{MG&V 5(4), 1996, 563–577 }

SOME ISSUES CONNECTED WITH A 3D REPRESENTATION OF MULTIVARIATE DATA POINTS

Anna Bartkowiak, Adam Szustalewicz

Institute of Computer Science, University of Wrocław,

Przesmyckiego 20, Wrocław 51–151, Poland

Abstract

The presented considerations complement our former paper (*B&Asz*) entitled "The augmented biplot and some examples of its use" (Machine Graphics and Vision, 4, 1995, 161–185). Now we consider some topics concerned with a 3D representation of n points–individuals and p points–variables included in a data matrix $\mathbf{X}_{n\times p}$. The representation is displayed in the form of a spinplot called also spinner. Of course, the 3D representation exhibits the true interrelation structure between the displayed points only to some degree of accuracy.

We are concerned with the display of two kinds of information: (i) the goodness of the representation of individual points, and (ii) recognizing the mutual position (the out–of–page position) of several points when viewed in a projection plane displaying the flat projection of the whole system which is spinned by performed rotations.

Our considerations and needs coming from statistical data analysis problems resulted in a computer program named ASZE (A Slice and siZE displaying plot).

We present 3D displays made by this program for the oat varieties data and locations of their growing and the farm production data considered formerly by *B&Asz*. We show on these examples what a variety of questions can be answered when constructing the spinner with appropriate enhancements.

Key words: biplot, 3D plot, spinner, interdependence between variables, reduction of dimensionality

1 Introduction. What is a biplot representation

When considering a data matrix $\mathbf{X}_{n\times p}$ containing values of p variables measured on n individuals we may interpret the rows of $\mathbf{X}$ as points in R^p and the columns of $\mathbf{X}$ as points in R^n. In the following we will assume that $rank(\mathbf{X}) = r \geq 3$, also that $min(n, p) \geq 3$.

If $\mathbf{X}$ is of rank r, then also $\mathbf{X}^T\mathbf{X}$ is of rank r and as such has r positive eigenvalues. Let $\lambda_1 \geq \ldots \geq \lambda_r > 0$ and $\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_r]$ denote the positive eigenvalues and associated with them eigenvectors of the matrix $\mathbf{X}^T\mathbf{X}$, and let $\mathbf{L} = diag(\lambda_1^{1/2}, \ldots, \lambda_r^{1/2})$ be the diagonal matrix containing square roots of the respective eigenvalues.

It is known [1, 3, 7] that – using the *svd* algorithm – the matrix $\mathbf{X}$ defined above can be decomposed into the product of two rank r matrices, $\mathbf{G}_{n\times r}$ and $\mathbf{H}_{r\times p}$ as:

$$\mathbf{X} = \mathbf{GH} = \begin{bmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_n \end{bmatrix} [\mathbf{h}_1, \ldots, \mathbf{h}_p], \tag{1}$$

where $\mathbf{G} = \mathbf{XAL}^{-1}$, $\mathbf{H} = \mathbf{LA}^T$. The above decomposition is called the GH decomposition. It is used for drawing a graph called biplot, which is a 2D representation of both the individuals and the variables appearing in $\mathbf{X}$.

The traditional biplot is drawn by taking the first two columns of $\mathbf{G}$ as coordinates of the n points–individuals, and the first two rows of $\mathbf{H}$ as coordinates of points–variables.

The overall goodness (truthworthness) of the 2D representation given by the biplot – as compared with the true configuration of the points–individuals in R^p and points–variables in R^n is given by the ratio $(\lambda_1 + \lambda_2)/\sum_{j=1}^{r} \lambda_j$.

If the matrix $\mathbf{X}$ is scaled so that $\mathbf{X}^T\mathbf{X} = \mathbf{R}$, the correlation matrix of the p variables, then the following decomposition formula holds:

$$\mathbf{R} = \mathbf{H}^T\mathbf{H}. \tag{2}$$

Considering only the diagonal of $\mathbf{R}$ (containing unit variances) we obtain:

$$diag(r_{11}, r_{22}, \ldots, r_{pp}) = diag(\underbrace{1, 1, \ldots, 1}_{\text{p times}}) = \sum_{k=1}^{r} diag(h_{k1}^2, h_{k2}^2, \ldots, h_{kp}^2). \tag{3}$$

The formula above says that the unit variances of the considered variables can be reproduced exactly from the squared row elements of $\mathbf{H}$ – when taking all of them. Thus, consequently, taking only the first two rows from $\mathbf{H}$ for drawing the biplot we know accurately for each of the p points–variables, which of them is represented fairly, and which is not. If some variables are represented fairly, then

also their mutual interrelations (configuration in R^p) can be seen quite accurately in the biplot.

2 Adding the 3rd dimension to a biplot representation

2.1 Is it worth to add the 3rd dimension?

Obviously, by taking the first 3 columns of **G** and the first 3 rows of **H** a better graphical representation of the individuals and variables is obtained as when taking only two of them. We consider then a 3 dimensional coordinate system. The overall goodness of representation of the true configuration in R^p (for individuals) and R^n (for variables) is then improved by the amount $\lambda_3/(\lambda_1 + \ldots + \lambda_r)$. When considering individually subsequent variables then, according to (3), we obtain for them improvements of the representation by amounts $h_{31}^2, h_{32}^2, \ldots, h_{3p}^2$ respectively.

Is it worth to consider the 3rd dimension? We will obtain an answer to this question by considering, how large is the value of λ_3 as compared to the sum $\lambda_1 + \ldots + \lambda_r$.

Say, we have decided to consider the 3–dimensional representation (3D). Which variables will then be represented in a satisfactory manner? We obtain an answer to this question by considering the partial cumulative sums

$$\sum_{k=1}^{3} h_{k1}^2, \quad \sum_{k=1}^{3} h_{k2}^2, \quad \ldots, \quad \sum_{k=1}^{3} h_{kp}^2 \tag{4}$$

in relation to the total cumulative sums characterizing a perfect (ideal) representation of the variables:

$$\sum_{k=1}^{r} h_{k1}^2, \quad \sum_{k=1}^{r} h_{k2}^2, \quad \ldots, \quad \sum_{k=1}^{r} h_{kp}^2. \tag{5}$$

If for a considered variable (j) the ratio of its partial cumulative sum $\sum_{k=1}^{3} h_{kj}^2$ to its total cumulative sum $\sum_{k=1}^{r} h_{kj}^2$ is close to 1.0, then this variable is quite fairly represented in the 3D structure; on the opposite, if the respective ratio is close to zero, then the 3D representation of this variable is poor, and we should be very

careful when drawing inference about the mutual interrelations of this variable with other variables.

Let $\mathbf{r}_i$ denote the ith row (individual) of the data matrix $\mathbf{X}_{n\times p}$. Let x_i, y_i, z_i denote the coordinates of the ith individual when displayed in the spinner. Then, according to (1):

$$x_i = \mathbf{r}_i\mathbf{a}_1\lambda_1^{-1},\ y_i = \mathbf{r}_i\mathbf{a}_2\lambda_2^{-1},\ z_i = \mathbf{r}_i\mathbf{a}_3\lambda_3^{-1}\ \ (i = 1, \ldots, n), \tag{6}$$

where $\lambda_k, \mathbf{a}_k$ $(k = 1, 2, 3)$ denote the largest eigenvalues and associated with them eigenvectors of $\mathbf{X}^T\mathbf{X}$.

When the calculations were carried out on data matrix $\mathbf{X}$ scaled so that $\mathbf{X}^T\mathbf{X} = \mathbf{R}$, the correlation matrix, then all variables have the same (unit) variances. In such case the coordinates x_i, y_i, z_i can be interpreted as transformed ("new") variables obtained by evaluating linear combinations of the original ("old") variables (columns of the data matrix $\mathbf{X}$) with equal weights for each "old" variable. Formula (6) above tells us, that the linear combinations forming the "new" variables are designated by components of the eigenvectors $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$.

We will use this property when commenting the display of the last example presented in next section.

2.2 Viewing the 3D representation

Elaborated details on viewing a 3D configuration of points by projecting it onto a plane are well known and may be found e.g. in [2]. Below we recall some elementary principles applied (used) in our computer program ASZE (A Slice and siZE displaying plot) for visualizing graphically a configuration of n points.

We assume that we have a basic orthogonal 3-dimensional coordinate system $B = (\vec{e}_1, \vec{e}_2, \vec{e}_3)$, with respect to which the coordinates $\{x_i, y_i, z_i\}$ $(i = 1, \ldots, n)$ of the given n points are referred to. The coordinate system B together with the n displayed points will be called in the following *the (3D) Configuration* – referred to by this name in upper case.

An observer is viewing this Configuration. What the observer sees is only the projection of the Configuration onto a plane perpendicular to the direction of the

observers viewing.

Let the initial (*Home*) position of the coordinate system be such, that the x-axis (determined by $\vec{e}_1$) is horizontal, the y-axis (determined by $\vec{e}_2$) is vertical, and the z axis (determined by $\vec{e}_3$) is out–of–page and directed to the observer. At such starting position the observer sees only the (x, y) values, and he does not perceive the depth (the third dimension) at all.

The observer may perform a rotation of the Configuration. The rotation may be performed round the horizontal axis, vertical axis, or out–of–page axis. By applying a sequence of subsequent rotations carried out quickly one after another the observer may perceive an impression of movements of the points and axes included in the Configuration.

The movements are carried out by hitting some keys on the computer keyboard, or by clicking some mouse buttons.

The movements of the Configuration can be faster or slower; there might be also a superpositioning of fast 3D rotation (gyring) with some slower tour motion (gimbling) [8].

2.3 How to perceive the depth

By rotating the Configuration one may obtain an idea on the mutual position of the displayed points. However – as we have observed it when showing the graphs to various people – this may be quite difficult. Therefore – apart from the simple display of points in the projection plane – some additional enhancements should be used. Some of these enhancements could be:

1. Hologramic views
2. Shadowing and using hues
3. Depth cuing
4. Slicing.

The first two methods are successfully launched by E. Wegman and his collaborators [5, 6, 10]). They use high quality computers with high resolution

screen and apply the technique mainly to producing graphs exhibiting some surfaces. Especially when projecting some blocks of solids these technique help the observer to get the impression which parts of the solids are nearer and which further from his standpoint. These techniques cannot be applied when displaying points which in principle are dimensionless, although they are usually displayed by dots or other symbols.

The method of depth cuing, used e.g. by L. Tierney in XLisp-Stat [4] differentiates the size of the displayed dots – or other symbols denoting the points – according to their out–of–page coordinate.

Slicing allows for subdividing all displayed points according to their out–of–page coordinate into several classes called slices. All points belonging to the same slice obtain the same symbol size. For slice which is the nearest of the observer the symbol size is the largest; for the furthest slice the symbol size is the smallest.

3 Presentation of some real examples

In the following we will consider 3 real examples. The data were already dealt with by Bartkowiak and Szustalewicz [7] who constructed for these data the augmented biplots. The data sets considered here concern:

1. Yields of 8 oat varieties – considered as "variables" – cultivated in 13 locations – considered as "individuals". Thus we will deal with a data matrix $\mathbf{X}$ of size 13×8.
2. The transposed data matrix $\mathbf{X}$ just introduced in point 1 above. Now we consider the 13 locations as "variables" and the 8 varieties as "individuals", thus we will deal with a data matrix $\breve{\mathbf{X}}$ of size 8×13.
3. Some economic indices characterizing 31 farms. In this example the farms are considered as "individuals". Each farm is described by 6 indices ("variables") with the following meaning: $X1$ – area of land (in ha), $X2$ – investment of human work (in working days), $X3$ – the livestock (number of big animals), $X4$ – inventory stock. $X5$ – expenditure for fertilizers, $X6$ – vegetable production of the farm.

We will deal here with a data matrix $\mathbf{X}$ of size 31×6.

3.1 The goodness of the representation – the GR index

The accuracy of the representation of the "variables" – when taking $k = 2$, $k = 3$ and $k = r(= all)$ components from the GH decomposition is given in Table 1. Tab 1 Remind, that for $k = 2$ we draw a biplot, and for $k = 3$ we construct a spinner.

The accuracy (goodness) of the graphical representation will be called in the following the GR index.

Looking at the values shown in Tab. 1 we note that for the f i r s t data set (8 oat varieties denoted as $V1, \ldots, V8$) the representation in the biplot is satisfactory. We note also that by adding the 3rd dimension we will obtain only a minuscule (1%) improvement of the GR index.

For the s e c o n d data set (13 Locations denoted as $L1, \ldots, L13$) the GR index is not satisfactory: neither for $k = 2$ nor for $k = 3$. The data contain a big outlier (location nb. 11). After dropping this outlier the GR index improves a little, nonetheless the representation still remains unsatisfactory for $k = 2$. When accounting also for the third dimension, we obtain a satisfactory representation for 5 variables only ($L8, L9, L10, L12, L13$). Thus by adding the 3rd dimension we improve considerably the GR index of two variables only ($L12$ and $L13$).

For the t h i r d example (Economic indices $X1, \ldots, X6$ characterizing 31 farms) the GR index for $k = 2$ (representation in the biplot) is nearly satisfactory: for 3 variables it lies in the range 79–89% and for the other 3 variables in the range 91–96%. When taking $k = 3$ (representation in the spinplot) the GR index increases for all variables to attain in the average the value 94%. Thus for these data adding the 3rd dimension helps really in obtaining a better visualization of the interdependence structure among the considered variables.

3.2 Graphical presentation in the form of a spinner

Our computer program ASZE offers a number of options which permit to display the spinner in various combinations.

The data are read by items. Each item may represent one "variable" or one "individual". The data for each item are read from a text file rowwise, each item in separate row. The data for an item contain firstly its coordinates, then also – as last two numbers in the row – its label and "group" number.

A group number "0" means that the data array should be interpreted as "variable" – then it will be displayed as a spike holding a (small) ball. Group numbers > 0 will be presented as isolated balls (circles) in differentiated colors and fillings.

The rotation in the spinner is performed by pushing some keys from the keyboard. We may perform 3 kinds of rotations. We may use the $\downarrow\uparrow$ keys for rotating round the horizontal axis, the $\leftarrow\rightarrow$ keys for rotating round the vertical axis, and the *PgUp,PgDn* keys for rotating round the out–of–page axis. Each push (hit of the key) performs a rotation by an angle α. The magnitude of α can be chosen: the default is $\alpha = 5^o$.

We assume that the observer is looking at the displayed configuration along the out–of–page axis. What the observer perceives is the projection of the displayed configuration onto the plane spanned by the horizontal and the vertical axes (we call this plane the "screen page").

The display consists of two panels.

The l e f t panel, a smaller one, describes possible actions (hot keys) which may be undertaken by the user. So after pressing the B key we obtain in this panel a *bar plot* exhibiting how much of the total variance (or: squared length of the $\mathbf{h}_j$ vector, see formula (5)) is accounted for (reproduced) by the 1st, 2nd and 3rd component of the GH decomposition. Variables, for which the reproduction is not heigh (say, less then 0.90) have a bad representation in the spinner and we should not trust the exhibited configuration.

The r i g h t panel, in the shape of a square, exhibits the spinner.

The program ASZE permits for two modes of exhibiting the spinner: sliced and non sliced.

(i) *Sliced.* The exhibited points appear as filled balls (circles) with size de-

pending on the out–of–page coordinate.

(ii) *Non sliced*. All points are presented as balls (circles). A filled circle means that the point is located before the screen page (the same where the standpoint of the observer is located); an empty circle indicate for a location behind the screen page.

Now we will present some snapshots from spinners constructed for the 3 mentioned data sets.

3.2.1 Yields of 8 oat varieties considered as variables characterizing different locations

The *home* position of the spin plot obtained for these data is shown in Fig. 1. The variables are presented as small balls located on spikes emanating from the point (0,0,0) of the coordinate system. The locations are presented as small isolated circles. In the same figure we have displayed the bar plot showing that the representation in the spinner is a very good one.

Among the points denoting locations (they appear here as belonging to group 1) we note a very big outlier: location nb. 11.

Some variables (no. 8 and 6, also no. 1, 5 and 9) seem to be nearly collinear, which means that the respective varieties give very similar yields in the considered locations. To find out whether this statement is not spurious (these varieties could still differ in the 3rd, i.e. in the "z" coordinate) we rotate the spinner so that the 3rd component is visible. The rotated position of the spinner is shown in Fig.2. One can see that also when looking at the 3rd coordinate, the balls representing the mentioned sets of varieties are very near. Hence we may infer, that the indicated varieties (i.e. no. 8 and no. 6, also no. 1, 5 and 9) are very similar with respect to their yields.

3.2.2 Comparing locations in which the oat varieties are cultivated

We have removed from the data the outlying location nb. 11 – thus the analysis was performed with 12 locations only. Now the yields in these locations were considered as "variables" characterizing the varieties.

The *Home* position of the spin plot, with the *Bar plot* and *Slicing* options on, is shown in Fig. 3. Looking at the bar plot one can see that for several items the representation in the spinner is not trustworthy. Looking despite this doubt at the plot we obtain the impression that the spikes indicating for the locations are subdivided into two sets – according to the second ("y") component. This impression is sustained after rotating the system – what can be seen in Fig. 4. A land specialist could perhaps tell why the locations {6 10 12 13} and {7 9 8 1 3 } appear in two clusters.

The bad representation of the locations in the spinner is signalled only by the bar plot in the left panel of the display. To make it more obvious and more appealing to the observer, we have made from the worst represented locations (which have the *GR* index < 0.80) a separate group (group no. 2). The spinner obtained for such layout is shown in Fig. 5. Now the badly represented locations are shown as isolated balls with dotted filling. Since *Slicing* is on, the sizes of the balls are differentiated according to their out–of–page coordinate.

3.2.3 Economic indices for 31 farms

We have for these data $p = 6$ variables denoting the economic indices, (items read in as belonging to group 0) and $n = 31$ farms characterized by these indices and classified as belonging to group 1. This makes together 37 items.

The *Home* position of the spinner constructed for these data is shown in Fig. 6. The exhibit was captured when the options: *Bar plot*, *Slicing*, *Labelling* group 1 were off; however the initial balls (circles) were increased by pushing several times the G key.

Looking at Fig. 6 one can see that the variables no. 5 and 6 are very close. This is sustained after rotating the system by 90^o, putting the *Slicing* on and removing all the slices but the first two. We come then to the display shown in Fig. 7. Again, the balls labelled $X5$ and $X6$ are nearly undistinguishable.

We have tried to rotate the system to such position that the balls $X5$ and $X6$ would be distant. We could not achieve it! The best for this purpose seemed to be the view reproduced in Fig. 8 (here again the *Slicing* is off). Let us allow

to give the comment, that the considered variables are in fact highly correlated ($r_{56} = 0.93$); thus the spinner displays really the true relation between these variables.

In Fig. 9 we show only points (balls) representing the farms. The display is in *Home* position of the spinner with *Slicing* on. In Fig. 9 we see all slices (i.e. all 31 points–farms).

The coordinates (x, y, z) of these points were obtained from the GH decomposition (see formula (1)) using the eigenvectors $(\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3)$ computed from the $\mathbf{X}^T\mathbf{X}$ matrix that was in our case the correlation matrix $\mathbf{R}$ of the 6 indices (see formula (6)). The eigenvectors $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ have the following structure:

	$X1$	$X2$	$X3$	$X4$	$X5$	$X6$
$\mathbf{a}_1$	0.42	0.34	0.41	0.41	0.42	0.44
$\mathbf{a}_2$	−0.28	0.79	0.33	−0.10	−0.31	−0.29
$\mathbf{a}_3$	0.48	−0.11	0.41	−0.77	−0.05	−0.01

Since the original indices were all normalized to unit variances, we may infer that:

1. the 'x' coordinate of the transformed points–farms expresses generally their magnitude with respect to all indices,

2. the 'y' coordinate expresses the prevalence of $X2$ (and eventually of $X3$) as opposed to remaining indices. Thus points (in the spinner) with large 'y' will indicate mainly farms with a great investment of human work ($X2$) and eventually great livestock ($X3$) as opposed to other indices,

3. the 'z' coordinate expresses the prevalence of small $X4$ as opposed to big $X1$ and $X3$. Thus points (in the spinner) with large 'z' will indicate farms with small inventory stock as compared to the area of the land and the livestock.

To exhibit only farms with small inventory stock (condition (3) above) we have firstly removed slices 3–8 and next we have put on labelling of group 1. The obtained display is shown in Fig. 10. One can see that these are farms numbered 20, 26, 29, 27, 17, 13.

4 Final remarks

A 3–dimensional structure with possibility of rotation is called spinner.

We have shown on some real examples, that adding the 3rd dimension to the biplot, and thus constructing a spinner, may result in a better representation of the data structure.

Adding some small enhancements to graphical visualization done by a spinner may greatly enrich the results of data analysis.

Generally, when programming our program ASZE (serving for displaying a spinner) we have put emphasis on goodness of representation of variables. As mentioned when analyzing the third example from this paper, it might be also interesting to look in more detail individually on the goodness of representation of the "individuals" (i.e. of representations of rows of a data matrix $\mathbf{X}_{n\times p}$), especially in such cases when we are concerned with drawing inference on the closeness or distances between these individuals. To deduce trustworthy and reliable conclusions on this topic we need more refined tools, which are not provided by the computer program ASZE.

Another example of using a spinner is shown in [9]. The authors have considered displays by a spinner in the context of regression analysis performed for some spirometric data. The spinner permitted to choose variables worth to be included into the regression equation used for predicting values of two regressands.

References

1978

[1] Gabriel K.P.: Least squares approximation of matrices by additive and multiplicative models. *J.R.Statist. Soc.* ser. B., 40, 186–196.

1986

[2] Berger M.: *Computer Graphics with Pascal.* Benjamin/Cunnings, Menlo Park Ca.

[3] Jolliffe I.T.: Principal Component Analysis. Springer, New York.

1990

[4] Tierney L.: *Lisp–Stat, An Object-Oriented Environment for Statistical Computing and Dynamic Graphics.* Wiley, N.Y.

1993

[5] Wegman E.J., Carr D.B.: Statistical graphics and visualization. In: C.R. Rao (ed), *Handbook of Statistics*, Vol. 9. Elsevier Sc., North Holland, Amsterdam, 857–958.

1994

[6] Wegman E.J., Luo Q.: Visualizing Densities. *Techn. Rep.*, No 100, George Mason Univ, Fairfax.

1995

[7] Bartkowiak A., Szustalewicz A.: The augmented biplot and some examples of its use. *Machine Graphics & Vision*, 4, no 3–4, 161–185.

[8] Buja A., Cook Di, Asimov D., Hurley C.: Theory and Computational Methods for Dynamic Projections in High–Dimensional Visualization. *Working Paper*, pp. 1–41. AT&T Bell Laboratories, Murray Hill N.Y. 1–41.

1996

[9] Bartkowiak A., Liebhart J., Szustalewicz A.: Visualizing the correlation structure of some spirometric data by a biplot extended to 3 dimensions. In: J. Doroszewski and L. Bobrowski (eds), *Statistics in Clinical practice*, 2nd Int. Seminar 25 – 30 June 1996, Warsaw, 50–53.

[10] Wegman E.J., Luo Q.: High Dimensional Clustering Using Parallel Coordinates and the Grand Tour. *Manuscript prepared for Proceedings of the 27nd Symposium on the Interface (Interface ´96)*, Sydney.