VISUALIZING THE CORRELATION STRUCTURE BY A BIPLOT EXTENDED TO 3 DIMENSIONS

A.Bartkowiak, J.Liebhart*, A.Szustalewicz

Institute of Computer Science, University of Wrocław, Wrocław, Poland
*Department of Internal Diseases and Allergology, Medical Academy of Wrocław, Wrocław, Poland

Motto

A single picture is worth a thousand numbers (after F.W.Young)

Plan

- Recalling the method of biplot as approached by Gabriel (1978,1990).
- Saying a few words on the uses and misuses of a biplot presentation.
- Presenting the augmented biplot (Bartkowiak & Szustalewicz 1995) as a method of incorporating into the drawn biplot some additional information on its representativeness.
- Showing how useful a biplot might be when the 3rd dimension is added.
- Some proposals and considerations how to graph a 3-dimensional structure effectively to allow the user to perceive the true distances between points-variables.
- Analysing in detail one medical data set containing observational data (results of spirometric examinations) for patients suffering from bronchial asthma or chronic obturative pulmonary disease (Liebhart et al. 1989).

Main points of elaboration

Gabriel (1978) has shown that a data table **X** of size n x p can be presented as the product of two matrices, **G** and **H**:

$$\mathbf{X} = \mathbf{GH}, \text{ size } \mathbf{G} = n \times r, \text{ size } \mathbf{H} = r \times p,$$

with r denoting the rank of **X**.

The biplot visualization is based on the first two columns of **G** (coordinates of points-individuals) and the first two rows of **H** (coordinates of points-variables) and results in a plot exhibiting in the same graph points-individuals and points-variables.

In the following we will be concerned only with the representation of variables, i.e. with the representation $\mathbf{H}=(\mathbf{h}_1,\ldots,\mathbf{h}_p)$, which is graphed in a plane as a set of vectors anchored at the point (0,0) of the plane.

It has been proved, that the correlation matrix **R** can be decomposed as:

$$R=H^TH=(h_i^Th_j)=(\sum_{\nu=1}^{r} h_{ri}h_{rj}), \quad i,j=1,\dots,p.$$

The presented formula shows that the entire correlation matrix **R** can be reproduced completely from elements of the vectors $\mathbf{h}_1,\dots,\mathbf{h}_p$. This is true for the diagonal elements of **R** and for each correlation coefficient r_{ij} as well.

If the rank of **X** (and consequently, that of **H**) were r=2, then we would obtain a perfect (fair) representation of correlations between all the variables.

However, in most cases, the rank r is greater then 2, thus the representation in the biplot is not fair.

For the analyzed data (containing patients with obturative function loss) we have been considering 11 variables: X1 - Residual Volume (RV), X2 - Total Lung Capacity (TLC), X3 - Age, X4 - Height, X5 - Vital Capacity (VC), X6 - Due Vital Capacity (in % DVC), X7 - Forced Expiratory Volume in the First Second (FEV1), X8 - Tiffany Index (Tiff), X9 - Forced Expiratory Flow at the level of 0.2-1.2 VC (FEV), X10 - Maximal Mid-Expiratory Flow Rate (MMFR), X11 - Maximal Mid-Expiratory Flow Time (MMFT).

The variables X3-X9 have been considered by Liebhart et al. (1989) as predictors for RV and TLC.

We will take for our illustrations the variables set (X3÷X9) - to show the correlations among the predictors, and next the variables set (X1÷X9) - to show the interrelations among the regressands and the predictors.

When considering the correlation matrix of the 9 predictors we obtained the following approximation to the diagonal of **R**:

Dimension	X3	X4	X5	X6	X7	X8	X9	X10	X11	Overall
$r=1$	.35	.27	.56	.26	.97	.54	.89	.69	.45	0.55
$r=2$	**.36**	**.58**	.97	**.39**	.97	.93	.89	.71	.85	0.74
$r=3$	**.42**	.92	.98	.98	.97	.93	.89	.74	.85	0.85
$r=9$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.00

From this Table it is evident that the traditional biplot is unsatisfactory for visualizing graphically the interrelations with and among the variables X3, X4, X6, and eventually X10 and X11.

However, looking at the values shown in the last row of the Table above, one can see clearly that by adding the 3rd dimension the representativeness of the variables X4 and X6 improves from 0.58 and 0.39 to 0.92 and 0.98 respectively - which is very satisfactory. Then only one variable (X3) remains badly represented.

The 3-dimensional visualization of variables can be realized by a spinner. This means that we imagine the variables as a swarm of points located in 3D space with coordinates (x,y,z). We are looking at this 3D system from one fixed point (the stand of the observer) - and what we see (and perceive) is the projection of the swarm of points on a projection plane. The whole system (coordinate axes with points) can be rotated (say, be clicking a mouse button) - then we obtain another projection of the swarm of points. Repeating the rotation we may perceive the impression of motion of the points. What is essential: the whole time we see only a 2D projection of the true 3D structure.

Some important topics to be considered here are:

1. How to make the observer to perceive - when looking at the projection plane only - the depth of the points (the 3rd dimension), i.e. whether the displayed points are near or far from the observer.

2. How to incorporate into the spinner the information on the representativeness (or non-representativeness) of the 3D structure as compared to the true r-dimensional structure of the column vectors of **H**.

The 1st and 3rd authors have designed and implemented a PC computer program in Turbo Pascal solving (at least partically) the problems connected with the mentioned two topics. During the presentation of the paper we intend to show some interesting views obtained by the use of the program.

Another very important property of the GH decomposition is that the cosine of the angle between the vectors $\mathbf{h}_i$, $\mathbf{h}_j$ is equal directly (or proportional) to the correlation coefficient r_{ij} between the variables X_i and X_j. So, if these variables are represented fairly in the projection plane, on is justified to infer from their proximity or distance, whether they are highly or poorly correlated.

This again will be illustrated on biplots constructed for the elaborated data sets.

The biplots considered so far were evaluated from a group containing observations n=125 patients. For some comparative purposes we have drawn from this group a representative subsample containing only n=28 patients. The biplot (its 3D representation constructed from the subsample was partically the same as that for the entire sample. This statement sounds quite optimistic, especially when taking into account the small number of observations as opposed to the relatively large number of variables.

Below we present some exemplary values of correlation coefficients r_{ij} observed in the entire sample and in the chosen subsample:

Pair of	Correlation coefficient between pair						
variables	(11,8)	(5,4)	(7,9)	(6,5)	(8,9)	(8,10)	(10,7)
n=125	-0.84	0.69	0.97	0.68	0.72	0.66	0.76
n= 28	-0.82	0.64	0.95	0.69	0.74	0.82	0.79

All these relations can be confirmed (read up) from the respective biplots.

So far we have visualized only the interrelations between the variables (X3÷X9) considered by Liebhart and al. as predictors for RV and TLC. Now let us add to the analysis the variables X1 (=RV) and X2 (=TLC) and construct the biplot from all the 11 variables. The structure for the predictors should remain practically the same (i.e. as that obtained from 9 variables only). Two additional vectors - corresponding to the variables X1 and X2 - appear in the plot. Now we should look at their position relatively to the vectors representing the predictors. With a fair representation of the considered points-variables in the biplot one can see directly whether the assumed predictors (X3,...,X9) are good for predicting X1 and X2. With a bit of luck we might then select the relevant predictors.

We were able to do this for X2, however were not lucky enough to do this for X1.

References

1. Bartkowiak A., Szustalewicz A.: The augmented biplot and some examples of its use, Machine Graphics & Vision, 1995, 161-185.

2. Gabriel K.P.: Least squares approximation of matrices by additive and multiplicative models, J.R. Statist. Soc., 1978, B, 40, 186-196.

3. Gabriel K.R., Odoroff Ch.L.: Biplots in medical research, Statistics & Medicine, 1990, 9, 469-485.

4. Liebhart J., Bartkowiak A., Liebhart E.: The impact of outliers in the regression estimating TLC from age and some spirometric observations, Modelling, Simulation & Control, C, AMSE Press, 1989, 15, 1-18.

ABA

POLSKA AKADEMIA NAUK
MIĘDZYNARODOWE CENTRUM BIOCYBERNETYKI
INTERNATIONAL CENTRE OF BIOCYBERNETICS
WORLD HEALTH ORGANIZATION
COLLABORATING CENTRE FOR RESEARCH AND TRAINING
IN BIOCYBERNETICS AND BIOMEDICAL ENGINEERING

Address: 02-109 Warszawa, ul. Trojdena 4, Poland, Tel: (+4822) 6599143, Fax: (+4822) 6582872
Director: Prof. Dr. Maciej Nałęcz

34th ICB Seminar

STATISTICS AND CLINICAL PRACTICE

Chairmen: Prof. Jan Doroszewski
Assoc. Prof. Leon Bobrowski

Warsaw, 24 - 28 June, 1996

CONTENTS