{file biaug.tex, Machine Graphics &Vision, Vol. 4 (1995), pp. 161–185 }

THE AUGMENTED BIPLOT AND SOME EXAMPLES OF ITS USE

Anna Bartkowiak, Adam Szustalewicz

Institute of Computer Science, University of Wrocław
Przesmyckiego 20, Wrocław 51-151, Poland

Abstract.

Biplot is an explorative method of data analysis permitting to represent graphically, usually in a plane, the interrelations among points-variables and points-individuals located in a multivariate space. This is done by making projections from the multivariate space onto two- or three-dimensional subspaces. The crucial issue is: to what extend the projections in the lower dimension subspaces reflect the true relations of points-variables and points-individuals in the full data space? It happens that sometimes the representation given by the biplot is a good one, however sometimes it is a bad one and certainly not sufficient.

We show exactly wherefrom (i.e. from which theorems) some inferential properties of a biplot can be deduced and under which circumstances the relations visualized in the biplot are trustworthy.

We propose to construct the biplot in an extended mode which permits to judge the adequacy of the two-dimensional approximation visualized by the classical biplot. We call the biplot drawn in the extended mode the *augmented biplot*.

Several real data examples illustrate the use of the augmented biplot and the broadness and diversity of problems which can be elucidated relatively simply by use of the elaborated technique.

Key words and phrases: Data matrix, exploratory data analysis, reduction of dimensionality, graphical representation of multivariate points and interrelations from multivariate space.

THE AUGMENTED BIPLOT AND SOME EXAMPLES OF ITS USE

Anna Bartkowiak, Adam Szustalewicz

Institute of Computer Science, University of Wrocław
Przesmyckiego 20, Wrocław 51-151, Poland

1 Introduction

A human can inspect easily two-dimensional presentations. With a little bit effort it is possible for him to inspect also 3-dimensional structures.

Viewing objects in more than 3 dimensions has seemed for long beyond of scope for human perception.

The contemporary developments of computer–aided methods made it possible to move into essentially multivariate spaces and make there "grand tours".

In the following we deal with such a situation, when the to be inspected objects are multivariate points clustered in some clouds. Usually each of the points denotes one object or individual, for whom several (or many) characteristics have been recorded. The collection of all the values for the observed objects is called "measurement data matrix", "data matrix", or simply "data".

One of the major problems faced by the contemporary data analyst is to find out the characteristic pattern hidden in the analysed data. For multivariate data – note, that the gathered observations are mostly such ones – this is not a trivial task, because the number of aspects, under which the data could be viewed, is rapidly increasing with the number of the analysed traits. On the other hand, very often the observed traits are correlated, what

implies that some redundancy is met in the recorded data values - which in turn gives hope that a fair representation in a space of lower dimension might be obtained.

Analyzing a multivariate data set we may be concerned with the following questions:

1. Is the data set redundant, i.e. could it be represented – perhaps by projections – in a space of lower dimension?

2. Suppose, we have observed p traits (variables) in n individuals. How much are these traits mutually correlated? Of course, there exist several techniques elaborated by mathematical statistics, which can give answer to this question by producing a huge output showing numerous tables of results and outcomes of various statistical tests relying on strict assumptions (that might be not valid for the considered data). Preferably, especially at the stage of exploratory data analysis, we would like to have a method which could provide in a simple way by a plain graphical output the answer: which are the basic relations amongst the considered variables.

3. Concentrating on the set of n individuals, each of them characterized by p traits, we would like to know which of these individuals are similar – in respect to the considered set of p traits – and which are not. Again it would be valuable to obtain a simple graph elucidating this point.

4. For further, more strict analysis of the data, it is extremely important to know, whether the data is relatively homogeneous, or – whether it contains some unusual features or outliers. Again we imagine, that some graphical procedures, if properly designed, might allow us to obtain some insight into the multivariate space containing the data values, and let us possibly discover some unusual features – if any – of the data, features which otherwise would be very difficult to detect.

By using the biplot technique we may obtain conspicuous answers to the aforementioned problems.

Biplot is a method of explorative data analysis primary proposed by Gabriel [3, 4, 10],

and independently by the French school of statisticians, see e.g. Lebart et al. [2]. The method starts from a data table $\mathbf{X}_{n\times p},\ n > p$, whose rows correspond to n individuals, and columns – to p variables measured in the n individuals. By making the singular value decomposition (SVD) of the table $\mathbf{X}$ some new matrices $\mathbf{Z}_{n\times 2}$ and $\mathbf{V}_{2\times p}$ are constructed, which are thought to be representative for the n individuals and the p variables under consideration. In turn the rows of the matrix $\mathbf{Z}$ and the columns of the matrix $\mathbf{V}$ can be viewed as coordinates of points located in a two-dimensional Euclidean space – and as such they can be plotted in a plane. Traditionally the n points $\{z_{i1},\, z_{i2}\},\ (i = 1, \ldots, n)$ are marked in the plane just as points, while the p points $\{v_{1j}, v_{2j}\},\ (j = 1, \ldots, p)$ are represented as a bundle of vectors originated in the point $(0, 0)$ of the coordinate system. Such simultaneous representation – in one common plane – of individuals and variables is called the biplot.

After drawing the biplot for the analyzed data and looking at the mutual location of points-individuals and points-variables, attempts are made to draw some inference (i) on the similarity of individuals, (ii) on the mutual correlations of the variables exhibited in the plot and (iii) on the interrelations amongst points of both kinds.

It should be emphasized that the matrices $\mathbf{Z}_{n\times 2}$ and $\mathbf{V}_{2\times p}$ generally do not exhibit all features of the interdependence between the considered individuals and variables – and therefore the plot can reflect their interdependence structure only crudely and partially. The representativeness by $\mathbf{Z}_{n\times 2}$ or $\mathbf{V}_{2\times p}$ – in relation to the whole data matrix $\mathbf{X}$ – may be a good one or a bad one, and the user should be aware of this fact. Very often the biplot – now available in several commercial software packages – is taken literally as it is, without remembering, that it is only a f l a t representation of true relations between the individuals and the variables located in some multidimensional space of observations.

We propose a modified method of drawing the biplot in a plane. Our algorithm draws the biplot in an extended mode and permits to judge to what degree the representation

of the variables in the plane reflects the true interrelation structure from the multivariate space of observations. We call the biplot constructed by our method the `augmented biplot`.

The plan of our paper is the following:

In Section 2 we recall firstly the SVD decomposition of the data table $\mathbf{X}_{n\times p}$ and methods of approximation of $\mathbf{X}$ by matrices of lower rank. Next we present the GH decomposition of the data table – as introduced by Gabriel. This decomposition is the basis of construction of the biplot. We show also formal proofs of some properties of the biplot constructed from the GH decomposition for a given data table $\mathbf{X}$. The presentations are necessary to understand wherefrom the construction of the biplot is originated. Further considerations will be based on statements established in that section. Our main point here is to show clearly, wherefrom some properties of the visualization by biplot can be deduced and under which circumstances they are valid.

In Section 3 we present the principles of drawing the traditional biplot with pointing to some its deficiencies. Then we make our proposal of drawing a biplot relaying on more information from the GH decomposition – the *augmented* biplot.

In Section 4 we present four examples of application of the augmented biplot considering some real problems and real data sets.

2 SVD and GH decompositions of a data table

Let $\mathbf{X}$ of size $n \times p$ be a data table containing observations of p variables on n individuals. Suppose that $n > p$, and $\text{rank}(\mathbf{X}) = r \leq p$. Without loosing the generality of our considerations we will assume throughout the paper that the columns of $\mathbf{X}$ are centered to zero, what means that $\mathbf{1}_n^T\mathbf{X} = \mathbf{0}_{1\times p}$, with $\mathbf{1}_n^T = \underbrace{[1, 1, \ldots, 1]}_{n \text{ times}}$.

2.1 The SVD decomposition

The SVD decomposition of a matrix $\mathbf{X}$ is a fundamental theorem appearing in many textbooks dealing with matrix computations (see, e.g., [11], [12]).

`Definition.` By the SVD decomposition of a data matrix $\mathbf{X}$ we mean its presentation in the form

$$\mathbf{X}_{n\times p} = \mathbf{U}_{n\times r}\ \mathbf{L}_{r\times r}\ (\mathbf{A}^T)_{r\times p}, \tag{1}$$

with matrices $\mathbf{U}_{n\times r}$, $\mathbf{L}_{r\times r}$, $\mathbf{A}_{p\times r}$ satisfying the following conditions:

$$\mathbf{U}^T\mathbf{U} = \mathbf{I}_r; \quad \mathbf{A}^T\mathbf{A} = \mathbf{I}_r; \quad \mathbf{L} = diag(l_1, \ldots, l_r), \quad \text{with} \quad l_1 \geq l_2 \geq \ldots \geq l_r > 0.$$

The symbol $\mathbf{I}_r$ denotes here the unit matrix of size $r \times r$, i.e. the matrix whose all diagonal elements are equal to one, and the off-diagonal elements are all equal to zero.

`Algorithm for constructing the SVD decomposition for a data table` $\mathbf{X}$.

Let $\lambda_1 \geq \ldots \geq \lambda_r > 0,\ (r \leq p)$ be the positive eigenvalues of the cross-product matrix $\mathbf{X}^T\mathbf{X}$, and let the column vectors $\mathbf{a}_1, \ldots, \mathbf{a_r}$ be the corresponding eigenvectors.

Let $\mathbf{L}$, $\mathbf{A}, \mathbf{U}$ be defined as:

$\mathbf{L} = diag(\lambda_1^{1/2}, \ldots, \lambda_r^{1/2}), \quad \mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_r], \quad \mathbf{U} = \mathbf{X}\mathbf{A}\mathbf{L}^{-1}$.

Then these matrices provide the SVD decomposition (1) of the data table $\mathbf{X}$.

The correctness of this algorithm is proved by multiplicating the matrices $\mathbf{U}, \mathbf{L}, \mathbf{A}$ as defined above. We should get then the primary matrix $\mathbf{X}$.

Using the formula for multiplication of two matrices, $\mathbf{A}$ and $\mathbf{B}$, the first of them given by its column vectors, and the second by its row vectors:

$$\mathbf{AB} = [\mathbf{a}_1, \ldots, \mathbf{a_m}] \begin{bmatrix} \mathbf{b}_1 \\ \ldots \\ \mathbf{b}_m \end{bmatrix} = \sum_{i=1}^{m} \mathbf{a}_i\mathbf{b}_i\,,$$

we obtain an equivalent representation of the SVD decomposition of the matrix $\mathbf{X}$:

$$\mathbf{X} = \mathbf{U}\mathbf{L}\mathbf{A}^T = \sum_{i=1}^{r} l_i\mathbf{u}_i\mathbf{a}_i^T\,. \tag{2}$$

Note, that each matrix $l_i \mathbf{u}_i \mathbf{a}_i^T$ appearing in the formula 2 above is a rank 1 matrix of size $n \times r$.

2.2 Approximation of a matrix X by lower rank matrices

Suppose now, we want to approximate the matrix $\mathbf{X}_{n\times p}$ of rank r by a matrix of lower rank, say $\mathbf{X}^{(m)}_{n\times p}$ of rank $m \leq r$.

To do this we should firstly establish a measure of goodness of fit or of the approximation error.

A commonly used measure of the approximation error is the so called Euclidean norm of the error matrix $\mathbf{E} = \mathbf{X} - \mathbf{X}^{(m)}$, which is defined as (see e.g. [12], p. 90)

$$\|\mathbf{E}\| = (tr(\mathbf{E}^T\mathbf{E}))^{1/2} = (\sum_{i=1}^{n}\sum_{j=1}^{p} e_{ij}^2)^{1/2}. \tag{3}$$

The problem, how to approximate the matrix $\mathbf{X}$ by lower rank matrices with minimal error – when using the Euclidean norm criterion – was solved firstly by Householder and Young in 1938 (see, e.g. [6]). Their result can be enounced as the following:

`Theorem on approximation of a matrix` $\mathbf{X}$ `by a matrix of lower rank.`

The best approximation of a matrix $\mathbf{X}_{n\times p}$ of rank r by a matrix $\mathbf{X}^{(m)}_{n\times p}$ of rank $m \leq r$, when minimizing the Euclidean norm of the error matrix $\mathbf{E} = \mathbf{X} - \mathbf{X}^{(m)}$, as obtained by taking as the approximation matrix the first m components of the SVD decomposition of $\mathbf{X}$, i.e. the matrix $\mathbf{X}^{(m)}_{n\times p}$ defined as follows:

$$\mathbf{X}^{(m)} = \sum_{i=1}^{m} l_i \mathbf{u}_i \mathbf{a}_i^T.$$

The error of this approximation equals:

$$\|\mathbf{X} - \mathbf{X}^{(m)}\| = (\sum_{i=m+1}^{r} \lambda_i)^{1/2}.$$

Obviously for $m = r$ the error is equal to zero.

The proof of the optimality of this approximation may be found in several numerical textbooks, e.g. [12].

The error of the approximation can be deduced as follows. By taking the SVD decomposition of the matrices $\mathbf{X}$ and $\mathbf{X}^{(m)}$ we obtain

$$\|\mathbf{E}\|^2 = \|\mathbf{X} - \mathbf{X}^{(m)}\|^2 = \|\sum_{i=1}^{r} l_i \mathbf{u}_i \mathbf{a}_i^T - \sum_{i=1}^{m} l_i \mathbf{u}_i \mathbf{a}_i^T\|^2 = \|\sum_{i=m+1}^{r} l_i \mathbf{u}_i \mathbf{a}_i^T\|^2 = \sum_{i=m+1}^{p} l_i^2 \|\mathbf{u}_i \mathbf{a}_i^T\|^2.$$

However, since the squared Euclidean norm of a matrix can be evaluated as the trace of its cross-product matrix, we obtain:

$$\|\mathbf{E}\|^2 = \sum_{i=m+1}^{p} l_i^2 \|\mathbf{u}_i \mathbf{a}_i^T\|^2 = \sum_{i=m+1}^{p} l_i^2 \, trace(\mathbf{a}_i \mathbf{u}_i^T \, \mathbf{u}_i \mathbf{a}_i^T) = \sum_{i=m+1}^{p} l_i^2 = \sum_{i=m+1}^{p} \lambda_i \,,$$

and the error of the approximation equals

$$Approximation\ error \ = \|\mathbf{E}\| = \|\mathbf{X} - \mathbf{X}^{(m)}\| = (\sum_{i=m+1}^{p} \lambda_i)^{1/2}. \tag{4}$$

Alternatively, instead of considering the error of the approximation, one could consider the goodness of the approximation, named also *goodness of fit*. This is defined as the ratio of the squared norm of $\mathbf{X}^{(m)}$ to the squared norm of $\mathbf{X}$:

$$Goodness\ of\ fit \ = \frac{\|\mathbf{X}^{(m)}\|^2}{\|\mathbf{X}\|^2} = \frac{(\lambda_1 + \ldots + \lambda_m)}{(\lambda_1 + \ldots + \lambda_r)}, \quad m \leq r. \tag{5}$$

The sum of all eigenvalues is called also (especially by French statisticians) the *total inertia* of the data cloud given by the matrix $\mathbf{X}$. Obviously

$$Total\ inertia \ = \|\mathbf{X}\|^2 = trace(\mathbf{X}^T\mathbf{X}) = \lambda_1 + \ldots + \lambda_r.$$

So the goodness of fit is established by taking into account how large part of the total inertia of $\mathbf{X}$ is reproduced by the approximating matrix $\mathbf{X}^{(m)}$.

2.3 The GH decomposition of a data table X

The GH decomposition was introduced and named by Gabriel [3, 4, 10], see also [6].

`Definition.` Let $\mathbf{U}, \mathbf{L}, \mathbf{A}$ be the matrices yielding the SVD decomposition of a data table $\mathbf{X}$ – as given by formula (1). Substitute: $\mathbf{G} := \mathbf{U}, \quad \mathbf{H} := \mathbf{L}\mathbf{A}^T$.

The decomposition:

$$\mathbf{X}_{n\times p} = \mathbf{G}_{n\times r}\mathbf{H}_{r\times p}\,. \tag{6}$$

is called the GH decomposition of the data table $\mathbf{X}$.

Let us denote the row vectors of $\mathbf{G}$ by $\mathbf{g}_1, \ldots, \mathbf{g}_n$; similarly let us denote the column vectors of $\mathbf{H}$ by $\mathbf{h}_1, \ldots, \mathbf{h}_p$. Using these denotations we can write down formula (6) in a slightly different form showing explicitly the row vectors of $\mathbf{G}$ and the column vectors of $\mathbf{H}$:

$$\mathbf{X}_{n\times p} = \begin{bmatrix} g_{11} & g_{12} & \ldots & g_{1r} \\ \vdots & \vdots & \vdots & \vdots \\ g_{n1} & g_{n2} & \ldots & g_{nr} \end{bmatrix} \begin{bmatrix} h_{11} & h_{12} & \ldots & h_{1p} \\ h_{21} & h_{22} & \ldots & h_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ h_{r1} & h_{r2} & \ldots & h_{rp} \end{bmatrix} = \begin{bmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_n \end{bmatrix} [\mathbf{h}_1, \mathbf{h}_2, \ldots, \mathbf{h}_p]\,. \tag{7}$$

Looking at this formula it can be seen that each element x_{ij} of the data matrix $\mathbf{X}$ can be obtained by multiplying the i-th row of the matrix $\mathbf{G}$ by the j-th column of the matrix $\mathbf{H}$; in other words, it can be obtained as the scalar product (inner product) of the vectors $\mathbf{g}_i$ (row vector) and $\mathbf{h}_j$ (column vector):

$$x_{ij} = \mathbf{g}_i\,\mathbf{h}_j\,. \tag{8}$$

This is true for each element x_{ij} of $\mathbf{X}$. Thus, with known vectors $\mathbf{g}_1, \ldots, \mathbf{g}_n$ and $\mathbf{h}_1, \ldots, \mathbf{h}_p$ we are able to reproduce – by use of them – the whole data matrix $\mathbf{X}$. This statement has enormous practical implication and we will use it later.

Beside of distinguishing the rows $\mathbf{g}_1, \ldots, \mathbf{g}_n$ of $\mathbf{G}$ we will use also sometimes the column vectors of $\mathbf{G}$ which we will denote as $\mathtt{G}_1, \ldots, \mathtt{G}_r$.

Thus, according to the circumstances, the matrix $\mathbf{G}$ will be viewed as composed either from

the row-vectors $\mathbf{g}_1, \ldots, \mathbf{g}_n$, either from the column-vectors $\mathtt{G}_1, \ldots, \mathtt{G}_r$:

$$\mathbf{G} = \begin{bmatrix} \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_n \end{bmatrix} = [\mathtt{G}_1, \ \ldots \ , \mathtt{G}_r] \, . \tag{9}$$

Similarly, $\mathbf{H}$ might be viewed as composed either from column vectors $\mathbf{h}_j, \ (j = 1, \ldots, p)$ either from row vectors $\mathtt{H}_i \ (i = 1, \ldots, r)$:

$$\mathbf{H} = [\mathbf{h}_1, \ \ldots \ , \mathbf{h}_p] \ = \begin{bmatrix} \mathtt{H}_1 \\ \vdots \\ \mathtt{H}_r \end{bmatrix} . \tag{10}$$

Let us look now at the similarities of the SVD and GH decomposition of a matrix $\mathbf{X}$. The respective formulae are:

$$\mathbf{X} = \sum_{i=1}^{r} \lambda_i \mathbf{u}_i \mathbf{a}_i^T \ = \sum_{i=1}^{r} \mathbf{u}_i (\lambda_i \mathbf{a}_i)^T \ = \sum_{i=1}^{r} \mathtt{G}_i \mathtt{H}_i \, . \tag{11}$$

It follows from the above, that all the considerations on approximating the matrix $\mathbf{X}$ (say, of rank r) by matrices of lower rank – enounced above for the SVD decomposition – are valid also for the GH decomposition. Thus, when using the Euclidean norm, the best approximation of $\mathbf{X}$ by a matrix $\mathbf{X}^{(h)}$ of rank $h \leq r$ is obtained when taking the first h columns of $\mathbf{G}$ and the first h rows from $\mathbf{H}$, which means that the approximating matrix $\mathbf{X}^{(h)}$ has the structure:

$$\mathbf{X}^{(h)} = \sum_{i=1}^{h} \mathtt{G}_i \mathtt{H}_i \, .$$

The vectors $\mathbf{g}_i$ and $\mathbf{h}_j$ appearing in the GH decomposition bear some properties which have strongly appealing practical implications. We consider these properties in next subsection.

2.4 Three important properties of vectors $\mathbf{g}_i$ and $\mathbf{h}_j$ obtained in the GH decomposition

Let us remind – what was stated already at the begin of Section 3 – that the data matrix $\mathbf{X}$ we deal with, has columns centered to zero.

`Property 1.` The value of the element x_{ij} is proportional to the length of the projection of the vector $\mathbf{g}_i$ onto the vector $\mathbf{h}_j$.

Proof. Let y_{ij} denote the vector obtained as the projection of $\mathbf{g}_i$ onto the vector $\mathbf{h}_j$. The length of this projection equals (see figure 1) : `Fig 1`

$$y_{ij} = \|\mathbf{g}_i\| \cos(\alpha_{ij}), \tag{12}$$

with α_{ij} being the angle between the vectors $\mathbf{g}_i$ and $\mathbf{h}_j$.

Formerly we have shown in formula (8) that every element x_{ij} of the data matrix $\mathbf{X}$ can be reproduced as the scalar product of the vectors $\mathbf{g}_i$ and $\mathbf{h}_j$.

Let $\|\mathbf{g}_i\|$ and $\|\mathbf{h}_j\|$ denote the lengths, i.e. the Euclidean norms, of the vectors $\mathbf{g}_i$ and $\mathbf{h}_j$ respectively.

Taking into account the general definition of a scalar product and also formula (12) we obtain:

$$x_{ij} = \mathbf{g}_i \, \mathbf{h}_j = \|\mathbf{g}_i\| \, \|\mathbf{h}_j\| \, \cos(\alpha_{ij}) = \|\mathbf{h}_j\| \, y_{ij} \, . \tag{13}$$

So it is proved that the value of the j-th variable observed in the i-th individual is proportional – with a proportionality factor $\|\mathbf{h}_i\|$ – to the value y_{ij} (remind, y_{ij} is the length of the projection of $\mathbf{g}_i$, the vector representing the i-th individual, onto $\mathbf{h}_j$, the vector representing the j-th variable). If $\|h_i\|$ were equal 1, then it would be exactly $x_{ij} = y_{ij}$.

Projecting the vectors $\mathbf{g}_1, \ldots, \mathbf{g}_n$ – representing subsequent individuals numbered $1, \ldots, n$ – onto the same vector $\mathbf{h}_j$ we obtain the projection values $y_{1j}, \ldots \, y_{nj}$ exhibiting the magnitude of the j-th variable as observed for these individuals.

`Property 2.` The matrix $\mathbf{H}\mathbf{H}^T$ reproduces – with an accuracy up to a constant that equals $\frac{1}{n-1}$ – the sample covariance matrix $\mathbf{S}$ evaluated from the data table $\mathbf{X}$.

Proof. This is easy to show by substitution into the formula for $\mathbf{S}$ instead of $\mathbf{X}$ its equivalent representation $\mathbf{GH}$:

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X} = \frac{1}{n-1}(\mathbf{GH})^T(\mathbf{GH}) = \frac{1}{n-1}\mathbf{H}^T\mathbf{H}.$$

This ends the proof of Property 2.

In calculating $\mathbf{S}$ above we have used the fact that the columns of the data table $\mathbf{X}$ were centered to zero; therefore the covariances could be derived simply as cross-products.

It follows from the derived equality, that the (i,j)-th element of $\mathbf{S}$ can be equivalently calculated as

$$s_{ij} = \frac{1}{n-1}\mathbf{h}_i^T\mathbf{h}_j\,. \tag{14}$$

Taking into account the definition of a scalar product, the element s_{ij} can be written down as

$$s_{ij} = \frac{1}{n-1}\|\mathbf{h}_i\|\ \|\mathbf{h}_j\|\ \cos(\alpha_{ij})\,, \tag{15}$$

with α_{ij} denoting the angle between the vectors $\mathbf{h}_i$ and $\mathbf{h}_j$.

In the special case of (15) when $i = j$ we obtain, that

$$s_{ii} = s_i^2 = \frac{1}{n-1}\|\mathbf{h}_i\|^2, \quad i = 1,\ldots,p\,, \tag{16}$$

which means that the length of each of the vectors $\mathbf{h}_j,\ (j = 1,\ldots,p)$ is proportional – with coefficient of proportionality equal to $\frac{1}{n-1}$, the same for all variables – to the variances of the p variables under consideration.

`Property 3.` Using the GH decomposition of the data matrix $\ddot{\mathbf{X}} = (\ddot{x}_{ij})$ with elements standardized according the formula

$$\ddot{x}_{ij} = \frac{x_{ij} - \overline{x}_{.j}}{s_j\sqrt{n-1}}\,, \tag{17}$$

we obtain a matrix $\mathbf{H}$ such that the matrix $\mathbf{H}\mathbf{H}^T$ reproduces *exactly* the correlation matrix $\mathbf{R}$ of the considered variables. The symbol $\overline{x}_{.j}$ denotes here the arithmetic mean of the j-th column of $\mathbf{X}$.

Proof. Let $\mathbf{R}$ be the correlation matrix of the variables $X_1,\ldots,X_p$. Let $\mathbf{P} = (p_{jk})$ denote the cross-product matrix $\ddot{\mathbf{X}}^T\ddot{\mathbf{X}}$, i.e. $\mathbf{P} = \ddot{\mathbf{X}}^T\ddot{\mathbf{X}}$. Then p_{jk}, the (j,k)-th element of $\mathbf{P}$

is obtained as

$$p_{jk} = \frac{\sum_{i=1}^{n}(x_{ij} - \bar{x}_{.j})(x_{ik} - \bar{x}_{.k})}{s_j s_k\,(n-1)} = r_{jk},$$

which is obviously the sample correlation coefficient between the j-th and the k-th variable.

Hence $\mathbf{P} \equiv \mathbf{R}$.

In the premise of property 3 we have assumed that the GH decomposition was made with the matrix $\ddot{\mathbf{X}}$, what means that

$$\ddot{\mathbf{X}} = \mathbf{GH}.$$

Substituting the above representation into the product $\ddot{\mathbf{X}}^T\ddot{\mathbf{X}}$ we obtain:

$$\mathbf{R} = (\ddot{\mathbf{X}}^T\ddot{\mathbf{X}}) = \mathbf{H}^T\mathbf{G}^T\mathbf{GH} = \mathbf{H}^T\mathbf{H}. \tag{18}$$

From the derived decomposition of $\mathbf{R}$ given above as formula (18) one can deduce two important properties of the vectors $\mathbf{h}_j, \mathbf{h}_k$ yielding r_{jk}, $j, k = 1, \ldots, p$.

1. Taking $j = k$ we obtain

   $$r_{jj} = \mathbf{h}_j^T\mathbf{h}_j = \|\mathbf{h}_j\|^2 = 1,$$

   which means that the vectors $\mathbf{h}_j$ are of unit length.

2. Taking $j \neq k$ we obtain

   $$r_{jk} = \mathbf{h}_j^T\mathbf{h}_k = \|\mathbf{h}_j\|\,\|\mathbf{h}_k\|\cos(\alpha_{jk}) = \cos(\alpha_{jk}),$$

   which means that the cosine of the angle between the vectors $\mathbf{h}_j$ and $\mathbf{h}_k$ is direct equal to the correlation coefficient r_{jk} between the j-th and the k-th variable.

3 Drawing the traditional and the augmented biplot

Both the traditional and the augmented biplots are drawn using values obtained as the GH decomposition of the considered data matrix $\mathbf{X}$. In this section we remind firstly principles of drawing the traditional biplot; next we proceed to our proposal of drawing the augmented biplot.

3.1 Drawing the traditional biplot

When drawing the traditional biplot we use only the information contained in the first two columns of the matrix $\mathbf{G}$ and in the first two rows of the matrix $\mathbf{H}$, i.e. we use the approximation:

$$\mathbf{X}^{(2)}_{n\times p} = \begin{bmatrix} g_{11} & g_{12} \\ \vdots & \vdots \\ g_{n1} & g_{n2} \end{bmatrix} \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1p} \\ h_{21} & h_{22} & \dots & h_{2p} \end{bmatrix} = \begin{bmatrix} \mathbf{g}^{(2)}_1 \\ \vdots \\ \mathbf{g}^{(2)}_n \end{bmatrix} \begin{bmatrix} \mathbf{h}^{(2)}_1 & \mathbf{h}^{(2)}_2 & \dots & \mathbf{h}^{(2)}_p \end{bmatrix} \qquad (19)$$

That means using by us the best approximation of $\mathbf{X}$ by a matrix of rank 2 (see formula (11)).

Obviously each row $[g_{k1}, g_{k2}]$ $(k = 1, \dots, n)$ of the matrix shown above can be considered as containing coordinates of a two-dimensional point and as such it can be drawn as a point in a plane. In such way we can obtain representations of the individuals.

Similarly, each column vector $\mathbf{h}^{(2)}_j$ appearing in (19) can be considered as containing coordinates of a two-dimensional point and as such can be also drawn in the plane. This time we obtain representations of the variables. Usually these representations are graphed in the form of vectors originating from the beginning of the coordinate system of the plane.

The error of the approximation given by (19) equals $\lambda_3 + \ldots + \lambda_r$ (see formula (4)), and the goodness of fit amounts $(\lambda_1 + \lambda_2)/(\sum_{i=1}^{r} \lambda_i)$ — see formula (5).

If the error of the approximation is small, then the representation of points-individuals and points-variables is a fair one, and we can relay on it. In such case we can utilize the 3 properties shown in subsection 2.4 above; in particular we can judge the relative magnitudes of the variables as realized for various individuals; also, we can "see" the correlations amongst the considered variables. This is great! Having multivariate data, i.e. points located in multivariate space, we can observe directly their interrelations.

However, all this is valid only then, when the representation in the biplot is a fair one. In the case, when that is not true, we may get quite a false visualization in the biplot – and then, of course, all the perceived interdependencies might not be true too.

When using the traditional biplot we are left at this point and nothing more can really be done: since the biplot is untrustworthy, it can not be taken serious, and truly speaking, it should be discarded.

3.2 Proposal of drawing the augmented biplot

Let us look at the representation of variables as given by the vectors $\mathbf{h}_1^{(2)}, \ldots, \mathbf{h}_p^{(2)}$ appearing in (19) and the vectors $\mathbf{h}_1, \ldots, \mathbf{h}_p$ appearing in (7).

For each vector $\mathbf{h}_j$ $(j = 1, \ldots, p)$ representing the j-th variable we have the following situation:

Length of the two-dimensional representation of $\mathbf{h}_j^{(2)}$: $(\sum_{i=1}^{2} h_{ij}^2)^{1/2}$;

True length of $\mathbf{h}_j$ as given by $\mathbf{h}_j^r$: $(\sum_{i=1}^{r} h_{ij}^2)^{1/2}$.

It is possible to graph in one plane the l e n g t h s both of $\mathbf{h}_j^{(2)}$ and $\mathbf{h}_j^{(r)}$.

`Our proposal` is the following: Draw firstly the traditional biplot, y including the vectors $\mathbf{h}_j^{(2)},\ j = 1, \ldots, p$.

Next prolong – say, by a dotted line – each vector $\mathbf{h}_j^{(2)}$ from its length given by the first two rows of the matrix $\mathbf{H}$ – to the length of the full vector $\mathbf{h}_j^{(r)}$.

We call a biplot exhibiting such a double representation of both the censored vector $\mathbf{h}_j^{(2)}$ and of the full length vector $\mathbf{h}_j = \mathbf{h}_j^{(r)}$ — the *augmented biplot*.

After drawing the biplot in the proposed way we can grasp at once – when looking at the graph – whether the representation is a good one. One can also see immediately, which variables are represented in a poor way.

`Algorithm for construction of the augmented biplot.`

Let $l_j = (h_{1j}^2 + h_{2j}^2)^{1/2}$ and $L_j = (\sum_{i=1}^{r} h_{ij}^2)^{1/2}$ denote the lengths of the vectors $\mathbf{h}_j^{(2)}$ and $\mathbf{h}_j^{(r)}$ as given by (19) and (7).

Let $(x_j, y_j) = (h_{1j}^{(2)}, h_{2j}^{(2)})$ be the coordinates of the j-th endpoint of the vector $\mathbf{h}_j^{(2)}$ representing in the biplot the j-th variable.

Then the coordinates (X_j, Y_j) representing the "full" vector-variable N^o j are obtained as:

$$X_j = x_j \frac{L_j}{l_j}, \quad Y_j = y_j \frac{L_j}{l_j}.$$

Proof. The proof follows immediately from relations visualized in Figure 2. Fig 2

From similarity of the triangles OAP and OBQ we have:

$$\frac{x_j}{l_j} = \frac{X_j}{L_j}, \quad \frac{y_j}{l_j} = \frac{Y_j}{L_j},$$

wherefrom we obtain the sought values X_j and Y_j of the augmented representation.

3.3 Reproduction of variances by subsequent approximations

Denoting by $S_{ii} = (n-1)s_{ii}$ the total variation of the i-th variable, evaluated as the total adjusted sum of squares, or as the (i,i)-th element of the cross-product matrix $\mathbf{X}^T\mathbf{X}$, with $\mathbf{X}$ centered to zero, we have (see (16)):

$$S_{ii} = \mathbf{h}_i^T \mathbf{h}_i = h_{1r}^2 + \ldots + h_{ri}^2.$$

This formula tells us that S_{ii}, the variation of the i-th variable can be reproduced by taking the squared components of $\mathbf{h}_i$.

To see the importance of the subsequent components of the vectors $\mathbf{h}_1, \ldots, \mathbf{h}_p$ in reproducing the considered variables $V_1, \ldots, V_p$ (whose measurements have been recorded in subsequent columns of the data matrix $\mathbf{X}$) we display the cumulative increments h_{1i}^2, $h_{1i}^2 + h_{2i}^2, \ldots,$ $h_{1i}^2 + \ldots + h_{ri}^2$ – for each of the variables V_i – using the following layout

$m \Downarrow$	V_1	V_2	$\ldots$	V_p
1	h_{11}^2	h_{12}^2	$\ldots$	h_{1p}^2
2	$\sum_{k=1}^{2} h_{k1}^2$	$\sum_{k=1}^{2} h_{k2}^2$	$\ldots$	$\sum_{k=1}^{2} h_{kp}^2$
3	$\sum_{k=1}^{3} h_{k1}^2$	$\sum_{k=1}^{3} h_{k2}^2$	$\ldots$	$\sum_{k=1}^{3} h_{kp}^2$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
r	S_{11}	S_{22}	$\ldots$	S_{pp}

From such layout it can be seen at once, how large amounts of the variation S_{ii} are reproduced by subsequent components of the vectors $\mathbf{h}_1, \ldots, \mathbf{h}_p$. Usually, the amounts showing the cumulative increments are adjusted to show the fractions or percentages of S_{ii}.

In the case when the calculations are done with data standardized according to formula (17), i.e. when the eigenvalues and the eigenvectors for the SVD decomposition are obtained from cross–products of data standardized according to (17) (what means that the cross–products yielded direct the correlation matrix), we have

$$S_{ii} \equiv 1 \quad (i = 1, \ldots, p),$$

and the rows of the proposed layout show at once, how large fraction of the total (unit) variation has been accounted for when considering m compounds of the approximation.

4 Some real examples of applications of the augmented biplot

It is much advised to carry out the calculations with data standardized to have the same variances – otherwise the results would strongly depend on units, in which the measurements were expressed.

In all the examples presented in this chapter the respective data were firstly standardized according to formula (17), what implies that the eigenvalues and eigenvectors needed for the GH decomposition can be seen as those computed from a correlation matrix. This implies in turn, that

$$trace(\ddot{\mathbf{X}}^T \ddot{\mathbf{X}}) = \sum_{i=1}^{p} \lambda_i = p,$$

and the derived vectors $\mathbf{h}_1, \ldots, \mathbf{h}_p$ are of unit length.

Table 1: Yields of 8 oat varieties (columns A-H) gathered in 13 locations

N^o	Location	A=V1	B=V2	C=V3	D=V4	E=V5	F=V6	G=V7	H=V8
1.	Wrocikowo	39.4	36.9	37.4	40.8	43.0	40.9	38.1	37.5
2.	Garbno	41.6	40.3	40.2	46.9	42.9	42.4	42.7	40.5
3.	Nikutowo	44.0	43.7	45.6	45.3	48.2	45.0	44.2	42.2
4.	Rychliki	53.4	55.9	54.7	55.3	57.1	58.6	57.2	55.3
5.	Chełchy	45.5	44.2	41.1	45.4	46.5	46.4	45.0	39.6
6.	Cicibór	37.6	38.8	35.7	35.8	38.6	38.9	39.1	38.8
7.	Uhnin	48.6	49.1	48.4	49.6	50.8	49.4	52.1	47.5
8.	Krzyżewo	54.5	53.5	52.7	56.0	57.3	56.4	54.7	52.0
9.	Marianowo	48.3	46.7	46.6	49.1	51.0	50.0	50.5	46.6
10.	Seroczyn	50.9	51.9	50.2	51.7	53.0	55.1	54.7	51.8
11.	Bezek	46.4	54.4	43.9	50.4	48.5	51.5	43.4	48.2
12.	Czesławice	45.2	44.1	42.9	46.4	45.5	48.1	45.1	46.0
13.	Jarosław	32.0	28.7	26.7	30.1	30.6	34.2	32.2	31.5

4.1 Similarities among oat varieties

The data contain yields of 8 varieties of oat as gathered in 13 experimental stations located in the north-eastern region of Poland. The respective yields were registered by the Cultivar Testing Center in Słupia Wielka. We have taken the data from the paper of Dmowski et al. [8], who analysed the data under different aspects. The now to be considered varieties are: A - Markus, B - Perona, C - Borek, D - Rumak, E - Dragon, F - Boruta, G - Ułan, H - Płatek.

The location of the experimental stations was as follows: 1 - Wrocikowo, 2 - Garbno, 3 - Nikutowo, 4 - Rychliki, 5 - Chełchy, 6 - Cicibór, 7 - Uhnin, 8 - Krzyżewo, 9 - Marianowo, 10 - Seroczyn, 11 - Bezek, 12 - Czesławice, 13 - Jarosławiec. In the following we will refer to the experimental stations in these locations simply as "locations".

The gathered data are shown in Table 1. The rows in this table correspond to the locations; the columns – to the varieties.

We assume the rows in Table 1 to be "subjects" or "individuals", and the columns A $\div$ H to be "variables" characterizing the subjects (each location is characterized by yields obtained

when growing the indicated oat varieties) – and so we are put in the classical situation considered in sections 1–4. Speaking precisely we have to deal now with a data matrix $\mathbf{X}$ of size 13×8.

The aim of our analysis is to present graphically in a plot the interdependencies between the yields (amount of oat grain per ha) obtained for different varieties. We would like to present in the same plot also the locations and show which of them give high, mediocre or low yield; none the less the emphasis in our analysis will be put on the interrelations between the varieties, which in the following will be considered as "variables" characterizing each of the locations.

For the data table $\mathbf{X}$ defined in such way we evaluate the correlation matrix $\mathbf{R}$ and next we make the GH decomposition according to formula (11). We obtain then the eigenvalues λ_i $(i = 1, \ldots, 8)$ shown in Table 2. Since the eigenvalues were computed from the correlation matrix, they sum up to 8.00, which is equal to p, the number of the considered variables.

In the same table (i.e. in Table 2) we show also the eigenvectors corresponding to the three highest eigenvalues. All the eigenvectors are of unit norm – which follows from the computational procedure.

Next we computed the rank 2 approximation of $\mathbf{X}$ – as given by formula (19) and so we obtained the basis for drawing the classical biplot.

The corresponding biplot is shown in Figure 3. In this figure the "variables", i.e. the varieties A ÷ H are shown as vectors $V1, \ldots, V8$, while the locations are shown as dots labelled by integers 1 through 13. Fig 3

Looking at the eigenvalues shown in Table 2 one can see that the first two of them contain 97.53% of the total inertia (trace) of the matrix $\mathbf{R}$, from which they were calculated. Hence we deduce that the representation shown in Figure 3 is a good one.

Because the evaluations were done on the basis of a correlation matrix – the vectors representing the varieties (assumed as variables in our data table) should be of unit length. Since the representation in the plane is a good one, the vectors appearing in the biplot are very near to unit length. Hence we can relay on the interdependencies shown in the biplot.

Table 2: The eigenvalues λ_i and the eigenvectors $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ for the data "Yields of 8 oat varieties", and reproduction of variances of the eight oat varieties when using approximation of rank m=1, 2 and 3

Nr	λ_i	$\sum(\lambda_i)$	% trace	$\mathbf{a_1}$	$\mathbf{a}_2$	$\mathbf{a}_3$
1	7.6554	7.6554	95.69	0.3586	-0.1653	0.1059
2	0.1473	7.8027	97.53	0.3479	0.6368	0.0546
3	0.0956	7.8983	98.73	0.3564	-0.2849	0.0601
4	0.0384	7.9366	99.21	0.3519	0.0484	0.5968
5	0.0337	7.9703	99.63	0.3559	-0.2446	0.3314
6	0.0226	7.9930	99.91	0.3567	0.2342	-0.1696
7	0.0043	7.9972	99.97	0.3487	-0.5237	-0.4756
8	0.0028	8.0000	100.00	0.3522	0.3077	-0.5111

Reproduction of variances of yield of eight varieties								
$m \Downarrow$	V1	V2	V3	V4	V5	V6	V7	V8
1	0.98	0.93	0.97	0.95	0.97	0.97	0.93	0.95
2	0.99	0.99	0.98	0.95	0.98	0.98	0.97	0.96
3	0.99	0.99	0.98	0.98	0.99	0.98	0.99	0.99

In particular we see in the plot that the angles between the vectors-varieties are all acute, which means that the yields of all the varieties are positively correlated. One can also see that the varieties $V6$ and $V8$ give very similar yields (i.e. the yields of these two varieties are highly correlated). The same can be said on the varieties $V1, V3$ i $V5$.

It is also interesting to look at the reproduction of the variances of subsequent varieties, which are considered in this subsections as "variables". The reproduction is shown in the lower part of Table 2. The variances of the oat varieties are reproduced by two components of the approximation at least in 95.69% (the variety V4), and for two of them in 97.53% (the varieties V1, V2). Taking an approximation of rank 3 we obtain a reproduction of the variance of five varieties in 98.73%, which is really very good.

We can conclude therefore: instead of dealing with a p-dimensional space it is sufficient to deal with a 3-dimensional space, in which all the essential information can be recovered.

Now let us look at the points representing the locations. One can see immediately that

the location numbered as "11" (Bezek) is truly an outlier in comparison to other locations. The outstanding position of this point is caused by a high value of its ordinate, which is given by $\mathtt{G}_2$, the second column of the matrix $\mathbf{G}$ from formula (19) ($\mathtt{G}_2 = \mathbf{X}\mathbf{a}_2$). In turn, looking at the elements of $\mathbf{a}_2$ shown in Table 2 one can see that its values exhibit very distinctly the difference between the yield of the variety "B" (denoted in the plot as V2) and the variety "G" (denoted in the plot as V7). Thus we deduce that the outlyingness of the location Bezek in the biplot is due to the fact, that it shows other interrelation between the varieties B and G – in comparison to the interrelations stated in other locations. This supposition can be checked directly by looking into the data table: in fact inspecting Table 1 one can see, that generally the difference in yield for the varieties B and G is negative; however in Bezek this difference is positive.

Looking at the biplot shown in Figure 3 one can also state that generally the locations N^o 4, 8, and 10 give relatively high yield; while the location N^o 13 provides a low yield. This can be easily seen when projecting the respective points-locations onto the vectors-variables representing in the biplot the yields of subsequent varieties (the justification of this action relies on formula (13)): the projections of the locations N^o 4, 8, 10 are utmost positive, while the projection of the location N^o 13 is utmost negative.

We could summarize the analysis of Table 1 as follows: The yields of all varieties are positively correlated. Location N^o 11 (Bezek) exhibits other type of interdependencies as the remaining locations. The yield in location N^o 13 (Jarosławiec) is extremely low.

All these conclusions could be obtained when looking at the data table – however, with a larger amount of data such regularities might be difficult to discover. Using the biplot technique we are able to discover the regularities of that kind immediately.

Table 3: Eigenvalues λ_i obtained for transposed data table 1 first (a) from entire data set and next (b) after omitting location N^o. 11. Below: Reproduction of variances of subsequent locations when using approximations of rank m=1, 2 and 3

	a) Analysing L1÷L13			b) L1÷ L10, L12, L13		
Nr	λ_i	$\sum\lambda_i$	% trace	λ_i	$\sum\lambda_i$	% trace
1	6.5586	6.5586	50.45	6.5296	6.5296	54.41
2	2.5776	9.1362	70.28	2.5415	9.0712	75.59
3	1.5443	10.6805	82.16	1.3441	10.4153	86.79
4	1.0163	11.6969	89.98	0.6846	11.0999	92.50
5	0.6141	12.3109	94.70	0.5228	11.6227	96.86
6	0.5157	12.8267	98.67	0.2351	11.8578	98.81
7	0.1733	13.0000	100.00	0.1422	12.0000	100.00
8	0.0000	13.0000	100.00	-0.0000	12.0000	100.00

Reproduction of variances of yield in 13 locations													
$m \Downarrow$	L1	L2	L3	L4	L5	L6	L7	L8	L9	L10	L11	L12	L13
1	.65	.36	.29	.55	.71	.14	.51	.86	.90	.64	.03	.48	.43
2	.81	.57	.77	.69	.76	.77	.52	.97	.92	.89	.10	.62	.75
3	.87	.68	.80	.73	.76	.83	.87	.98	.97	.95	.54	.90	.78

4.2 Similarities among locations of cultivation of oat varieties.

Now we will look at the table dealt with in the previous section in a different way: we will consider a transposed table $\breve{\mathbf{X}}$, whose rows represent the varieties and columns – the locations. After standardizing this transposed table by use of formula (17) we evaluate the eigenvalues and the eigenvectors of the cross-product matrix $\breve{\mathbf{X}}^T\breve{\mathbf{X}}$ – which is equal to the correlation matrix of the standardized data. The obtained eigenvalues, their cumulated sums and percents of reproduction of the total inertia are shown in Table 3. The corresponding (augmented) biplot is shown in Figure 4. Fig 4

Looking at the eigenvalues shown in Table 3 we state that the first two eigenvalues account only for 70.28% of the total inertia. We conclude therefore that the representation in the shown biplot is not very satisfactory. Adding the third dimension to the biplot we would get a better representation (82.16%), although still not satisfactory. Taking into account formula (18) we

are able to state which of the considered variables have in the biplot a fair representation, and which have not.

Table 3 in its lower part shows the goodness of representation of subsequent 13 vectors-variables (which are now interpreted as locations) when taking into account the 1-, 2- or 3-dimensional approximation. Considering the two-dimensional approximation, as used in the biplot, we state that the worst approximation is obtained for location N^o 11 (reproduction only in 10 %), N^o 2 and 7 (reproduction 57 % and 52 %) and location N^o 12 (reproduction in 62 %). Looking at the biplot shown in Fig. 4 one can see that the vectors representing the mentioned locations are much shorter as the remaining ones.

In previous subsection – when analyzing the untransposed data table $\mathbf{X}_{13\times 8}$ – we have observed that the location N^o 11 is distinctly different, i.e. it occupies an outlying position. Because the sample size of the considered data is small, this location could have a larger impact on the cross-product matrix, from which the eigenvalues and eigenvectors were evaluated. Such extreme outlying points might be very influential in the computations, in particular they could distort the directions of the eigenvectors constructed on the basis of the cross-product matrix $\mathbf{X}^T\mathbf{X}$ (see the algorithm for constructing the SVD decomposition, subsection 2.1).

To find out the impact of location N^o 11 on the obtained graphical representation we have removed this location from the data matrix and next we have repeated our analysis with the reduced data set. We got similar results as previously. In Table 3 in part (b) we show the eigenvalues obtained when considering the reduced data set. One can see that the values and percents of exhaustion obtained from the reduced data set (from 12 locations considered as "individuals") are quite similar to those obtained from the entire data set represented by 13 locations.

Table 4: Eigenvalues λ_i and eigenvectors $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ for the data "Vegetable Production of Farms", and reproduction of variances of the 6 characteristics when using approximation of rank m=1, 2 and 3

Nr	λ_i	$\sum(\lambda_i)$	% trace	$\mathbf{a_1}$	$\mathbf{a}_2$	$\mathbf{a}_3$
1	4.7574	4.7574	79.29	0.4219	-0.2757	0.4755
2	0.6178	5.3752	89.59	0.3450	0.7930	-0.1144
3	0.2808	5.6560	94.27	0.4099	0.3322	0.4087
4	0.2026	5.8586	97.64	0.4067	-0.0966	-0.7691
5	0.0908	5.9494	99.16	0.4222	-0.3072	-0.0463
6	0.0506	6.0000	100.00	0.4373	-0.2847	0.0084

Reproduction of variances of 6 farm characteristics						
$m \Downarrow$	V1	V2	V3	V4	V5	V6
1	0.85	0.57	0.80	0.79	0.85	0.91
2	0.89	0.95	0.87	0.79	0.91	0.96
3	0.96	0.96	0.91	0.96	0.91	0.96

4.3 Interrelations between characteristics of farms

Now we will consider the data published by Zeliaś [1] and analyzed later by Bartkowiak and Liśkiewicz [9]. The data table contains values of 6 variables gathered for 31 representative farms located in mid-west region of Poland. The considered variables are:

V1 – area of land (in ha), V2 – investment of human work (in working days), V3 – the livestock (number of big animals), V4 – inventory stock, V5 – expenditure for fertilizers. V6 – vegetable production of the farm.

The above mentioned authors investigated the regression of V6 (taken as the predicted variable) from V1÷V5 taken as predictors. Our aim now is to present graphically the interrelations amongst the considered variables.

Proceeding along the lines as shown in previous sections we have evaluated from the correlation matrix the eigenvalues and eigenvectors (shown in Table 4), the reproduction of the variables (shown in the same table), and next we have constructed the biplot shown in Figure 5. Fig 5

The goodness of the reproduction of the total inertia is also shown in Table 4. One can see, that for these data the representation of the variables in the biplot is a fair one: the first two components reproduce 89.59% of total variability (inertia) of the considered variables. Adding the third dimension we obtain a reproduction in 94.27%. The worst representation is met for variable V4 (only 79%), which is graphed in the plot by a shorter vector. Looking at the biplot one can see at once that the highest correlation with V6 is exhibited by V1 i V5; the lowest – by V2 i V3.

It might be interesting for the reader to know that after computing the multiple regression of V6 from V1,...,V5 we obtained the following Studentized regression coefficients: t[1]=2.94, t[2]=–0.10, t[3]=0.21, t[4]=2.24, t[5]=3.54.

It follows from these values, that the variables V2 and V3 are in the evaluated regression equation statistically nonsignificant. The squared multiple correlation coefficient for the considered regression amounts: RR=0.9296.

4.4 Similarities among blood pressure measurements

Another example illustrating the usefulness of data representation by a biplot will be given by dealing with the data considered in [5]. The authors have investigated the problem whether the measurements of arterial blood pressure (systolic and diastolic) – when recorded by use of the automated device AVIONICS 1900 and the traditional mercurial sphygmomanometer – yield the same values. To clarify this problem the mentioned authors have used data from an trial in which – for about 120 men and 120 women – the measurements were recorded by both methods, in a randomized sequence, and with an appropriate washout period. The two accounted methods were: automatic (i.e. using the AVIONICS 1900 device) and traditional (i.e. using the traditional mercurial sphygmomanometer). The measurements of the systolic and diastolic blood pressure were recorded for each patient in 3 positions: standing, sitting and lying. From this trial we take now only the measurements of the systolic blood pressure

Table 5: Eigenvalues λ_i and eigenvectors $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$ evaluated for the data "Systolic Blood Pressure for 117 women", and reproduction of variances of the measurements when considering approximations of rank m=1, 2 and 3

Nr	λ_i	$\sum(\lambda_i)$	% trace	$\mathbf{a_1}$	$\mathbf{a}_2$	$\mathbf{a}_3$
1	5.6127	5.6127	93.54	0.4066	-0.4615	0.3988
2	0.2190	5.8317	97.19	0.4118	-0.3048	-0.3780
3	0.0596	5.8913	98.19	0.4073	-0.4466	-0.0796
4	0.0427	5.9340	98.90	0.4072	0.4011	0.6871
5	0.0369	5.9709	99.51	0.4074	0.4495	-0.1951
6	0.0291	6.0000	100.00	0.4092	0.3631	-0.4261

Reproduction of variances of subsequent measurements						
$m \Downarrow$	V1	V2	V3	V4	V5	V6
1	0.93	0.95	0.93	0.93	0.93	0.94
2	0.97	0.97	0.97	0.97	0.98	0.97
3	0.98	0.98	0.98	0.99	0.98	0.98

recorded for n=117 women. The considered variables are denoted as follows:

V1 – method: automatic, position: standing,

V2 – method: automatic, position: sitting,

V3 – method: automatic, position: lying,

V4 – method: traditional, position: standing,

V5 – method: traditional, position: sitting,

V6 – method: traditional, position: lying.

The eigenvalues and the eigenvectors obtained for these data (which were also standardized according to formula (17)) are shown in Table 5. The respective biplot is shown in Figure 6. Fig 6

Looking at the eigenvalues shown in Table 5 one can state that the representation of the data visible in the biplot is a good one. Taking approximation of rank m=2 we obtain a 97.19% reproduction of the total inertia. Also the variances of each of the 6 individual measurements are then reproduced at least in 97%.

When looking at the mutual location of the vectors-variables in the biplot we find there

Table 6: Correlation coefficients between the considered blood pressure measurements

	V1	V2	V3	V4	V5	V6
V 1	1.0000					
V 2	0.9585	1.0000				
V 3	0.9590	0.9601	1.0000			
V 4	0.8968	0.9042	0.8917	1.0000		
V 5	0.8834	0.9153	0.8871	0.9556	1.0000	
V 6	0.8924	0.9225	0.9021	0.9496	0.9595	1.0000
	V1	V2	V3	V4	V5	V6

a characteristic pattern: The vectors-variables representing the "automated" recording are completely separated from those representing the "traditional" method.

Looking at the components of the eigenvectors shown in Table 5 we state also a specific pattern. The first eigenvector has components that are similar in their values – we can conclude, that this vector constructs a new variable (a new value) that depends generally from the magnitude of the object (here: from the level of the blood pressure stated in the measured women). The second eigenvector has the first three elements negative, and the next three positive – one could say, that this vector constructs a new variable which expresses the contrast or the difference between the automated and the traditional method of blood pressure recording.

Another presentation of the same biplot is shown in Fig. 7. Here we have assumed a scale ranging from -1 to +1 for both coordinates. This yields generally the impression that the data structures are more compact, in particular – that the vectors representing the variables are closer. Fig 7

It might be interesting for the reader to know the true correlations (Pearsonian correlation coefficients) between the considered variables. They are given in Table 6. One can see there distinctly that the correlations amongst the first 3 measurements are at least equal to 0.95; so are also the correlations amongst the last 3 measurements. That means, that measurements within a method are really highly correlated. The correlations between measurements belonging

to different methods exhibit values about 0.90 or less. This features can be seen at once when looking at the constructed biplot.

5 Final conclusions and remarks

1. The presented method, i.e. the augmented biplot, permits in many cases to view quite accurately the structure of data clouds located in multivariate space.

2. In many cases the problem for choosing appropriate scale for the biplot presentation may arise. This problem was signalized when dealing with the blood pressure data (subsection 5.4). The problem needs some thorough considerations.

3. For many data structures the accuracy and the informativeness of the presentation would improve when taking a rank 3 approximation of the $\mathbf{GH}$ matrix and presenting the interrelations utilizing a 3-dimensional biplot (remind, **bi**plot denotes a simultaneous representation of two dimensions of the data, meaning here e.g. individuals and variables, or rows and columns; hence – using this convention – a 3 dimensional biplot can not be called a "triplot"). We did not consider here the problems connected with presentations of a three-dimensional biplot.

References

[1] Zeliaś A. (1970). Uwagi o problemie optymalnego wyboru zmiennych objaśniających. {Remarks on the problem of optimal choice of predictors} *Przegląd Statystyczny* 17, 193-210.

[1977]

[2] Lebart L., Morineau A., Tabard N. (1977): *Techniques de la Description Statistique.* Dunod, Paris.

[1978]

[3] Gabriel K.R. (1978): Least squares approximation of matrices by additive and multiplicative models. *J. R. Statist. Soc., Ser. B., 40, 186-196.*

[1982]

[4] Gabriel K.R. (1982): Biplot. In: S. Kotz i N.L. Johnson (Ed.), *Encyclopedia od Statistical Sciences*, V.1, Wiley, New York, 263-271.

[1985]

[5] Bartkowiak A., Ruta R., Włodarczyk W. (1985). Zależności statystyczne między pomiarami ciśnienia tętniczego krwi metodą automatyczną i tradycyjną w różnych pozycjach pomiaru. {Statistical dependencies between measurements of arterial blood pressure performed automatically and in the traditional way in various positions} *Człowiek, Populacja, Środowisko*, z.12., 49-92.

[1986]

[6] Jolliffe I.T. (1986): *Principal Component Analysis.* Springer, New York.

[1988]

[7] Bartkowiak A. (1988). Testing equality of variances for pairwise dependent observations on the example of blood pressure readings. *Modelling, Simulation & Control*, C, AMSE Press, 12,No.3, 21-29.

[1989]

[8] Dmowski K., Błażczak P., Wagner W. (1989): Przydatność analizy dominacji stochastycznej dla wyboru czołowej grupy odmian. Cz. II. Przykład liczbowy. {Usefulness of stochastic dominance analysis for choosing the priority set of variates. Part II. Numerical example.} *Dziewiętnaste Colloquium Metodologiczne z Agrobiometrii.* PAN, 284 - 306.

[1990]

[9] Bartkowiak A., Liśkiewicz T. (1990). A dilemma: Good model fit or stability of coefficients in the model equation. *AMSE Review*, 12, No.4, 45-58.

[10] Gabriel K.R. i Odoroff Ch.L. (1990): Biplots in biomedical research. *Statistics in Medicine* 9, 469-485.

[1991]

[11] Golub G.H., Van Loan Ch.F. (1991): *Matrix Computation.* J. Hopkins University Press, Baltimore and London, 1991.

[12] Watkins D.S. (1991): *Fundamentals of Matrix Computations.* Wiley, New York.

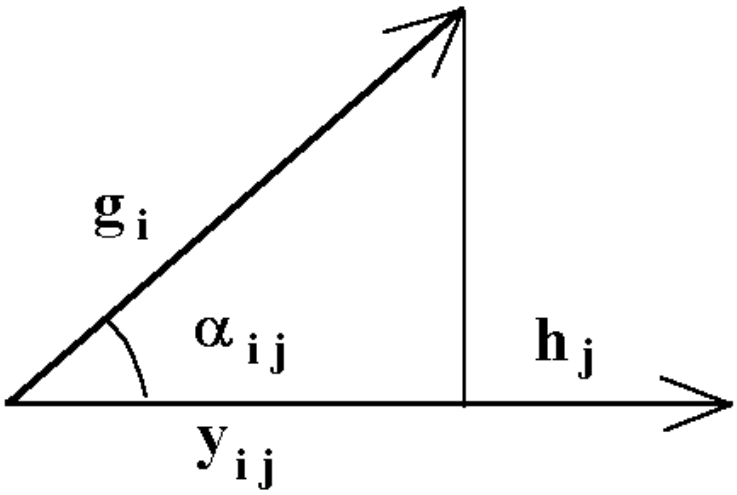

Figure 1: Projection of the vector $\mathbf{g}_i$ onto the vector $\mathbf{h}_j$

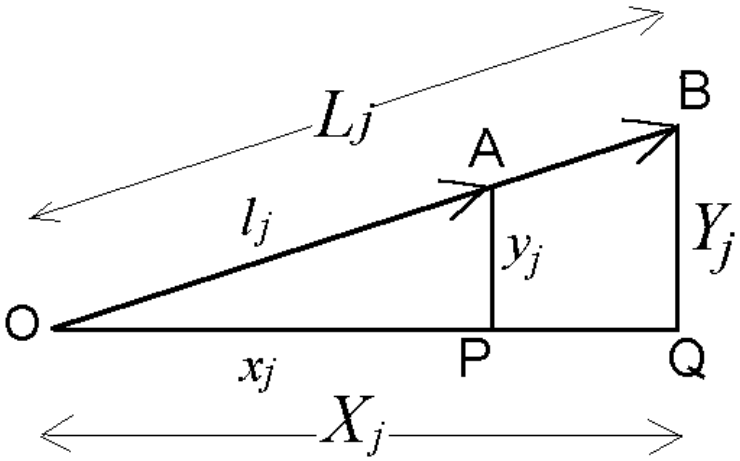

Figure 2: Augmenting the vector OA representing the j-th variable

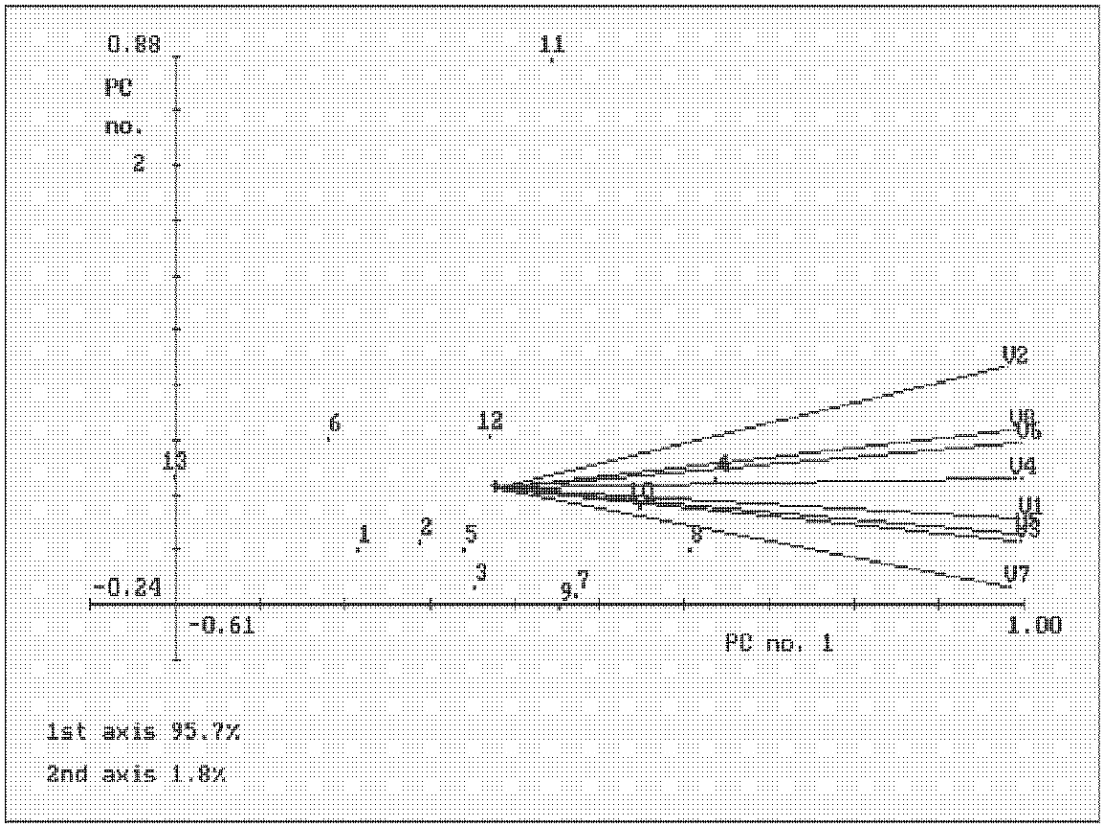

Figure 3: Augmented biplot visualizing proximities between 8 oat varieties (V1÷V8) when considering yields obtained in 13 locations. Yields of all 8 varieties are positively correlated. Note the outstanding position of location N^o 11, which means a different yield correlation structure in this location

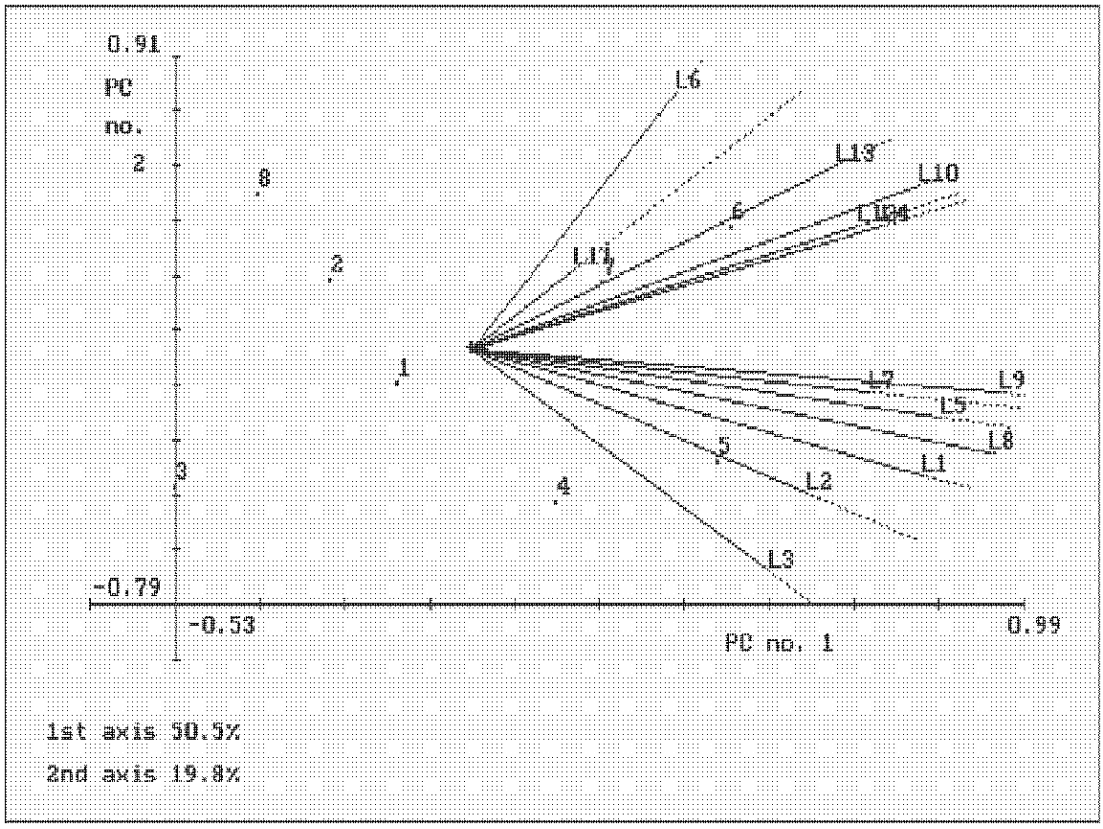

Figure 4: Augmented biplot visualizing proximities between 13 locations L1÷L13 when considering yields of eight oat varieties obtained in these locations. The representation in the biplot is not satisfactory. Note the specially poor representation of locations L11, L7 and L2

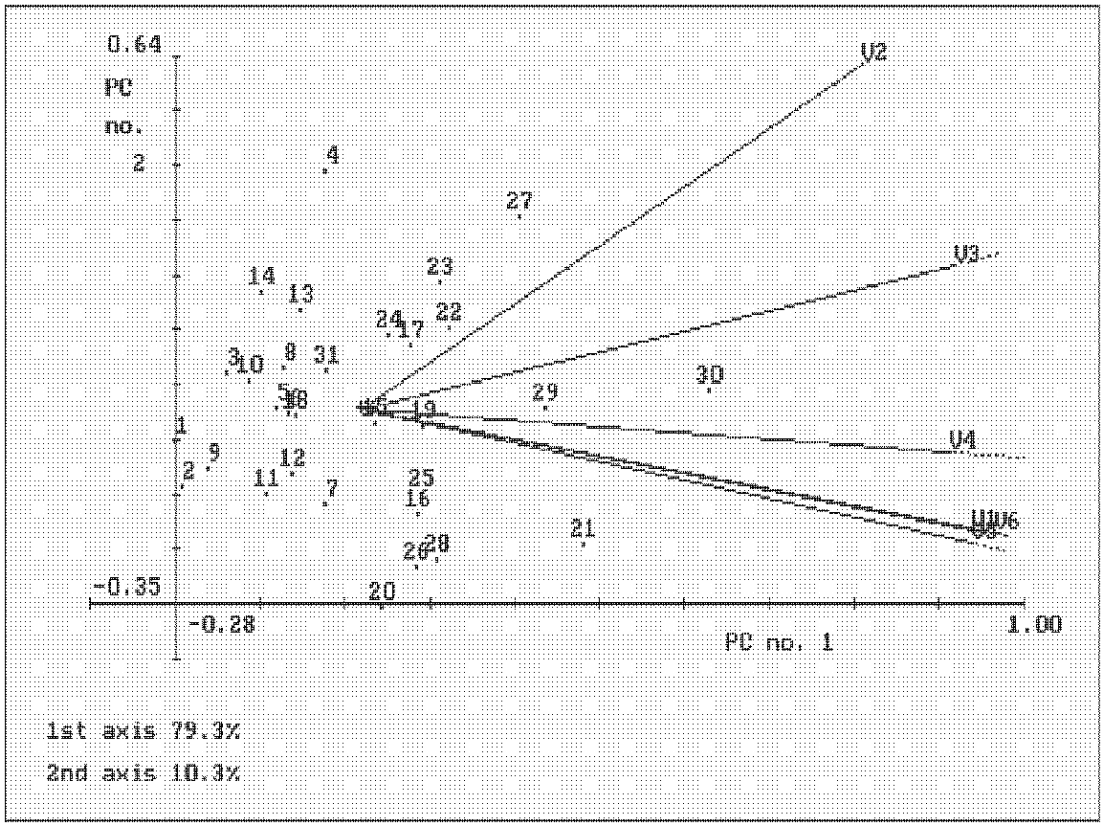

Figure 5: Augmented biplot visualizing correlations between variables V1÷V6 characterizing 31 farms. The representation is a fair one, except perhaps for variable V4. Farm N^o 30 has the highest values of all six variables, farm N^o 1 – the lowest

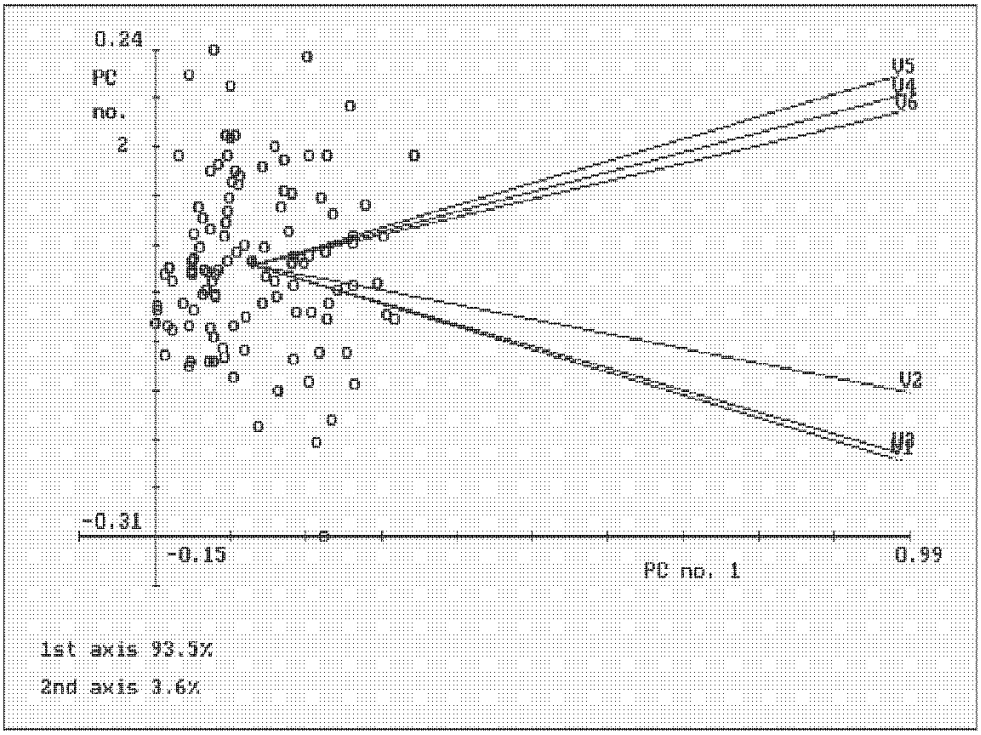

Figure 6: Augmented biplot characterizing correlations between measurements of arterial blood pressure recorded using 2 methods (automatic and traditional), each measurement taken in 3 positions (standing, sitting and lying) – when examining 117 women. The representation is very good. Note the separation of the automated (V1,V2,V3) and the traditional (V4,V5,V6) measurements

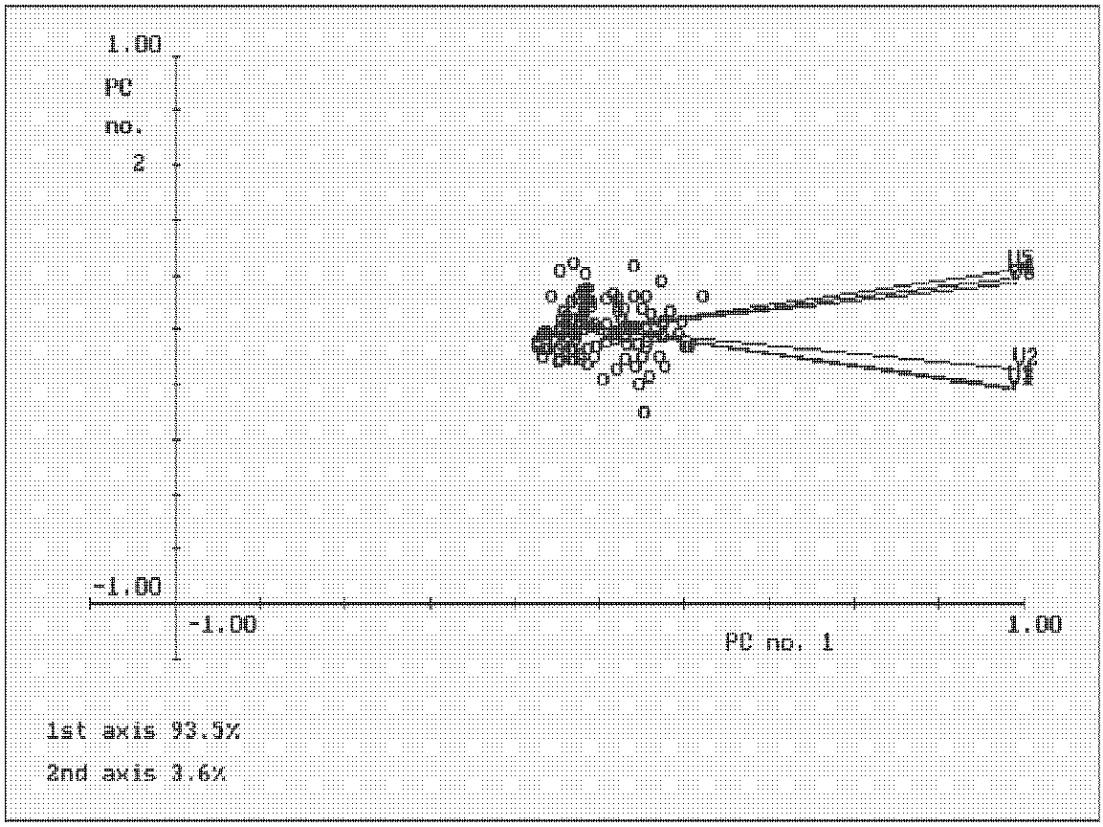

Figure 7: As Fig. 6, however using a different scale