycentric coordinates (see e.g. Coxeter 1969) with respect to the archetypes. The archetypes themselves have barycentric coordinates $(1, 0, \ldots, 0), (0, 1, \ldots, 0), \ldots, (0, 0, \ldots, 1)$, as they are the associated space basis. We note that, from the geometric properties of the barycentric coordinates, in the space spanned by the archetypes the data points actually belong to an $(m - 1)$ dimensional subspace of this associated space.

The reconstructed data points $\tilde{\mathbf{x}}'_i$ have barycentric coordinates in the archetype associated space as well. In particular, the equation $(\lambda_{i1} + \cdots + \lambda_{im})\,\tilde{\mathbf{x}}'_i = \lambda_{i1}\,\mathbf{a}'_1 + \cdots + \lambda_{im}\,\mathbf{a}'_m$, is exactly solved for $\lambda_{ij} = \gamma_{ij}$, $j = 1, \ldots, m$. Hence, it turns out that the $\gamma_{ij}(m)$ coefficients are the barycentric coordinates for the reconstructed points in the associated space (see Porzio et al. 2008 for details).

Consequently, they may be exploited to map the original data into a lower dimensional space. Given the relationship between each $\mathbf{x}'_i$ and its corresponding $\tilde{\mathbf{x}}'_i$, each

original point will be represented by the coefficients γ_{ij} into the archetype associates space. Note that the same dataset may be mapped into many archetype spaces, one for each value of m. Analyzing data in these associated spaces provides further insights into the data structure from a different perspective.

4.3 Exploring the Peripheries of the Data Scatter

As previously stated, when the aim is to explore the data, it can be useful to look at the archetypes for different values of m. Indeed, it is worth noticing that looking at a set of values of m may be necessary – even if the $RSS(m)$ indicates that a small m is sufficient to well reconstruct the data – to search for outliers and peripheral groups. In such cases, it could happen that outliers will coincide with some archetypes not corresponding to a sharp decrease in the $RSS(m)$. Indeed, the first archetypes will catch the majority of data, while additional archetypes will point to outliers or to small peripheral groups, if any.

The exploration task can be pursued exploiting static and dynamic graphical representation, such as parallel coordinates and tourplot, in the associated space provided by the barycentric coordinates. In the parallel coordinate plot of such associated spaces, multivariate skewness could be highlighted as marginal asymmetry along some directions. Moreover interaction tools – brushing, slicing and coloring – can offer the user a thorough exploration of the data periphery. Brushing and coloring data around the archetypes in the parallel coordinate plot can highlight gaps in the data structures, small groups and outliers. To enhance the detection of interesting patterns the parallel coordinate plot can also be empirically linked to the corresponding data in the tourplot graph.

For example, in Fig. 4, which correspond to $m = 4$ archetypes, the parallel coordinate plot depicts all the data in a 4-D space where all the coordinates are the γ coefficients as previously stated (for the geometric properties of γ coefficients it is actually a 3-D space). The plot shows that, for three out of the four axes, some points are isolated even if not very far from the others, while for the $a_3(4)$ axis there are three data groups. By pointing to the observations close to $a_3(4)$ – i.e. those points with coordinate values close to the pattern (0, 0, 1, 0) – we highlight three data points in the tourplot along the Cycle time dimension. Further investigations can be pursued to analyze the other groups appearing on the same archetype-axis. Some skewness and clusters appear by looking along the directions of the first two archetypes.

While in Sect. 4.1 we observed that going from $m = 4$ to $m = 6$ there was not much gain in interpretability and synthesizing power, when exploring the peripheries of the data scatter, the analysis is enhanced at $m = 6$ archetypes. In fact in Fig. 5, more peripheral groups appear: in four out of six dimensions small peripheral groups can be detected. For example, by pointing to the observations close to $a_3(6)$ – i.e. those points with coordinate values close to the pattern (0, 0, 1, 0, 0, 0) – we highlight in the tourplot twelve isolated data points along the Cycle time

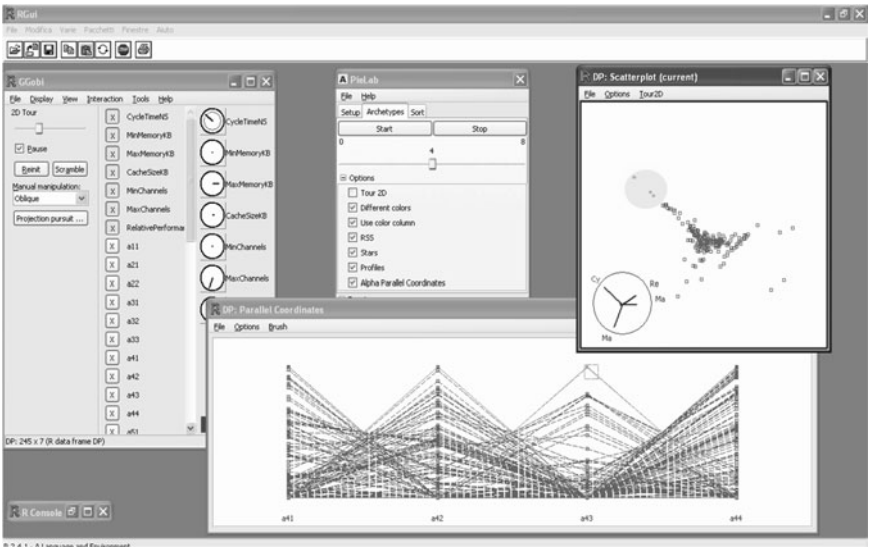

Fig. 4 The strategy in action for $m = 4$ archetypes: acting through brushing and coloring in the parallel coordinate plot in the γ space involves the highlighting of the corresponding points in the tourplot

dimension. Remember that with $m = 4$ archetypes we succeeded in detecting only three isolated data points. Furthermore, the first three archetypes and the last one show severe skewness. It is worth noticing that the tourplot view in Fig. 5 spotting the isolated group is not straightforward. In our case, with the aid of data representation in the space spanned by the archetypes and of interactive graphics, it was easier for us to capture this particular projection.

5 Concluding Remarks

The approach we have used in this paper to explore multidimensional data scatter seems promising in "finding and revealing the clues", mainly in uncovering gaps in the data structures, small groups, outlying values, asymmetries and irregularities in the shape.

The whole procedure, being based on open source softwares, can be easily replicated and perhaps enhanced. Moreover it is not computationally intensive and it does not require huge hardware capabilities.

Finally, note that for the sake of presentation we split the procedure into several spreadplots. However, all the different views can be merged in a unique spreadplot and inspected at the same time.

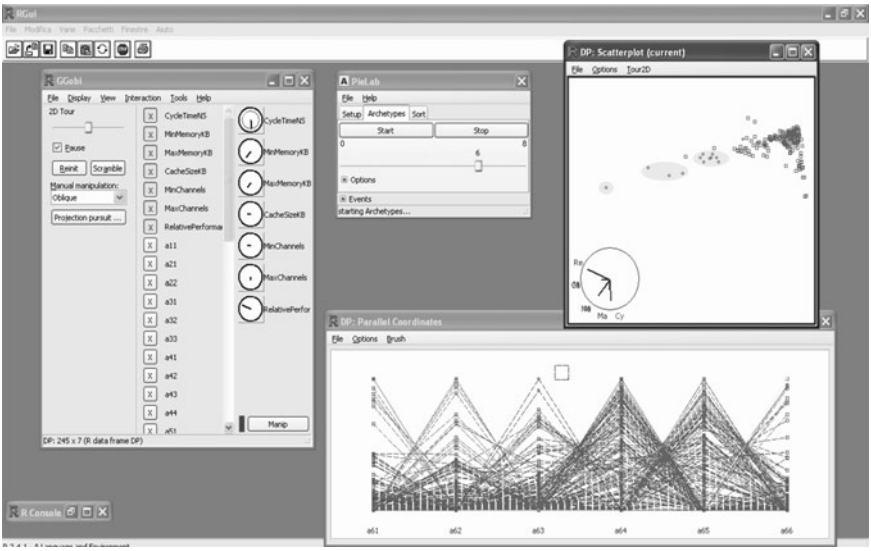

Fig. 5 The strategy in action for $m = 6$ archetypes: acting through brushing on the parallel coordinate plot in the γ space involves the highlighting of the corresponding points in the tourplot

Acknowledgements The authors wish to thank Michele Risi from University of Salerno for the design and implementation of the software architecture. The research work of Maria Rosaria D'Esposito benefits from the research structures of the STATLAB at the Department of Economics and Statistics, University of Salerno. The research work of Domenico Vistocco is supported by Laboratorio di Calcolo ed Analisi Quantitative, Department of Economics, University of Cassino.

References

Buja, A., Lang, D. T., & Swayne, D. F. (2003). GGobi: Evolving from XGobi into an extensible framework for interactive data visualization. *Journal of Computational Statistics and Data Analysis, 43*, 423–444.

Buschmann, F., Meunier, R., Rohnert, H., Sommerlad, P., & Stal, M. (1996). *A system of patterns: Pattern-oriented software architecture*. West Sussex, England: Wiley.

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphical methods for data analysis*. Wadsworth, Monterey CA.

Chan, B. H. P., Mitchell, D. A., & Cram, L. E. (2003). Archetypal analysis of galaxy spectra. *Monthly Notice of the Royal Astronomical Society, 338*, 790–795.

Cook, D., & Swayne, D. F. (2007). *Interactive and dynamic graphics for data analysis: With R and GGobi*. Berlin, Heidelberg: Springer, Use R Series.

Coxeter, H. S. M. (1969). *Introduction to geometry* (2nd ed., pp. 216–221). §13.7, Barycentric coordinates. New York: Wiley.

Cutler, A., & Breiman, L. (1994). Archetypal analysis. *Technometrics, 36*, 338–347.

D'Esposito, M. R., & Ragozini, G. (2008). A new R-ordering procedure to rank multivariate performances. In *Quaderni di Statistica* (Vol. 10, pp. 5–21). Italy: Liguori Editore.

D'Esposito, M. R., Palumbo, F., & Ragozini, G. (2006). Archetypal analysis for interval data in marketing research. *Italian Journal of Applied Statistics, 18*, 343–358.

Ein-Dor, P., & Feldmesser, J. (1987). Attributes of the performance of central processing units: A relative performance prediction model. *Communications of the ACM, 30*, 308–317.

Elder, A., & Pinnel, J. (2003). Archetypal analysis: An alternative approach to finding defining segments. In *2003 Sawtooth Software Conference Proceedings* (pp. 113–129). Sequim, WA.

Gamma, E., Helm, R., Johnson, R., & Vlissides, J. (1995). *Design patterns: Elements of reusable object-oriented software*. Reading Mass.: Addison Wesley.

Heavlin, W. D. (2007). Archetypal analysis of computer performance. *38th Symposium on the INTERFACE on Massive Data Sets and Stream*, Pasedena, CA.

Lang, D. T., Swayne, D., Wickham, H., & Lawrence, M. (2008). *rggobi: Interface between R and GGobi*. R package version 2.1.10, URL http://www.ggobi.org/rggobi.

Li, S., Wang, P., Louviere, J., & Carson, R. (2003). Archetypal analysis: A new way to segment markets based on extreme individuals. *A Celebration of Ehrenberg and Bass: Marketing Knowledge, Discoveries and Contribution, ANZMAC 2003 Conference Proceedings* (pp. 1674–1679). Adelaide.

Marinetti, S., Finesso, L., & Marsilio, E. (2006). Matrix factorization methods: Application to thermal NDT/E. *NDT&E International, 39*, 611–616.

Marinetti, S., Finesso, L., & Marsilio, E. (2007). Archetypes and principal components of an IR image sequence. *Infrared Physics & Technology, 49*, 272–276.

Porzio, G. C., Ragozini, G., & Vistocco, D. (2008). On the use of archetypes as benchmarks. *Applied Stochastic Models in Business and Industry, 24*, 419–437.

R Development Core Team. (2008). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Stone, E. (2002). Exploring archetypal dynamics of pattern formation in cellular flames. *Physica D, 161*, 163–186.

Stone, E., & Cutler, A. (1996). Introduction to archetypal analysis of spatio-temporal dynamics. *Physica D, 96*, 110–131.

Stuetzle, W. (1987). Plot windows. *Journal of American Statistical Association, 82*, 466–475.

Tukey, J. (1977). *Exploratory Data Analysis*. Reading, MA, U.S.A.: Addison-Wesley.

Wegman, E. J., & Carr, D. B. (1993). Statistical graphics and visualization. In C. R. Rao (Ed.), *Handbook of Statistics: Computational Statistics* (Vol. 9, pp. 857–958). NY, U.S.A.: Elsevier, North Holland.

Wilhelm, A. (2005). Interactive statistical graphics: The paradigm of linked views. In C. R. Rao, E. J. Wegman and J. L. Solka (Eds.), *Data mining and data visualization* (Vol. 24 of Handbook of Statistics, pp. 437–537). NY, U.S.A.: Elsevier, North-Holland.

Yang, D., Rundensteiner, E. A., & Ward, M. O. (2007). Analysis guided visual exploration to multivariate data. *IEEE Symposium on Visual Analytics Science and Technology*. Sacramento, California.

Young, F. W., Faldowski, R. A., & Harris, D. F. (1992). The spreadplot: A graphical spreadsheet of algebraic linked of dynamic plot. In: *ASA Proceedings Section on Statistical Graphics*. American Statistical Association, Alexandria, VA.

Young, F. W., Faldowski, R. A., McFarlane, M. M. (1993). Multivariate statistical visualization. In C. R. Rao (Ed.), *Handbook of Statistics: Computational Statistics* (Vol. 9, pp. 959–998). NY, U.S.A.: Elsevier, North Holland.