Advancing from unsupervised, single variable-based to supervised, multivariate-based methods: A challenge for qualitative analysis

Bernhard Lendl, Bo Karlberg

This article reviews and describes the open challenges of defining the unreliability limit or region when advancing from unsupervised single variable-based to supervised, multivariate-based methods applied for the purpose of qualitative analysis. An unambiguous definition of unreliability regions is difficult to make when dealing with multivariate methods, although useful additional information, such as increased selectivity, may be gained when applying such methods.

Keywords: Multivariate; Qualitative analysis; Selectivity; Single variable; Unreliability limit; Unreliability region

Bernhard Lendl
Institute of Chemical
Technologies and Analytics,
Vienna University of
Technology, Getreidemarkt
9-164/AC, A-1060 Vienna,
Austria

Bo Karlberg*
Department of Analytical
Chemistry, Stockholm
University, Svante Arrhenius
Vag 10-12, SE-10691
Stockholm, Sweden

1. Qualitative analysis

Qualitative chemical analysis is the branch of analytical chemistry that is principally concerned with detecting and identifying one or more constituents of a sample with the overall purpose of producing a binary response (i.e., a "Yes" or a "No" regarding the presence of the constituent(s) in the sample concerned). The underlying analytical problem is formulated in such a way that a specified property is either assigned or not assigned to the sample. Thus, this type of analysis should enable a decision to be made as to whether or not a sample contains a certain category or class of compounds. Therefore, from a logical perspective, qualitative chemical analysis is a simple, fundamental *classification* methodology. Either the sample has a defined property A (and thus belongs to class A) or the sample does not (and does not belong to class A).

The way that a classification of this type is performed depends on the data that the method or technique used generates. A set of ballots cast in an election can, for example, be manually sorted through visual inspection into two categories (e.g., ballot papers with a punched hole and those with no hole). However, if the instrumental method adopted produces a wealth of multivariate data for each sample, elaborate, chemometric methods are required to make the desired Yes/No classification. It is important to recognize at this point that any method or technique used for classification purposes, no matter how simple it may be to perform, will eventually fail to classify all samples correctly. Consequently, the term *unreliability of a test* has been proposed for use in connection with qualitative analysis [1].

It is important to determine the factors that can make qualitative analytical tests unreliable. This is not straightforward, and at least two concepts have to be considered: *calibration* and *analytical selectivity*. An erroneous calibration and/or poor selectivity will definitely lead to unreliability.

2. Unsupervised and supervised classification

A classification can be unsupervised or supervised. Unsupervised classifications do not require any *a priori* knowledge about the samples that are going to be classified. The property that is assigned to the "Yes" category of samples is, as the ballot example given above, easily measured and

*Corresponding author.
Tel.: +46 8 16 43 16;
Fax: +46 8 15 63 91;
E-mail:
bo.karlberg@anchem.su.se.

0165-9936/$ - see front matter © 2005 Elsevier Ltd. All rights reserved. doi:10.1016/j.trac.2005.03.010

the test method used has proved to be sufficiently selective.

However, supervised classification requires *a priori* truth knowledge about the set of samples that is intended to be used for calibration purposes. A typical example of a supervised classification would be the determination of the origin of a particular wine through Fourier-transform infrared spectroscopy, FTIR (e.g., addressing the question: is this a Merlot wine?). FTIR data obtained for a large set of Merlot wines and FTIR data obtained for other types of wine may then be processed by an artificial neural network (ANN) training procedure to create a model that can be applied to new samples for which *a priori* knowledge about their origin is not available.

These two examples, classification of ballots and wine samples, represent two extremes of the widely complex spectrum of applications developed for qualitative analysis. To classify all these applications is challenging, as is the selection and/or development of suitable classification methods when going from simple, unsupervised methods to more complex classifications requiring a supervised calibration strategy. This article briefly elucidates and describes this urgent challenge for analytical chemists.

3. Zero-order instruments – single-variable methods

Instruments used in analytical chemistry are categorized according to the type of data they provide [2]. Zero-order instruments produce one datum per sample. Examples of such instruments are balances, pH meters and filter photometers. The observed variables are continuous. However, in this context, it will be necessary to discuss the concept of calibration. According to the 2004 draft revision of the VIM [3], *calibration* has two definitions:

(1) an operation establishing the relation between quantity values provided by measurement standards and corresponding indications of a measuring system, carried out under specified conditions and including evaluation of measurement uncertainty; and

(2) an operation that establishes the relation, obtained by reference to one or more measurement standards, that exists under specified conditions, between the indication of a measuring system and the measurement result that would be obtained using the measuring system.

The calibration of a balance differs substantially from the calibration procedure for a wet chemistry method, but the two definitions above cover both these examples.

Let us consider a Soxhlet extraction method for determining the fat content in a sample. The final analytical step is weighing the fat residue obtained after evaporation of the organic solvent. A calibrated balance will then produce data needing just a simple conversion to a percentage value, or no conversion at all. The underlying analytical problem might be to determine whether or not the sample has a low fat content. Samples having a fat content below a certain specified cut-off level, C_L, should be classified as ''low-fat products'', while all other samples do not belong to this category. This method can be compared with the spectrophotometric molybdenum blue method for determining phosphate-yielding absorbance values that must be converted to concentration values through a calibration function, preferably a linear function. The ''converted'' or ''derived'' variable values are then used for the final classification of samples.

The transfer of the observed or derived continuous variable values to binary 0/1-variable values has been treated thoroughly elsewhere [4,5]. Fig. 1 gives an excellent summary of the problems associated with moving from a continuous single-variable system to the binary system. C_L, defined as the cut-off limit, can be a concentration, a weight, a total index (e.g., phenol index or iodine number) or, in principle, any other parameter related to the sample. The classification is based on the very simple criterion that samples for which the continuous variable responses are larger than C_L will have the binary value 1 and consequently the assigned property A; all other samples will not have the assigned property. The uncertainty that this single-variable system possesses will then determine the *unreliability interval*. In Fig. 1, this interval is C_0–C_1, subdivided into the zone of false positives, C_0–C_L, and the zone of false negatives, C_L–C_1. Setting the limits of C_0 and C_1 will also

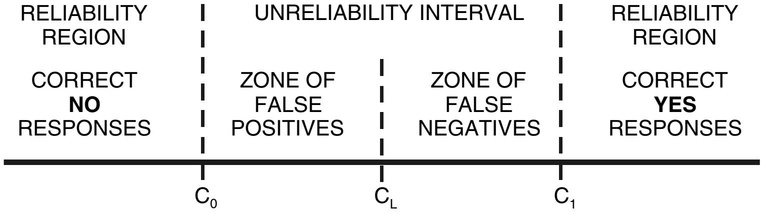

Figure 1. Illustration of the unreliability interval C_0–C_1 for single-variable methods for qualitative analysis. C_L is the cut-off limit.

determine the probability of reporting a correct answer when the obtained variable value lies in the reliability regions. In this context, the unreliability of a test has been defined as the property of a test method that characterizes the range of values of the measurand, for which the probability of committing errors of the first (false positive) or second kind (false negative) exceeds accepted limits.

Zero-order instruments require full analytical selectivity to give accurate values for the target variable. Interferences cannot be detected, since only one datum per sample is obtained. This datum usually represents, after suitable conversion through calculation or calibration, an analyte concentration. However, as mentioned above, many single-variable analytical methods aim at determining some type of total index that gives an estimated concentration of a group of analytes. A common feature of all single-variable methods is that they enable unsupervised classification when prior data are used to set or validate decision criteria.

4. First-order instruments – several variables

A first-order instrument is in principle an array of zero-order sensors. Diode-array spectrometers are typical examples of first-order instruments. However, the instrumental set-up for producing first-order data does not necessarily consist of a set of similar sensors; such set-ups may also consist of several different instruments, which analyze different properties of the sample.

Let us consider some typical practical examples for which first-order instruments are used when generating data.

The first example is the determination of protein in wheat. The reference method is the classical Kjeldahl method based on digestion, distillation and titration. This method is tedious, expensive and produces a lot of hazardous waste. Consequently, near infrared (NIR) analysis has today largely replaced it, but it is still the general reference method used to enable multivariate calibration. Such calibration is often applied to large data sets comprising thousands of samples for which both NIR spectra have been acquired and Kjeldahl analyses have been performed. Commonly applied regression approaches in this context are partial least squares (PLS) and ANN analyses. The regression vector, **b**, obtained from the multivariate calibration is then used to convert spectral data for the new samples with unknown protein contents into protein content values. Assume that we have an NIR analyzer calibrated in this way so that it displays the calculated protein content for each sample. For the analyst, such an instrument would resemble and perform like a zero-order instrument and, consequently, it could be analogously used for qualitative purposes. The underlying question asked might be: does the protein content in this wheat lot exceed 12.0%? The uncertainty of this classification method cannot be exactly determined, but it can be estimated. This means that the unreliability interval can also be estimated. The statistical tools that must be applied in such cases are somewhat more complex than those applied for zero-order instrument data. Comprehensive descriptions of state-of-the-art statistics can be found in textbooks on fundamental chemometrics. It is important to note that first-order instrument data can reveal the presence of interfering constituents in the sample, unlike zero-order instrument data. However, the interfering constituents cannot be quantified.

The second example is the classification of minced meat, as being of poultry (Yes) or non-poultry (No) origin. Again, NIR spectroscopy may be used. If it is used for calibration purposes, spectra of a large number of samples of all kinds of meat with known origin are first acquired. They are then classified, using an appropriate chemometric method. Several such methods are available and the optimal choice depends on the application. In many cases, there is no ‘‘best’’ choice. Principal component analysis (PCA) is a commonly applied classification method. If we are lucky, distinct clusters may be observed, and all poultry meat samples (and no others) may fall into a single cluster. In reality, this very rarely happens. To perform PCA on a set of NIR spectra, no *a priori* knowledge about the samples is required (unsupervised classification). However, in the minced meat example, we have relevant knowledge, so a supervised classification method, such as partial least squares discriminant analysis (PLS-DA), might be considered. The spectral data would then be contained in the X matrix and the binary response values in the Y matrix. The Y matrix contains two variables. The values of these two variables are (1;0) for poultry samples and (0;1) for all other meat samples. Prediction of Y for new samples with unknown origin will thus yield two *y*-values that are not necessarily either of the integers 0 and 1. The definition of an unreliability interval is clearly a challenge in such cases.

5. Second-order instrumentation – several variables

Second-order instruments are steadily gaining in importance in analytical chemistry, mainly because their prices are falling to affordable levels. An example of a second-order instrument is a liquid chromatograph equipped with a diode-array detector. An important feature of second-order instrumentation is that analysis can be performed in the presence of any component in the sample that is not included in the calibration model (the second-order advantage). The data treatment required for classification and qualitative analysis is, of course, more complex than the treatments required for first-order instrumentation. Each sample generates a huge number of data points that will eventually lead to a Yes/No answer.

6. The unreliability interval and the unreliability region

As concluded above (referring again to Fig. 1), zero-order data yield a well-defined unreliability interval, since uncertainty data can be generated to provide the basis for a straightforward calculation once the significance levels of errors of the first and second kinds have been agreed. The cut-off limit, C_L, represents the only datum for which we give both a Yes and a No answer. First and higher order instrumentation generates data for which the number of variables is reduced by an appropriate projection method to enable classification. In the wheat example discussed above, the number of variables is reduced to one and the classification is straightforward, even though determination of an unambiguous unreliability interval is more complex. However, this example differs in another distinct way from the zero-order (i.e., Soxhlet) example. Suppose that we are inadvertently testing a peanut sample instead of a wheat sample. The multivariate calibration model will then give us a predicted protein value and thus also the basis for a Yes or No answer to the qualitative question asked (i.e., whether or not the protein value is greater than 12.0%). However, the X residual will most likely be large for this peanut sample, giving us a new dimension of unreliability. This information is vital and enables us to detect that the given sample falls outside the established calibration. The fact that we can detect such cases will ascertain the achievement of correct classification results. This is an inherent property of first and higher order instrumentation, namely that outliers can be detected.

When using PCA for data reduction and classification purposes, the data are typically projected onto a plane (formed by two variables or 'orthogonal components') in a score plot. In such cases, a discriminant function may give a line that separates the two clusters (i.e., the Yes and No clusters), as in Fig. 2. Measurements projecting data points exactly on this line (after the PCA data treatment) belong to either of these two clusters, analogously to the case when the C_L value was obtained for zero-order instruments. However, again, the unreliability *region*, as we may call this entity, is no longer an interval but rather a region, projected onto a part of the plane, and is not easily calculated. Again, a ''new'' type of unreliability appears, for samples with score values far away from any of the two clusters and far away from the discriminant function line.

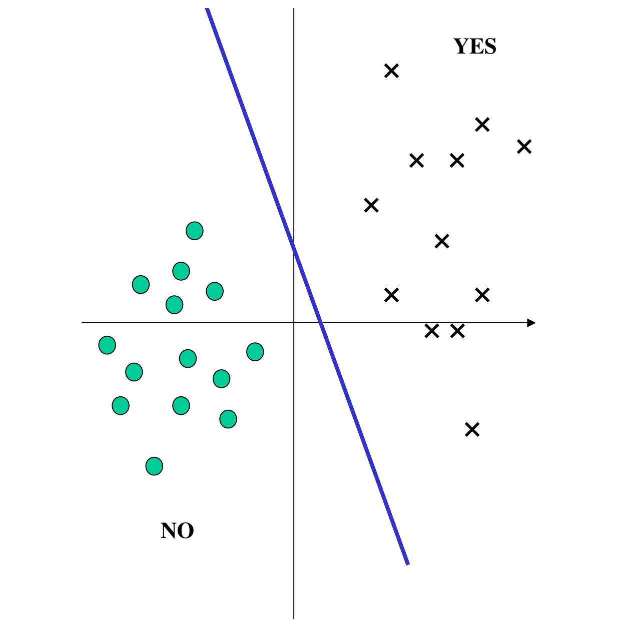

Figure 2. Use of a linear discriminant function to classify Yes/No samples.

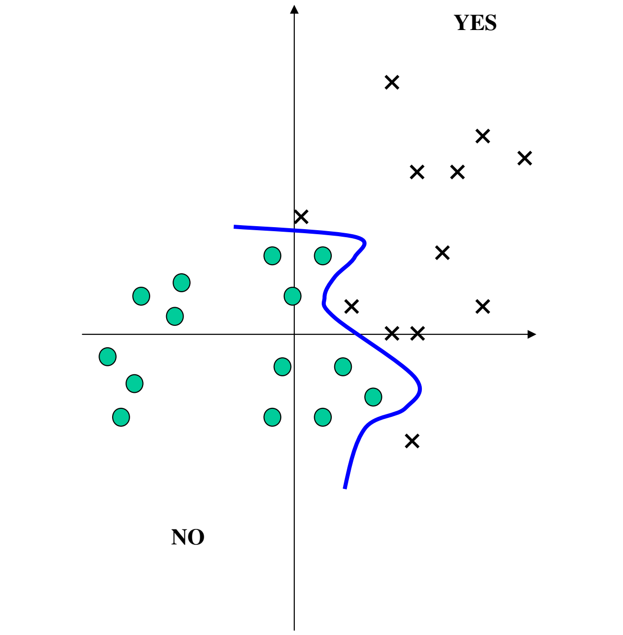

Figure 3. Use of a non-linear discriminant function to classify Yes/No samples.

Fig. 3 illustrates a fairly complex classification problem (data reduction to two component variables). No straight line can be utilized as the discriminant function in this case. Instead, non-linear discriminant function approaches, such as back-propagation ANNs or support-vector machines (SVMs), may be considered [6]. The unreliability concept is even more difficult to define for applications of this kind.

Soft independent modeling of class analogy (SIMCA) is a commonly applied supervised classification method [7]. Two or more separate PCA models are constructed. For qualitative purposes, the two classes are called Yes and No, and each class is enclosed in a defined hyper-volume. The boundaries of these hyper-volumes are defined in terms of the calculated standard deviations of points in each class. Any new measurement will be assigned probabilities for belonging to either of the two classes, so, in this case, there will be a consistent estimation of the unreliability. Consequently, SIMCA somewhat differs in this respect from the other mentioned classification methods for first and higher order data.

In conclusion, first and higher orders of instrumentation generate data that can be subjected to a supervised classification method that will give us the desired Yes or No answers. However, defining the unreliability region in such cases remains a challenge for analytical chemists. Nevertheless, it is worth repeating that, based on the nature of a multivariate response, additional sources leading to unreliable classification results can be detected. As an example, outlier detection has been discussed in the text. Such additional information is in principle not available when dealing with zero-order data.

References

[1] European Commission, Metrology of Qualitative Chemical Analysis, EUR 20605 EN, European Commission, Brussels, Belgium, 2002.

[2] K.S. Booksh, B.R. Kowalski, Anal. Chem. 66 (1994) 782A.

[3] International Standards Organization, International Vocabulary of Basic and General Terms on Metrology (VIM), 3rd Edition, Draft, ISO, Geneva, Switzerland, April 2004.

[4] W.A. Hardcastle, S. Ellison, Measurement Uncertainty and Traceability Workshop 2002, Expression of Uncertainty in Qualitative Testing, Approaches to the Problem, Eurachem/CITAC, LGC/VAM/2002/021, S.L.R. Ellison, S. Gregory, W.A. Hardcastle, Analyst (Cambridge, UK) 123 (1998) 1155.

[5] A. Ríos, D. Barceló, L. Buydens, S. Cardenas, K. Heydorn, B. Karlberg, K. Klemm, B. Lendl, B. Milman, B. Neidhart, R.W. Stephany, A. Townshend, A. Zschunke, M. Valcárcel, Accred. Qual. Assur. 8 (2003) 68.

[6] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines, Cambridge University Press, Cambridge, UK, 2000.

[7] S. Wold, M. Sjöström, ACS Symp. Ser. 52 (1977) 243.