JOURNAL OF CHEMOMETRICS
J. Chemometrics 2003; **17**: 503–511
Published online in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cem.814

Scatter plotting in multivariate data analysis

Paul Geladi[1]*, Marena Manley[2] and Torbjörn Lestander[3]

[1]Unit of Biomass Technology and Chemistry, Swedish University of Agricultural Sciences, Röbäcksdalen, PO Box 4097, SE-904 03 Umeå, Sweden
[2]Department of Food Science, University of Stellenbosch, Private Bag X1, Matieland 7602, South Africa
[3]Department of Silviculture, Swedish University of Agricultural Sciences, SE-901 83 Umeå, Sweden

Received 8 July 2002; Revised 29 April 2003; Accepted 9 May 2003

In data analysis, many situations arise where plotting and visualization are helpful or an absolute requirement for understanding. There are many techniques of plotting data/parameters/residuals. These have to be understood and visualization has to be made clearly and interpreted correctly. In this paper the classical favourites in chemometrics, scatter plots, are looked into more deeply and some criticism based on recent literature references is formulated for situations of principal component analysis, PARAFAC three-way analysis and regression by partial least squares. Biplots are also afforded some attention. Examples from near-infrared spectroscopy are given as illustrations. Copyright © 2003 John Wiley & Sons, Ltd.

KEYWORDS: plotting; visualization; multivariate data analysis; line plots; interpretation of scatter plots; biplots; near-infrared spectroscopy; principal component analysis; PARAFAC; partial least squares regression

1. INTRODUCTION

Ever since the start of chemometrics in the late 1960s and early 1970s, visualization of raw and transformed data, model parameters, residuals and other results has had an important place in it. One may even argue that visualization and chemical interpretation of plots form the true nature of chemometrics. Plotting, graphs and visualization of results, often problem-based, can be used in many ways, and this usually causes no big problem, but sometimes care has to be taken not to misrepresent or misinterpret results. There has not been much in the chemometrics literature on how to systematically make and interpret plots. The statistical literature has also not offered very much material. It has always been assumed that everybody knows how to make plots and does it right.

There are classical books on how to lie with simple graphs and maps [1,2]. These show how disproporionate plotting, zooming in on details and/or choice of scales and axes can be used to make the uninitiated or badly informed reader come to false conclusions. Chemometricians have always used line and scatter plots in huge numbers for showing results from principal component analysis and partial least squares regression. Maybe it is time to look deeper into what visualization in especially scatter plots really is. Many publications and popular software programs use plots without really describing how they were made and why they were made so. There is a whole gamut of sloppily made or slightly misleading plots. This stems from the early days of chemometrics, when scientists were happy to just fill their computer screens or lineprinter pages with symbols, but the tradition continues in many publications.

This paper takes up the subjects: what we plot in chemometrics and why; scatter plots from principal component analysis; scatter plots from three-way analysis; scatter plots from regression and calibration; and biplots. A few examples are used to illustrate the concepts. The terms plotting, graphing and visualization are often used to mean the same thing. In this paper they are used interchangeably.

2. WHAT DATA ARE SHOWN AND WHY?

The first question to ask is: what does a chemometrician want to do with data? The answer is complicated. A list is given in Table I. Most of the activities in the table need visualization or are made easier by using plots. Table II gives some suggestions of easily made and useful plots.

Visualization of data in general is old and some good general and historical references are available in References [3–5]. In Cartesian plotting, plots are made in squares or rectangles, with the scales and ticks just outside (or inside) the rectangular frame. This is the Cartesian co-ordinate system. A description of terms is given in Reference [3]. Descartes has also invented parallel co-ordinates, a technique that may be useful when not too many variables are to be studied simultaneously [6].

Geographical maps are familiar to most individuals. In such maps the horizontal and vertical co-ordinates are the same (distance in meters or kilometers) and distances and

*Correspondence to: P. Geladi, Unit of Biomass Technology and Chemistry, SLU Röbäcksdalen, PO Box 4097, SE-904 03, Sweden.
E-mail: paul.geladi@chem.umu.se

Contract/grant sponsor: EU Unizon–Kvarken project, subproject NIRCE.

Table I. What to do with data in a data analysis situation

- Inspect raw and transformed data
- Detect outliers/groupings/gradients
- Select a model
- Build the model, calculate the model parameters
- Choose a pseudorank
- Inspect and use the model parameters
- Study the residuals
- Prediction diagnostics

angles can be readily interpreted. This assumes that the map is over an area small enough to ignore curvature. In a geographical map, every point has a physical meaning. Contour plots in experimental design [7] follow the same principle, but here the horizontal and vertical scales do not necessarily mean the same thing and distances and angles do not always mean anything physical. Images [8] are a special form of maps with square pixels. Geograpical maps and satellite images are very similar, but with complementary information.

Scatter plots are square or rectangular plots and are quite different from maps or contour plots. Scatter plots can be made by plotting some measured result against some parameter in a Cartesian co-ordinate system. Examples are phosphate ions in a lake against depth, speed of a synthetic reaction against autoclave pressure, current against voltage, etc. The principle of scatter plots is that the entries from two vectors of the same size are plotted pairwise in the Cartesian co-ordinate system. The area outside the plotted points is not used and may never become useful. Regions within the plots may be physically impossible and therefore always remain empty. This does not happen for maps. In plots such as the ones described here, it is possible to construct a model (a continuous line or curve) from the plotted points and to show this model in the plot. In this way, data and model are shown together, which increases the possibilities of interpretation. Line plots are really scatter plots where one of the plotted variables is increasing or decreasing in regular intervals.

Most plots traditionally made in the sciences show measured values in SI units, practical empirical units or indices derived from these, e.g. by taking ratios, products, sums, etc. If the units on the horizontal and vertical axes are different, an appropriate scaling to obtain a square or rectangular size of plot is chosen. This is often based on range scaling, with some 5%–10% extra shrinkage to keep points from falling on the frame of the plot. This technique is so well established that one usually looks at the plots without considering the basic principles that led to the result.

Latent variable methods produce scores and loadings and these rarely use physical units: the plots become virtually unit-free. The items plotted against each other (scores, loadings) are based on the same measured data but projected differently. The units become very complicated unless one variable has a huge loading and the others have small ones, because this would be a return to a univariate situation. In data analysis by principal component analysis the scores are orthogonal and the loadings are orthonormal. This makes the co-ordinate systems used for the score plots orthonormal. With correct horizontal and vertical scaling, such plots can be used to show distances and angles in a meaningful way. A combined way of showing scores and loadings with a meaningful interpretation of distances and angles is the biplot or joint plot. Latent variables are produced in a number of models: PCA, PLS regression, factor analysis, PARAFAC, etc. The way in which the latent variables are used for plotting is not always the same and depends a lot on the properties of the latent variables in question.

Table II. Some plotting techniques that are useful and used frequently in chemometrics

- Line plots (spectra, chromatograms, concentration profiles)
- Two-dimensional plots (score and loading plots in PCA, PLS and for three-way analysis)
- Three-dimensional plots
- Response and other surfaces
- Quantile plots
- Biplots and joint plots in two and three dimensions
- Imaging or mapping of results and parameters
- Dendrograms

Table II gives some visualization techniques used frequently in chemometrics. Many of these can be used without great difficulty even by inexperienced users and need no further explanation, but the scatter plots, especially score and loading plots from latent variable methods, need some special attention.

3. PLOTTING IN COMPONENT MODELS

The scatter plot is not problem-free. This is most easily shown with principal components, but problems become more complicated with regression models and three-way models. Some equations for principal component analysis and PARAFAC analysis are given.

Principal component analysis of $\mathbf{X}$ is also given as a singular value decomposition. A model with two principal components is

$$\mathbf{X} = \mathbf{u}_1 s_1 \mathbf{v}_1' + \mathbf{u}_2 s_2 \mathbf{v}_2' + \mathbf{E} = \mathbf{t}_1 \mathbf{p}_1' + \mathbf{t}_2 \mathbf{p}_2' + \mathbf{E} \qquad (1)$$

These two types of equation are the ones that are most frequently used. A more general equation for R components is

$$\mathbf{X} = \mathbf{U}\,\mathbf{S}\,\mathbf{V}' + \mathbf{E} = \mathbf{T}\,\mathbf{P}' + \mathbf{E} \qquad (2)$$

where

$\mathbf{X}$	data matrix ($I \times K$), properly scaled, centered, etc.
$\mathbf{E}$	residuals ($I \times K$)
$\mathbf{U}$	orthogonal matrix of scores, each column normalized to one ($I \times R$)
$\mathbf{V}$	orthogonal matrix of loadings, each column normalized to one ($K \times R$)
$\mathbf{S}$	diagonal matrix of singular values ($R \times R$)
$\mathbf{u}_1$, $\mathbf{u}_2$	first two columns of $\mathbf{U}$, normalized scores
$\mathbf{v}_1$, $\mathbf{v}_2$	first two columns of $\mathbf{V}$, loadings
$\mathbf{t}_1$, $\mathbf{t}_2$	first two columns of $\mathbf{T}$, with $\mathbf{T} = \mathbf{US}$, scores
$\mathbf{p}_1$, $\mathbf{p}_2$	first two columns of $\mathbf{P}$, with $\mathbf{P} = \mathbf{V}$.

The vectors $\mathbf{p}_1$ and $\mathbf{p}_2$ can be shown as bar plots for a few variables or as line plots if many variables are used. They can also be shown as scatter plots, usually called loading plots. The vectors $\mathbf{t}_1$ and $\mathbf{t}_2$ (or $\mathbf{u}_1$ and $\mathbf{u}_2$) can be shown as line plots

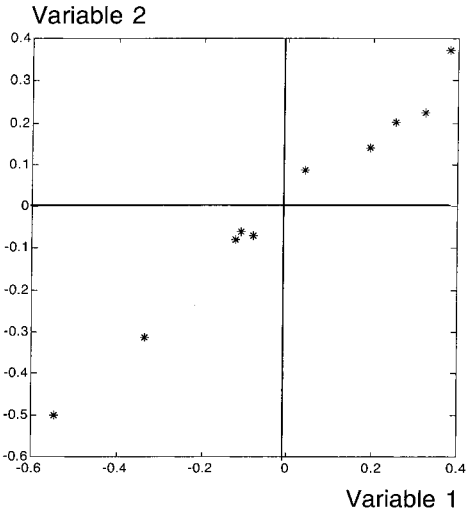

Figure 1. The mean-centered data of the example in a scatter plot.

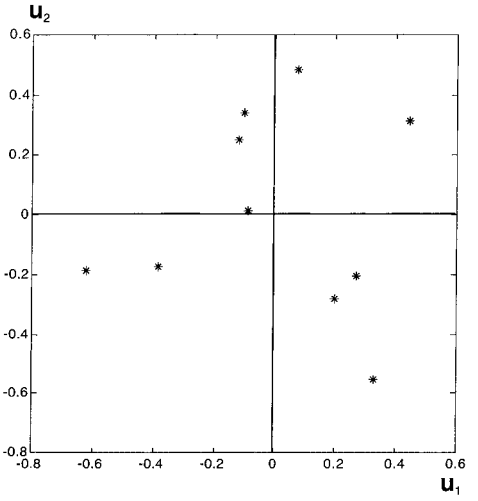

Figure 3. The same score plot as in Figure 2, but now just filling the square with points and ignoring sizes of scores by using normalized scores. Distances in different directions are different. Angles are wrong.

if there are many objects and they are ordered, e.g. according to sampling time. They can also be shown as scatter plots, called score plots.

Sign inversion is well known in PCA. It comes from the equality

$$\begin{aligned}\mathbf{X} &= \mathbf{u}_1 s_1 \mathbf{v}_1' + \mathbf{u}_2 s_2 \mathbf{v}_2' + \mathbf{E} \\ &= (-\mathbf{u}_1) s_1 (-\mathbf{v}_1)' + \mathbf{u}_2 s_2 \mathbf{v}_2' + \mathbf{E} \end{aligned} \quad (3)$$

This means that any pair of scores and loadings can be mirrored. It also means that small differences between algorithms for calculating PCA may give mirrored results. This is not a big problem, but it sometimes confuses newcomers trying to make an interpretation of their data. The good thing is that PCA scores and loadings are mirrored together.

An easy example of a score plot is illustrated in Figures 1–3. A similar example is also found in Reference [9]. There are two correlated variables ($r = 0.99$) and 10 objects. The raw data are variable 1 [0.9501, 0.2311, 0.6068, 0.4860, 0.8913, 0.7621, 0.4565, 0.0185, 0.8214, 0.4447] and variable 2 [1.0732, 0.3895, 0.7912, 0.6336, 0.9266, 0.8432, 0.6436, 0.2019, 0.9035, 0.6234]. Figure 1 shows the mean-centered raw data in a plot of variable 1 (horizontal) and variable 2 (vertical). PCA on the mean-centered data gives two components explaining 99.5% and 0.5% of the total sum of squares. Figure 2 shows the score plot of the two PCA scores ($\mathbf{t}_1$ and $\mathbf{t}_2$ in Equation (1)). It can be seen that all distances and angles between objects are kept correct. The structure inside the multivariate space is not changed. It can be seen that the two variables are correlated and that this gives a strong first component. The second component is only noise around this first component. One should notice that the figures are made square and with equal scales and that zero is indicated. This is good practice. Figure 3 shows the plot of the normalized scores ($\mathbf{u}_1$ and $\mathbf{u}_2$ from Equation (1)). This is the same as filling the square with points. The structure in the data is lost and it is easy to conclude that the data were just noise. Unfortunately, this is how many score plots are shown in the literature. The reason why Figure 2 gives the correct geometrical interpretation is that the basis ($\mathbf{p}_1$ and $\mathbf{p}_2$ from Equation (1)) for the figure is orthonormal and that it is made with identical scales.

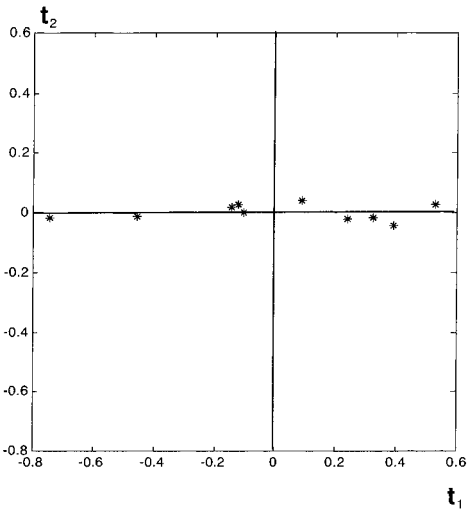

Figure 2. The score plot after PCA on the data of Figure 1. The geometry of the multivariate space is preserved. Distances in all directions are correct. Angles are correct.

Figure 4 shows a PCA score plot for a more complicated situation: a 112×1501 data matrix of FT-NIR spectra of wheat samples [10]. PCA was done on mean-centered data and gave four components explaining 79%, 18%, 1.6% and 0.8% of the total sum of squares, a total of 99.4%. The samples represent five irrigation treatments, one of which is highlighted by using different symbols. The percentage of the sum of squares is indicated in the plot. Figure 5 is the same score plot as the one in Figure 4, but now made by filling the rectangle. Both figures have their good points and drawbacks. In Figure 5 the separation of one of the irrigation classes (crosses) from the others (asterisks) is clearer, but in Figure 4 the geometry is correct. Interpretation of the variability of the cluster (crosses) is more correct in Figure 4 than in Figure 5. Indicating the % SS explained on the axes helps in avoiding erroneous conclusions. In the book by Beebe *et al.*

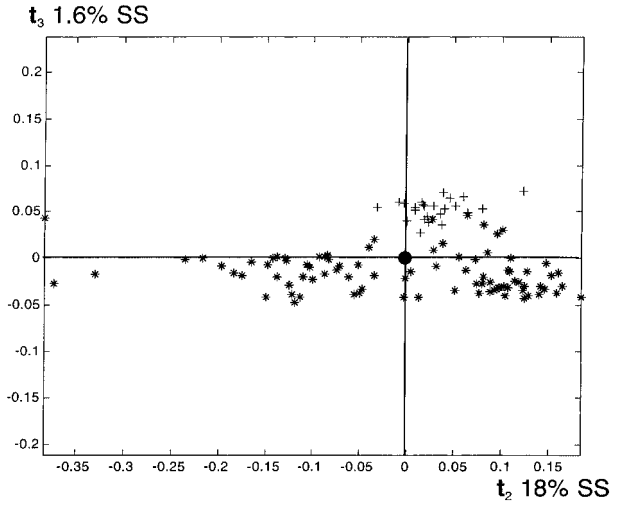

Figure 4. Score plots for the second and third principal components for the wheat example. One of the irrigation locations is indicated by crosses. Notice that zero is clearly indicated. The percentage of the sum of squares (SS) explained is indicated on the axes.

[11] the score and loading plots are introduced properly and some of the principles are explained. The two ways of plotting introduced are Euclidean correct, with equal scales, and 'just filling the square'. The scales and their properties are discussed in a clear manner, so this is *the* book for finding out how to make score and loading plots.

Zero should always be indicated in a plot. In many models, zero has a special function. In the PCA model a zero loading means that the variable does not contribute to the component shown and a zero score means that the object shown is close to the mean (center) of the data set, if mean-centered data are used. Not knowing where zero is in a plot or what a zero value means is very risky for the interpretation of plots.

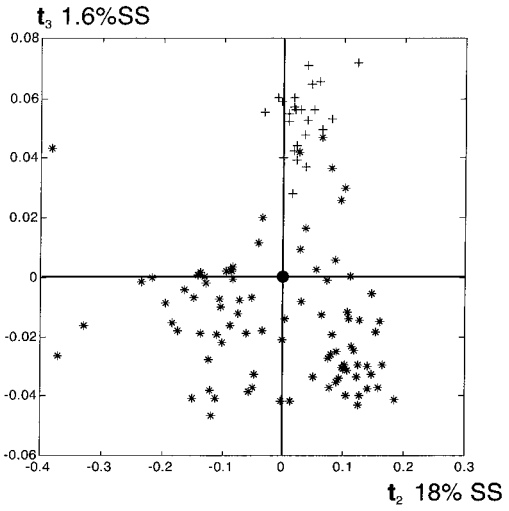

Figure 5. A version of Figure 4 made by filling the rectangle with points. The separation of the crosses from the asterisks is seen more clearly, but the variability in the cluster of crosses can be misinterpreted.

In three-way analysis by PARAFAC [12–16] the equation is

$$\begin{aligned} x_{ijk} &= a_{i1}b_{j1}c_{k1} + a_{i2}b_{j2}c_{k2} + e_{ijk} \\ &= (-a_{i1})(-b_{j1})c_{k1} + a_{i2}b_{j2}c_{k2} + e_{ijk} \\ &= (-a_{i1})b_{j1}(-c_{k1}) + a_{i2}b_{j2}c_{k2} + e_{ijk} \\ &= a_{i1}(-b_{j1})(-c_{k1}) + a_{i2}b_{j2}c_{k2} + e_{ijk} = \cdots \end{aligned} \tag{4}$$

where

x_{ijk} element of three-way array $\underline{\mathbf{X}}$
e_{ijk} the residual, element of residual array $\underline{\mathbf{E}}$
a_{i1} ith element of first A-loading vector
b_{j1} jth element of first B-loading vector
c_{k1} kth element of first C-loading vector.

The loadings of the PARAFAC model are not orthogonal and the nice properties of PCA loadings are not always available for them. This makes scatter plots of PARAFAC loadings different and sometimes line plots are preferred.

Equation (4) means that in PARAFAC models for three-way data any pair of loadings may be mirrored, leaving the remaining loading unchanged [17–19]. This can really cause confusion in the interpretation of loading plots, because moving in one direction in one loading plot may correspond to moving in the opposite direction in another.

Kiers [20] has shown that it is always possible to rotate factors or components to an orthogonal basis and get Euclidean properties for the plots. This can be done in a much simpler way than explained earlier [20]. For any bilinear or trilinear model, write the model as

$$\mathbf{X} = \hat{\mathbf{X}} + \mathbf{E} \tag{5}$$

where $\hat{\mathbf{X}}$ is the model and $\mathbf{E}$ is the residual matrix. Decompose $\hat{\mathbf{X}}$ by the singular value decomposition

$$\hat{\mathbf{X}} = \mathbf{USV}' \tag{6}$$

$\mathbf{V}$ is the orthonormal basis and the elements of $\mathbf{US}$ can always be plotted in a Euclidean manner. $\hat{\mathbf{X}}$ may be a factor model, results from curve resolution or a matricized trilinear model, e.g. PARAFAC.

An example is taken from Reference [21]. Pine tree seeds are produced by crossing known mothers and fathers in a two-way ANOVA layout of six mothers (called A–F) and 10 fathers. For each cell in the ANOVA an NIR spectrum of 1050 wavelengths is measured, giving a $6 \times 10 \times 1050$ three-way array. This array is analyzed with PARAFAC and three PARAFAC components are retained, of which the first one only explains the average spectrum and the second and third ones explain differences between mothers and fathers. It is also possible to produce components according to Equations (5) and (6).

Figure 6 is a scatter plot of PARAFAC A-loadings 2 and 3 with the mothers indicated. PARAFAC loadings are normalized to length one, but the figure was made with a rescaling to take into account the size difference of components 2 and 3. A plot according to Equations (5) and (6) is shown in Figure 7. The two figures are slightly different, but the main content is the same: mother C is different from the other ones. Conclusions about mothers A, B and D–F are different for the two figures. The choice of which figure is best is

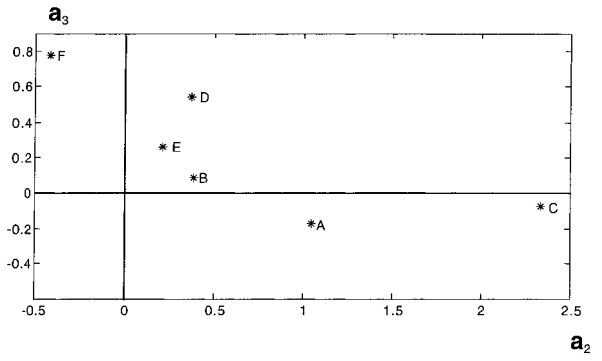

Figure 6. A scatter plot for A-loadings 2 and 3 of the seeds example. The mothers are named A–F.

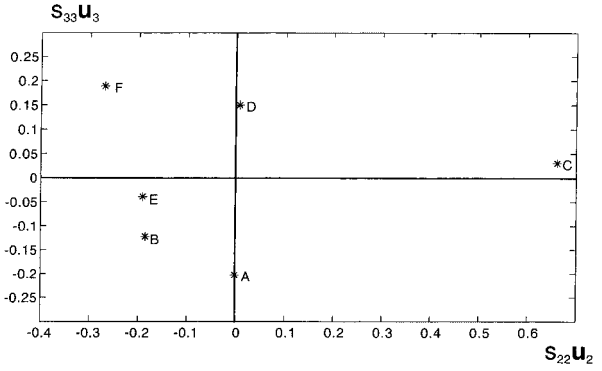

Figure 7. A scatter plot for the A-loadings of the seeds example made as in Equations (5) and (6). Components 2 and 3 of the SVD are used.

difficult to make. Those who have a strong belief in latent variables (A-loadings) would prefer Figure 6. Figure 7 is more geometrically correct; however, by calculating the SVD, some information may have been rotated into the unused first component.

One last remark is that the interpretation in score and loading plots is also dependent on the assumption that the scaling and mean-centering of $\mathbf{X}$ in Equation (1) are appropriate. Changing the scaling of the variables influences the plots. This is also valid for the next section on partial least squares modeling.

4. PLOTTING IN PARTIAL LEAST SQUARES REGRESSION

Another source of vectors and their line and scatter plots in chemometrics is regression models. In particular, the partial least squares (PLS) algorithm [11,22,23] gives many vectors to be plotted. Partial least squares regression makes the model

$$\mathbf{y} = \mathbf{X}\mathbf{b}_{\mathrm{pls},R} + \mathbf{f}_R \tag{7}$$

where

$\mathbf{y}$	vector of responses, mean centered ($I \times 1$)
$\mathbf{X}$	matrix ($I \times K$) of predictor variables, mean centered and properly scaled
$\mathbf{b}_{\mathrm{pls},R}$	vector of K regression constants
$\mathbf{f}_R$	the vector of I residuals.

A PLS model is constructed for a number R of PLS components. The pseudorank R is very important: if R is too small, there is underfitting; if R is too large, the model becomes very good but gives bad predictions, a situation of overfitting. There are two different PLS algorithms that give the same $\mathbf{b}_{\mathrm{pls},R}$ but many and also different model parameters as shown below. This makes studying line and scatter plots in PLS a challenge.

Calculating one PLS component works as follows.

1. $\mathbf{w}_1' = \mathbf{y}'\mathbf{X}$ (8)
2. Normalize $\mathbf{w}_1$ to length one, giving the first weight vector of $\mathbf{X}$.
3. $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$, the first score vector of $\mathbf{X}$ (9)
4. $q_1 = (\mathbf{t}_1'\mathbf{t}_1)^{-1}\mathbf{t}_1'\mathbf{y}$, the first loading of $\mathbf{y}$ (10)
5. $\mathbf{g}_1 = (1/q_1)\mathbf{y}$, the first score vector of $\mathbf{y}$ (11)

Different algorithms can be given depending on how $\mathbf{X}$ and $\mathbf{y}$ are reduced (deflated) in order to find the next component.

$$\text{Alt. 1} \quad \mathbf{E}_1 = \mathbf{X} - \mathbf{t}_1\mathbf{p}_1' \tag{12}$$

where $\mathbf{p}_1$ (the first PLS loading of $\mathbf{X}$) is calculated so that $\mathbf{E}_1$ is orthogonal to $\mathbf{t}_1$: $\mathbf{t}_1'\mathbf{E}_1 = \mathbf{0}'$.

$$\text{Alt. 2} \quad \mathbf{E}_1^* = \mathbf{X} - \mathbf{t}_1\mathbf{w}_1' \tag{13}$$

The deflation of $\mathbf{y}$ is given as

$$\mathbf{f}_1 = \mathbf{y} - q_1\mathbf{t}_1 \tag{14}$$

The operation in Equation (14) is only needed for the calculation of y-scores $\mathbf{g}_r$. If the $\mathbf{g}_r$ are not needed, it can be left out.

The next PLS component is calculated using $\mathbf{E}_1$ and $\mathbf{f}_1$ instead of $\mathbf{X}$ and $\mathbf{y}$. This gives $\mathbf{E}_2$ and $\mathbf{f}_2$ and so on until the correct number of components (R) is found. Then $\mathbf{b}_{\mathrm{pls},R}$ can be given as

$$\mathbf{b}_{\mathrm{pls},R} = \mathbf{W}(\mathbf{P}'\mathbf{W})^{-1}\mathbf{q} \tag{15}$$

where

$\mathbf{P}$	($K \times R$) PLS X-loadings
$\mathbf{W}$	($K \times R$) PLS X-weights
$\mathbf{q}$	($R \times 1$) PLS y-loadings.

The PLS model is used for prediction in a test set:

$$\hat{\mathbf{y}} = \mathbf{X}_\mathrm{t}\,\mathbf{b}_{\mathrm{pls},A} \tag{16}$$

where

$\mathbf{X}_\mathrm{t}$	test set ($J \times K$)
$\hat{\mathbf{y}}$	predicted responses ($J \times 1$).

The known responses for the test set $\mathbf{y}_\mathrm{t}$ (J × 1) are then used for checking whether the prediction really works by making a scatter plot of $\hat{\mathbf{y}}$ against $\mathbf{y}_\mathrm{t}$. An explanation of what can be seen in such plots is given in Reference [24]. Another good test of a PLS regression model is the inner relation [25] obtained by plotting $\mathbf{g}_r$ (the score vector for $\mathbf{y}$ in the PLS algorithm) against $\mathbf{t}_r$ for each PLS component ($r = 1, \ldots, R$). This inner relation can show outliers, non-linearity, grouping of data, etc.

A PLS model produces a large number of vectors and these can all be used to make line and scatter plots. A list is

Table III. A list of common scatter plots used for PLS models, with comments

• PLS scores $\mathbf{t}_2$ against $\mathbf{t}_1$, $\mathbf{t}_3$ against $\mathbf{t}_1$ . . . are not the same as for PCA
• Inner relation $\mathbf{g}_1$ against $\mathbf{t}_1$, $\mathbf{g}_2$ against $\mathbf{t}_2$ . . . should be made square with equal scales
• PLS loadings $\mathbf{p}_2$ against $\mathbf{p}_1$, $\mathbf{p}_3$ against $\mathbf{p}_1$ . . . choose $\mathbf{p}$ or $\mathbf{w}$ or a modified $\mathbf{w}$?
• PLS weights $\mathbf{w}_2$ against $\mathbf{w}_1$, $\mathbf{w}_3$ against $\mathbf{w}_1$ . . . choose $\mathbf{p}$ or $\mathbf{w}$ or a modified $\mathbf{w}$?
• Prediction quality $\hat{\mathbf{y}}$ against $\mathbf{y}_t$ . . . has to be made square with equal scales

given in Table III, with some comments. For the score plots it should be remembered that these are not the same plots as for PCA scores (Equations (12) and (13) are different from Equation (1)). For something that resembles the loading plots for PCA loadings, sometimes $\mathbf{p}_r$, sometimes $\mathbf{w}_r$ is used and the choice is rarely properly explained [22,25]. The inner relation in PLS ideally gives a linear relationship with slope one [25], so a square plot with equal scales is the least misleading option. The same goes for the plot of $\hat{\mathbf{y}}$ against $\mathbf{y}_t$ [24].

Ergon [26] has studied Equations (12) and (13) and given useful comments. For many PLS components they may be rewritten as

$$\mathbf{X} = \mathbf{T}\mathbf{P}' + \mathbf{E} \tag{17}$$

or

$$\mathbf{X} = \mathbf{T}^*\mathbf{W}' + \mathbf{E}^* \tag{18}$$

where

$\mathbf{T}$, $\mathbf{T}^*$ PLS X-scores
$\mathbf{E}$, $\mathbf{E}^*$ X-residuals.

There has always been confusion about which vectors to use for PLS loading plots. This confusion is not reduced by the use of modified versions of $\mathbf{W}$ in the literature [25]. Both Equations (17) and (18) have their drawbacks. Equation (17) gives an orthogonal $\mathbf{T}$, but a non-orthonormal plotting basis $\mathbf{P}$ and extra vectors in $\mathbf{W}$. Equation (18) has an orthonormal plotting basis in $\mathbf{W}$, but $\mathbf{T}^*$ is not orthogonal. Luckily, it is always possible to use Equation (18) in the same decomposition as in Equations (5) and (6) to get scores and loadings with nice properties. In this case, $\mathbf{T}^*\mathbf{W}'$ would be the matrix $\hat{\mathbf{X}}$. More details about this are given by Ergon [26]. This technique also has a disadvantage. The order of the PLS components is changed compared with the traditional PLS model and this may require some sorting out of results.

5. BIPLOTTING

Loadings and scores may be studied separately, but more information can sometimes be extracted by plotting them together in a biplot. One may imagine that score and loading plots can be made on transparent sheets and that these can be superimposed and interpreted in that way. This is intuitively appealing and useful, but the geometry is complicated and misinterpretations tend to be made.

When making biplots (overlayed scatter plots), a few rules of thumb have to be remembered:

(a) the relative scaling of the different components for a scatter plot of scores with a large and a small eigenvalue;
(b) the relative scaling of the scores and loadings—variables and objects do not use the same scale and then each axis has two different scales;
(c) compensation for an inbalance in the number of objects and variables;
(d) the physical scales used on the horizontal and vertical axes.

The biplot was developed by Gabriel [27] in 1971 and later also described by other authors [9,28,29]. Gabriel's original biplot was only meant for rank-two models. The use of three-dimensional biplots and components other than the first and second ones was added later and may require a different interpretation. By changing Equation (1) the following is obtained:

$$\mathbf{X} = (\mathbf{u}_1 s_1^{1/2})(s_1^{1/2}\mathbf{v}_1)' + (\mathbf{u}_2 s_2^{1/2})(s_2^{1/2}\mathbf{v}_2)' + \mathbf{E} \tag{19}$$

The singular value is distributed equally among the $\mathbf{u}$ and $\mathbf{v}$ parts (scores and loadings) for the purpose of forming new variables $\mathbf{h}$ and $\mathbf{g}$ to be plotted:

$$\mathbf{X} = \mathbf{h}_1\mathbf{g}_1' + \mathbf{h}_2\mathbf{g}_2' + \mathbf{E} \tag{20}$$

Equation (20) can also be rewritten as

$$\mathbf{X} = (\mathbf{u}_1 s_1^c)(s_1^{1-c}\mathbf{v}_1)' + (\mathbf{u}_2 s_2^c)(s_2^{1-c}\mathbf{v}_2)' + \mathbf{E} \quad 0 \le c \le 1 \tag{21}$$

The special cases of $c = 0$, 0.5 and 1 are described by Jackson [9], but other values of c may also be chosen for a subjective interpretation. The case of $c = 1$ is the row metric-preserving version. It gives Euclidean distances between objects and Mahalanobis distances between variables. Setting $c = 0$ gives the column metric-preserving version. It gives Euclidean distances between variables and Mahalanobis distances between objects. Equations (19) and (21) give scores and loadings that fit the SVD model, but for pure plotting purposes it is also possible to choose exponents that do not exactly sum to one in equation (21) [30]. Good examples of biplots are shown in Reference [29].

Equation (19) works well when the number of objects and variables is almost equal. With 1000 variables and 10 objects, all variables would end up close to the origin of the biplot, and the objects would be more spread out. This is so because of the constraint of length one of $\mathbf{u}$ and $\mathbf{v}$ in the singular value decomposition. A compensation for number of objects (I) and variables (K) is made by introducing a fudge or zoom factor z:

$$\mathbf{X} = z(\mathbf{u}_1 s_1^c)(1/z)(s_1^{1-c}\mathbf{v}_1)' + z(\mathbf{u}_2 s_2^c)(1/z)(s_2^{1-c}\mathbf{v}_2)' + \mathbf{E} \tag{22}$$

Biplots can be expanded to the use of three-way loadings, especially for Tucker3 models. Then they get the name joint plots [28]. A good example can be found in Reference [31].

For making biplots, some practical rules may be given.

(a) Check the score ($\mathbf{u}_1 s_1$ and $\mathbf{u}_2 s_2$) and loading ($\mathbf{v}_1$ and $\mathbf{v}_2$) plots separately. One needs equal scales on the horizontal and vertical axes. Extreme outliers may have

Table IV. Interpretation of scatter plots for PCA

Interpretation of a Euclidean scatter plot (score plot)
• Distances are correct in vertical, horizontal and diagonal directions
• Objects close to zero: not important for the components shown in the plot, close to the data set center
• Objects close to each other: close to each other in multivariate space, similar
• Dense cluster: low variability in the components shown
• Spread-out cluster: large variability in the components shown
• Cluster shape: correct for the components shown
Interpretation of a Mahalanobis scatter plot (loading plot)
• Distances are not easily interpreted unless the components have equal importance
• Variables close to zero: not important for the components shown in the plot
• Variables close to each other: close to each other in multivariate space, similar
• Variables on opposite sides from zero: similar except for opposite influence on the model

to be removed in both score and loading plots. Variables with small loading values may have to be marked and remembered as unimportant, possibly removed, especially in cases with many variables.

(b) Use the vectors $[(\mathbf{u}_1\sqrt{s_1})'\,(\sqrt{s_1}\mathbf{v}_1)']$ and $[(\mathbf{u}_2\sqrt{s_2})'\times(\sqrt{s_2}\mathbf{v}_2)']$ and make the scatter plot with equal scales on the horizontal and vertical axes. This plot gives a quick overview of how scores and loadings behave. No distances are Euclidean in this plot unless the component sizes are equal.

(c) Use the vectors $[(\mathbf{u}_1 s_1)'\mathbf{v}_1']$ and $[(\mathbf{u}_2 s_2)'\mathbf{v}_2']$ and make the scatter plot with equal scales on the horizontal and vertical axes. In this plot the distances between objects are Euclidean. It is possible to draw lines from the origin of the plot to the variables and to project the objects on these lines. Some properties of Euclidean and Mahalanobis plots are given in Table VI.

(d) If the variables or the objects end up too close to the center of the plot, some compensation in the scale may be given by using a zoom factor.

The example of Figures 1–3 was augmented to size 10×6 by adding four small noisy variables. After mean-centering, PCA was done, giving two components explaining 96% and 2% of the total sum of squares. The plot according to Equation (19) is shown in Figure 8. A plot according to Equation (21) with $c = 1$ is shown in Figure 9. In Figure 8 it is clearly seen that variables v1 and v2 are important for the first score and that variables v3–v6 are less important. This figure is not Euclidean for either objects or variables. In Figure 9 the objects are shown in a Euclidean projection and the variables are shown as a Mahalanobis projection. The conclusion is that the first component is the important one, with correlated variables v1 and v2 determining the component direction. Variables v3 and v4 would not be interpreted as important, because the second component is small and the spread in objects in that direction is small in the figure. This can also be seen by drawing a line from the plot origin to v3 and v4 and projecting the objects on that line.

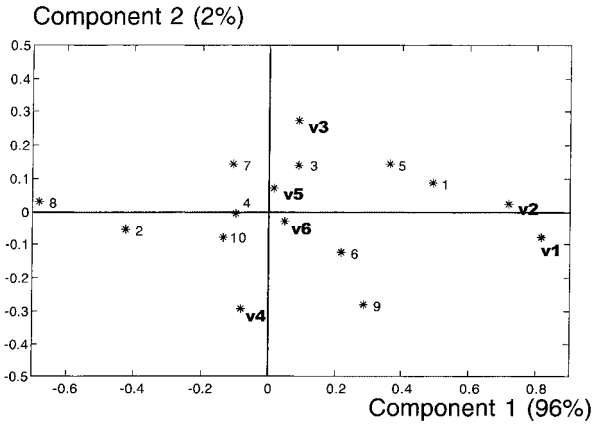

Figure 8. A biplot according to Equation (19) for the example as in Figures 1–3, but now with four extra noise variables. Objects are numbered and variables are called v1–v6.

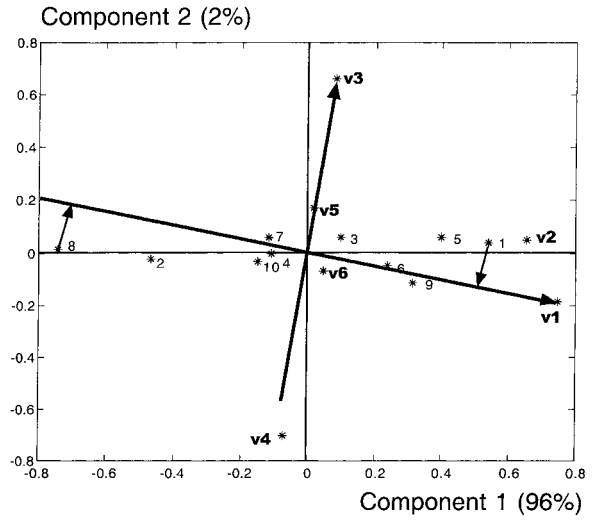

Figure 9. The same example as in Figure 8 used for a plot according to Equation (21) with $c = 1$. Objects are numbered and variables are called v1–v6. It is possible to draw arrows through the origin to the variables, here shown for v1 and v3. The objects can be projected orthogonally on these arrows for an interpretation, as shown for objects 1 and 8. The same projection of objects on the arrow pointing to v3 would only give small values.

The wheat data set used for Figures 4 and 5 was reduced for demonstration purposes to a 20×15 matrix. The objects were 10 each from two different irrigation locations. The wavenumbers used were 10 000, 6600, 6400, 6200, 6000, 5800, 5600, 5400, 5200, 5000, 4800, 4600, 4400, 4200 and 4000 cm^{-1}. PCA on mean-centered data gave three to four meaningful components explaining 97.4%, 2.1%, 0.28% and 0.16% SS, for a total of 99.9%. Figure 10 shows a biplot according to Equation (19). This is the quickest way to get objects and variables in the same plot. One may notice a possible outlier in the objects, on the negative side of the first component. All variables are on the positive side of the first component, meaning that they are all negative for this outlier. One may also notice one variable (10 000 cm^{-1}) that ends up close to zero, meaning that in the two-component model explaining 99.5% SS it does not contribute very much. Figure 11 shows

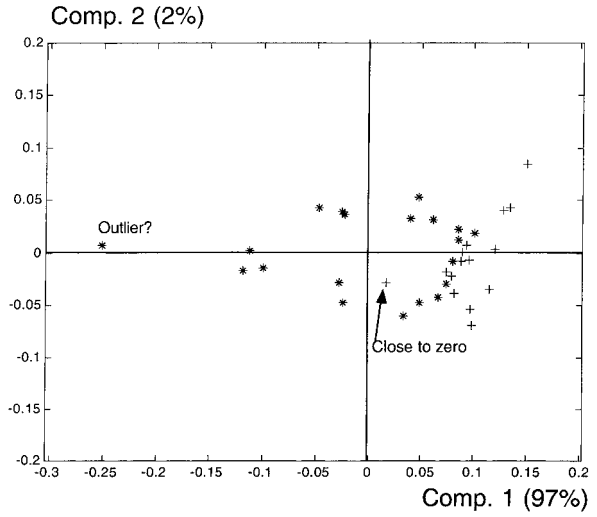

Figure 10. A biplot for the reduced (20 × 15) cereal data made according to Equation (21) with $c = 0.5$. (* = object, + = variable).

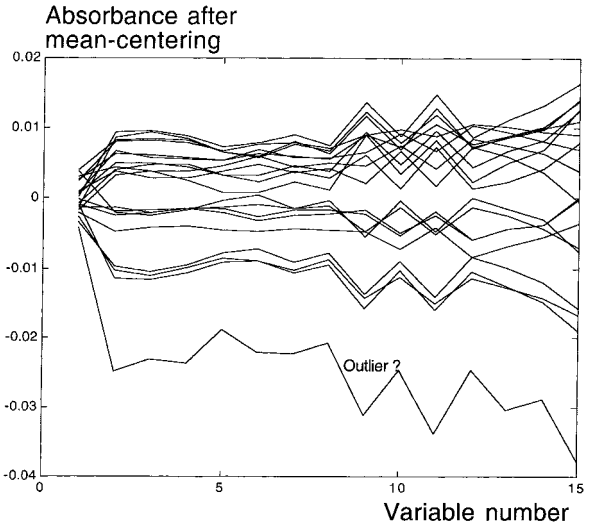

Figure 11. The data of the reduced (20 × 15) cereal data, with indication of the possible outlier. This figure is also a primitive form of parallel co-ordinates.

how the physical reality agrees with what could be seen in the biplot. One (mean-centered) spectrum is negative in all variables, and variable number 1 is the least important in describing differences between the objects. One may also see that the spectra cluster in two groups corresponding to the irrigation locations.

A last example is about viability of pine seeds [32], in other words answering the question: if this seed is planted, will it give a living tree? This is a binary classification problem. NIR transmittance spectra were collected from single seeds in the range 850–1048 nm. Spectra from each seed were measured at four different drying times, giving 874 spectra. Figure 12 shows a PCA score plot for this example, where the axes in the score plot have been interpreted as viability and drying time. The assignment of physical phenomena to the principal components takes away the need for having an exact Euclidean representation.

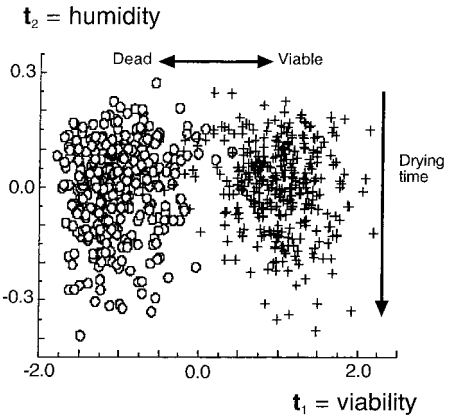

Figure 12. This is a score plot where the components have been assigned to physical phenomena. The need for a correct Euclidean representation is not present anymore.

Data analysis is not only geometry. The scores in a score plot also have a latent variable interpretation. The example in Figure 12 is also a situation where the components have been identified and named and Euclidean properties of the original multivariate space are not that important anymore. The sizes of some of the eigenvalues may also be the result of sampling, and then they change drastically if samples are added or deleted. The above shows that there is a duality in interpreting scatter plots. Principal components do not always show reality. In curve resolution situations the goal is to have pure spectra, chromatograms, etc. as loadings and proportions or concentrations as scores. This is a much stronger argument than needing an orthonormal basis for getting Euclidean properties in the plots.

6. CONCLUSIONS

Many types of plot can be used in data analysis. Line and scatter plots have always been a favourite for the latent variable or factor models that are popular in chemometrics. Line plots are actually also scatter plots. The scatter plot can show different things dependent on what is used to make it. It can also be made to look in a number of ways. Some of them can easily be misinterpreted.

A scatter plot can be made for two different variables measured in different physical units, and that is quite problem-free, but it can also be made for two principal component scores, in which case measures have to be taken to make it a real projection from multidimensional space onto a plane. This projection can have specified properties such as Euclidean or Mahalanobis distance. Loadings and weights from PLS form an extra problem, but a simple modification using singular value decomposition can give Euclidean correct plots. The same SVD modification works for the loadings from PARAFAC analysis.

When components are identified as specific latent variables, the rules of making score plots are not as strict. Principal components are an artificial construction, and real factors, e.g. from curve resolution, may have to be used instead, giving rise to yet another way of interpreting scatter plots.

Biplots and joint plots have specific rules of construction and interpretation. They are not made by just putting a score plot and a loading plot on top of each other. The secret is in partitioning the singular values correctly between the normalized scores and loadings.

There are many ways of making scatter plots and biplots and they all have their advantages. There is nothing wrong with a personal interpretation of how scatter plots and biplots should be made as long as the principle of construction is clearly indicated and the interpretation is done according to that principle. It is also important to realize that conclusions about a scatter plot of two components are not conclusions about the whole model. This is only the case if the two components together explain a large proportion of the total sum of squares of the data.

Acknowledgements

The EU Unizon–Kvarken project, subproject NIRCE is acknowledged for financial support.

REFERENCES

1. Huff D. *How to Lie with Statistics*. Victor Gollancz: London, 1954.
2. Monmonier MS. *How to Lie with Maps*. University of Chicago Press: Chicago, IL, 1991.
3. Cleveland WS. *The Elements of Graphing Data*. Wadsworth Advanced Books and Software: Monterey, CA, 1985.
4. Cleveland WS, McGill ME (eds). *Dynamic Graphics for Statistics*. Wadsworth and Brooks/Cole: Belmont, CA, 1988.
5. Tufte ER. *The Visual Display of Quantitative Data*. Graphics Press: Cheshire, CT, 1983.
6. Wegman EJ. Hyperdimensional data analysis using parallel coordinates. *J. Am. Statist. Assoc.* 1990; **411**: 664–675.
7. Meyers RH, Montgomery DC. *Response Surface Methodology. Process and Product Optimization Using Designed Experiments*. Wiley-Interscience: New York, 1995.
8. Geladi P, Grahn H. Multivariate image analysis. In *Encyclopedia of Analytical Chemistry*, Meyers R (ed.). Wiley: Chichester, 2000; 13540–13562.
9. Jackson JE. *A User's Guide to Principal Components*. Wiley: New York, 1991.
10. Van Zyl L, Manley M, Osborne B. Using different sample holders in determining protein and moisture content in whole wheat flour by menas of Fourier transform near infrared (FT-NIR) spectroscopy. *S. Afr. J. Plant Soil* 2001; **18**: 50–55.
11. Beebe KR, Pell RJ, Seasholtz M-B. *Chemometrics. A Practical Guide*. Wiley: New York, 1998.
12. Smilde AK. Three-way analyses. Problems and prospects. *Chemometrics Intell. Lab. Syst.* 1992; **15**: 143–157.
13. Bro R. PARAFAC: tutorial and applications. *Chemometrics Intell. Lab. Syst*. 1997; **38**: 149–171.
14. Law HG, Snyder CW, Hattie J, McDonald RK (eds). *Research Methods for Multimode Data Analysis*. Praeger: New York, 1984.
15. Coppi R, Bolasco S (eds). *Multiway Data Analysis*. North Holland: Amsterdam, 1989.
16. Andersson C, Bro R (eds). Special issue. Multiway analysis. *J. Chemometrics* 2000; **14**: 103–331.
17. Geladi P, Bergner H, Ringqvist L. From experiments to images to particle size histograms to multiway analysis. An example of peat dewatering. *J. Chemometrics* 2000; **14**: 197–211.
18. Geladi P, Åberg P. Three-way modeling of a batch organic synthesis process monitored by near infrared spectroscopy. *J. Near Infrared Spectrosc.* 2001; **9**: 1–9.
19. Geladi P, Forsström J. Monitoring of a batch organic synthesis by infrared spectroscopy: modeling and interpretation of three-way data. *J. Chemometrics* 2002; **16**: 329–338.
20. Kiers HAL. Some procedures for displaying results from three-way methods. *J. Chemometrics* 2000; **14**: 151–170.
21. Lestander TA, Odén P-C, Geladi P. 2- and 3-way analysis of NIR scans from seed crossings. In *Near Infrared Spectroscopy. Proceedings of the 10th International Conference*, Davies A, Cho R (eds). NIR Publications: Chichester, 2002; 385–388.
22. Martens H, Næs T. *Multivariate Calibration*. Wiley: Chichester, 1989.
23. Brown PJ. *Measurement, Regression and Calibration*. Clarendon: Oxford, 1993.
24. Geladi P. Some recent trends in the calibration literature. *Chemometrics Intell. Lab. Syst.* 2002; **60**: 211–224.
25. Eriksson L, Johansson E, Kettaneh-Wold N, Wold S. *Multi- and Megavariate Data Analysis. Principles and Applications.* Umetrics: Umeå, 2001.
26. Ergon R. PLS score–loading correspondence and bi-orthogonal factorization. *J. Chemometrics* 2002; **16**: 368–373.
27. Gabriel KR. The biplot—graphic display of matrices with applications to principal component analysis. *Biometrika* 1971; **58**: 453–467.
28. Kroonenberg PM. *Three-mode Principal Component Analysis*. DSWO Press: Leiden, 1983.
29. Brereton RG (ed.). *Multivariate Pattern Recognition in Chemometrics, Illustrated by Case Studies*. Elsevier: Amsterdam, 1992.
30. Martens H, Martens M. *Multivariate Analysis of Quality. An Introduction.* Wiley: Chichester, 2001.
31. Gemperline PJ, Miller KH, West TL, Weinstein JE, Hamilton JC, Bray JT. Principal component analysis, trace elements, and blue crab shell disease. *Anal. Chem.* 1992; **64**: 523A–532A.
32. Lestander TA, Odén P-C. Separation of viable and nonviable filled Scots pine seeds by differentiating between drying rates using single seed near infrared transmittance spectroscopy. *Seed Sci. Technol.* 2002; **30**: 383–392.