# Robust calibration reveals secret

*Sabine Verboven shows how archaeologists can benefit from using a robust statistical method to determine the historical origins of ancient glassware. She explains the power of robust calibration using principal components regression*

Many scientific methods are available to discover the historical origins of archaeological artefacts. Good indicators for glassware are the proportions of chemical components such as silicon dioxide (SiO2). A data set of electron probe X-ray microanalysis (EPXMA) spectra over 750 wavelengths was collected on 180 different archaeological glass samples and plotted as in Figure 1. A calibration model for the equipment was needed before individual samples could be analysed. The conventional approach (evaluation of spectra) demands much knowledge from the operator and it is both time-consuming and difficult to automate. A faster and more automatic approach is to use the *principal component regression* (PCR) method to build the calibration model. I explain PCR in a separate box.

In this model, the spectra are the explanatory variables; the response variable is the measured SiO2. Once all model parameters are calculated, the amount of SiO2 of a new sample can be predicted easily. For this, we need only to measure the EPXMA spectrum of the new sample, plug the measured values into the calibration model and calculate its corresponding SiO2-concentration. But we have to be aware of the existence of contaminated samples, called outliers, in the calibration data set. These can ruin the calibration model and lead to incorrect predictions.

Multivariate calibration is widely used in quantitative spectroscopy. A compound's concentration is predicted by the spectra, measured for the corresponding samples over a range of hundreds or even thousands of wavelengths. The number of concentrations on the other hand is limited to five at most. In a univariate approach, only one compound concentration will be modelled at a time as a function of the spectra, whereas a multivariate calibration model tries to predict all concentrations at once. To analyse such measurements and to build models for prediction, the classical calibration methods use multivariate statistical methods such as least squares regression (LS), principal component analysis (PCA), principal component regression (PCR), and partial least squares regression (PLS). All of these techniques are sensitive to outliers and, therefore, robust alternatives were developed. I explain robust statistics in a separate box.

Especially in the area of on-line process analysis, outliers have become nearly unavoidable because the analyst has limited control over the process that generates the calibration data and because data sets are often highly dimensional. It is impossible to discard outlying observations manually.The technological progress in analytical equipment, which allows the collection of high-dimensional data sets, therefore necessitates (semi-)automated procedures to detect outliers. The difficulty in detecting outliers in multivariate datasets is illustrated in Figure 2. This bivariate data set could, for example, show the measurements of the carbon chain length as a function of boiling temperature of organic compounds. The red data points represent ring structures, such as cyclohexane; the regular blue data points represent non ring structures, such as n-hexane. None of the outlying observations (indicated in red) has abnormal values in one of the two coordinate directions. For this reason, the outliers cannot be spotted by looking at each of the dimensions separately. They can be found only by taking into account the *robust covariance structure* (represented by the green ellipse (solid line)) of this bivariate data set.

High correlation between variables (multicollinearity), as well as outliers, can complicate the building of a calibration model for a high-dimensional multivariate data set with many covariates.

This is often the case in the calibration of chemometrical data, where the X-variables correspond to spectra measured at many frequencies such as in the glass data set. The classical least squares estimators have large variances in such instances, and, when the number of spectra measured is larger than the number of samples in the data set, these cannot even be calculated. Many biased estimators have been proposed to overcome these difficulties. An appealing method is principal components regression (PCR) since it is easy to understand and to compute.

The classical PCR procedure (see box) is very sensitive to contaminations of the data. Therefore, a robust PCR (RPCR) method is needed. This robust alternative also has two steps:

- **The reduction step:** a robust principal component analysis is applied to the spectral data, yielding a smaller set of uncorrelated regressors. This is:
- **The regression step:** a robust regression method (such as the LTS- or MCD-regression method) is used to estimate the relationship between the robust principal components and the response(s) of interest.

Let's revisit the glass data set. We built a robust calibration model with PCR using both the classical and robust approaches. Since the method has two stages, we can inspect outlying spectra in the reduction step as well as in the regression step. To inspect the outlying spectra of the glass data, we look at some diagnostic plots, called the *outlier maps* (see figures 3a to 3d). The classical maps, based on the classical PCR, are at the top of Figure 3. The outlier maps at the bottom are based on robust PCR.

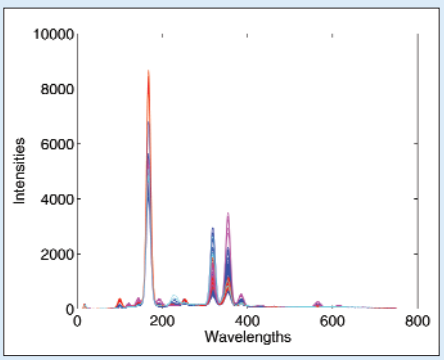


*Fig 1.* **Raw spectra of the Glass data set**

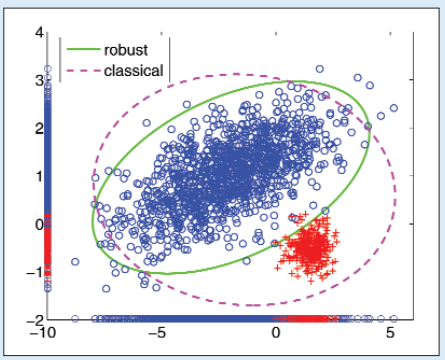


*Fig 2.* **Detecting outliers in a bivariate data set**

# of ancient glassware

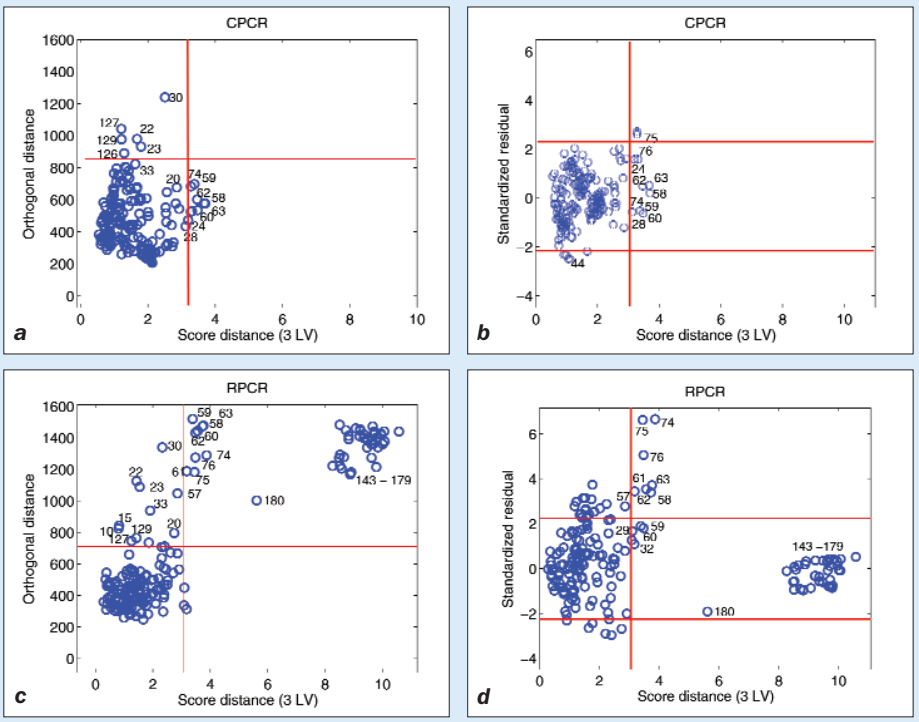


*Fig 3.* ***Outlier maps for the glass data set***

Consider the outlier maps at the reduction stage (Figures 3a and 3c). The classical method does not find important outliers whereas the robust method clearly distinguishes three major groups in the data: a group of regular data points (in the left, bottom corner) and two large groups of contaminated spectra.

The large group of spectra (numbers 143 to 179) and the smaller group with spectra numbers 58 to 63, 74 to 76 are called bad leverage points. The group in the upper left corner of the picture, containing spectra numbers 20, 22, 23, 30, 33,…, is a group of orthogonal outliers. There is also an isolated bad leverage point, glass spectrum number 180, lying in between the outlier groups. The idea behind this classification is simple. For each observation in the data set its distance to the lower dimensional subspace (orthogonal distance) and its distance to the other data points inside the subspace (score distance) are calculated. The data points are then divided into four classes based on the magnitude of these distances.

In Figure 4, all possible types of outliers for the reduction stage of the RPCR algorithm are visualized on a simulated three-dimensional data cloud. The simulated data set is built by two principal components and can be reduced into a two-dimensional plane depicted by the parallelogram. The marked data points are different types of outliers. So, the group of bad leverage points 143-179, 58-63, and 74-76 are comparable to the situations two and three in Figure 4. Each of these data points is far from the estimated subspace but its projection onto the subspace also lies far away from the bulk of regular data. The orthogonal outliers, glass spectra numbers 20, 22, 23,…, are similar to situation number five in Figure 4. Each of these observations is also far from the estimated subspace but its projection onto the subspace nicely falls inside the cloud of regular data points. Situations one and four (Figure 4), which we call

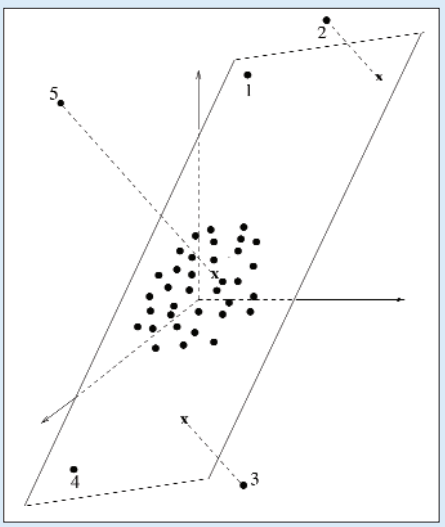


*Fig 4.* ***Different types of outliers in the reduction stage of PCR***

## WHAT IS ROBUST STATISTICS?

The most frequently used statistical techniques are very efficient but extremely sensitive to aberrant observations, also called *outliers*. Robust statistical methods are able to deal with these outliers. Suppose, for example, that recorded values (expressed in weight percentage) of silicon dioxide (SiO2) in a glass sample are: 68.52, 68.23, 67.42, 68.94 and 68.34. Assume that the first measurement was wrongly recorded as 18.52. This could be due to a malfunction of the measurement equipment or perhaps the production process was exposed to some abnormal external or internal changes. Hence, the average value for the five original measurements is 68.29, but for the contaminated array it becomes 58.29. This clearly shows that a single erroneous observation strongly affects the estimate. The harmful effect of one or more outliers on the estimate is a consequence of the non-robustness of the estimator of location which was, in this case, the mean. A robust estimator of location, such as the median, is influenced by outliers to a smaller extent. In the example, the median of the original values equals 68.34 and for the contaminated case changes into 68.23. Thus the median, being a robust estimator of location, gives a representative value of the amount of SiO2 in the glass sample – even when there are some abnormal values. There are many papers and books describing the need and effectiveness of robust methods, such as: Huber, PJ *Robust Statistics* (1981), Wiley: New York; Hampel, FR, Ronchetti, EM, Rousseeuw, PJ, and Stahel, WA *Robust Statistics: The Approach Based on Influence Functions* (1986), Wiley: NewYork.

All robust methods described in this article are part of LIBRA: the MATLAB LIBrary for Robust Analysis. The toolbox currently contains implementations of robust methods for location and scale estimation, covariance estimation (FAST-MCD), regression (FAST-LTS, MCD-regression), principal component analysis (RAPCA, ROBPCA), principal component regression (RPCR), partial least squares (RSIMPLS) and classification (RDA) and it provides many graphical tools to detect and classify the outliers.

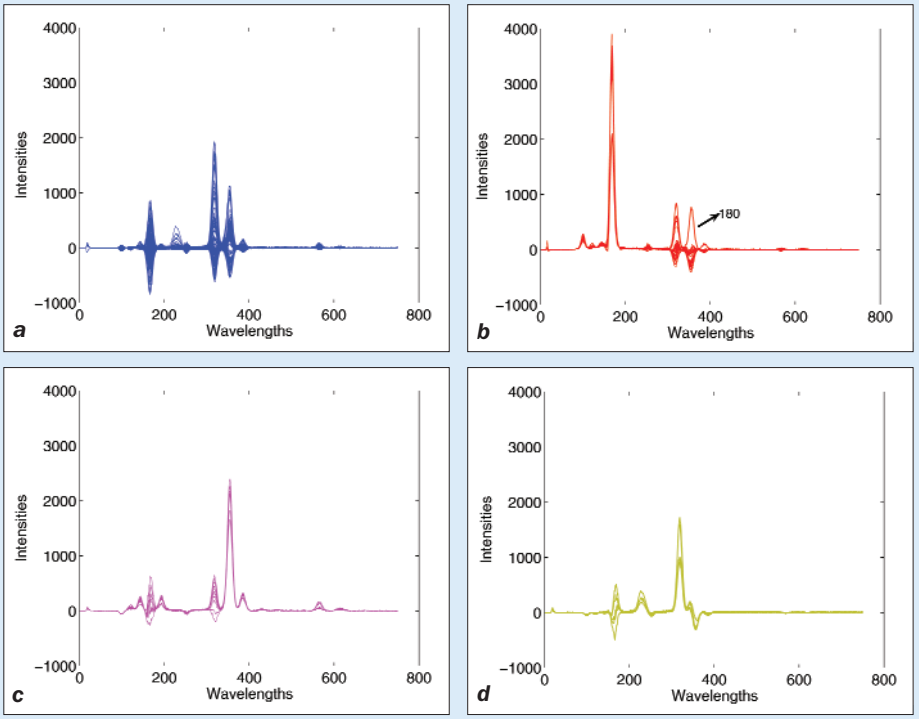


*Fig 5.* ***The spectra of the glass data set classified according to the robust outlier maps: (a) regular spectra; (b) bad leverage spectra 143 to 180; (c) bad leverage spectra 58 to 64 and 74 to 76; and (d) orthogonal outlying spectra 20, 22, 23, 30, 33.***

*good leverage points*, are not present at this stage in the glass data.

The classical PCR was fooled by the presence of the outliers. The bad leverage points influenced the estimation of the subspace in such a way that the outliers were completely masked.

At the regression stage (Figures 3(b) and 3(d)) the classical analysis lets us believe once again that there aren't many outliers. All spectra lie close to the borderline and the region between the two horizontal cut-off lines indicates the good regression leverage points. This means that each residual distance (which is, for a one-dimensional response model, equal to the standardised estimation error) is small, but the score distance is large. So, these spectra confirm the regression model at higher X-values.

However, looking at the robust regression output, another classification pops up. Although the spectra 143 to 180 are not harmful for the regression stage, as they are good leverage points, the smaller group (numbers 58 to 63 and 74 to 76) has had a bad influence. Notice that this group of observations affected the classical calculations by tilting the regression hyperplane in their direction such that the spectra are closer to the regression subspace. It therefore yielded smaller residuals and thus appeared as a set of good leverages points in figure 3b.

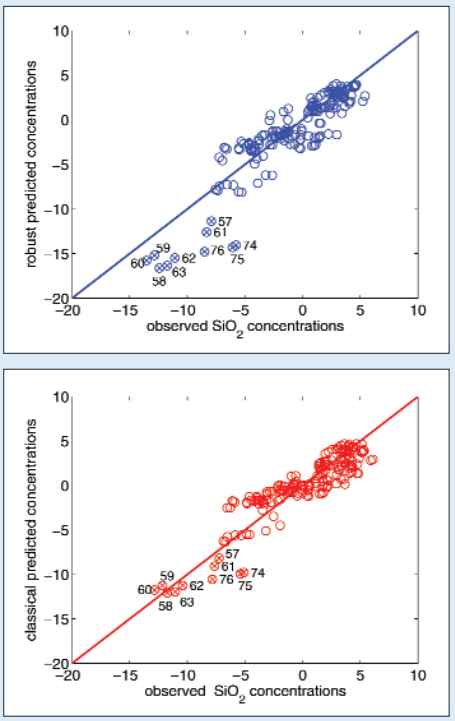


*Fig 6.* ***Original response versus estimated response for the glass data***

To identify outliers is one thing, but why these spectra are spotted by the robust PCR method is another. We must look closer at the spectra

## CLASSICAL PRINCIPAL COMPONENT REGRESSION (PCR) EXPLAINED

A two-dimensional data set, Figure 7, illustrates the simplest situation for the used of PCR. The coordinates $X_1$ and $X_2$ represent two explanatory variables such as the concentration of an antibiotic substance and the number of bacteria in a sample. These are negatively correlated to each other. We can combine them as a single explanatory variable to use in a simple linear regression model linking only that one variable to the response $y$, the amount of penicillin in the sample. To do this, we reduce the dimension of this dataset. We must do it without losing too much information. Intuitively we see that, by projecting all our data points in the direction $T_1$, we shall have the best answer to this problem. PCR will be the statistical technique to do that: it compresses this bivariate data set into its dominant direction $T_1$ and successively calculates the parameters of a simple linear regression model of $T_1$ on a response $y$.

PCR can be applied to situations with many more explanatory variables, to reduce the dimensions and to simplify the regression analysis accordingly.

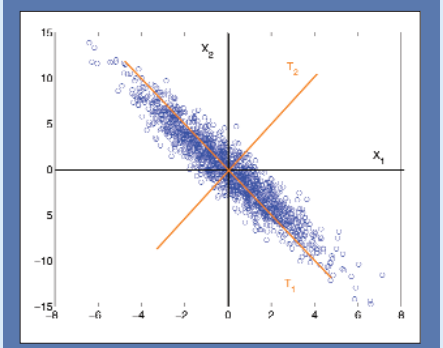


***Fig7. Principal components $T_1$ and $T_2$ of a simulated bivariate dataset.***

Sabine Verboven is with the Department of Mathematics, Statistics and Actuarial Sciences, University of Antwerp, Prinsstraat 13, B-2000 Antwerp, Belgium
sabine.verboven@ua.ac.be

1 Data from Lemberge, P. et al. (2000), "Quantitative Z-Analysis of the 16–17th Century Archaeological Glass Vessels using PLS Regression of EPXMA and μ-XRF Data," *Journal of Chemometrics*, 14, 751–763.

themselves. Comparing the spectra 143–180 (Figure 5b) with the regular spectra (Figure 5a) shows that they look alike but the intensities of the outlying group are much higher than the regular ones at the wavelengths 150 to180. The window of the detector system had been cleaned before the last 38 spectra were measured. As a result of this, less radiation (X-rays) was absorbed and more could be detected.

The other bad leverage points 58 to 63 and 74 to 76 are samples with a large concentration of calcium. In Figure 5(c) we see that their calcium Kα peak (around wavelengths 340 to 370) and calcium Kß peak (wavelengths 375 to 400) is higher than for the other glass vessels. The orthogonal outliers (20, 22, 23, 30 and 33) whose spectra are shown in Figure 5(d) have larger measurements at the wavelengths 215 to 245. This might indicate a larger concentration of phosphor.

Using the classical PCR model for predictions is not a good idea. Figure 5b shows that working with estimated parameters from the classical PCR results in modelling the outliers as well. The bad regression leverage points 58 to 63 and 74 to 76, also indicated in figure 3c, attracted the regression plane towards them. Therefore, it looks as if the predictions are almost perfect for every spectrum but the outliers are masked. The robust method on the other hand (figure 6a) exposes the outliers and the robust predictions of the regular spectra are still perfect.

The example demonstrates the outliers' influence on the classical analysis and in the meantime shows the resistance towards outliers of its robust counterpart. So, robust PCR identifies and classifies all outliers correctly and automatically, whereas classical PCR is clearly influenced by them.

It is important to identify the outliers in your data before moving to further analysis because, and I quote one of our contacts in industry: '*Our experience is that without robust multivariate calibration methods it would at best take considerably longer in constructing our calibration models and at worst it would take longer to construct the wrong model! We sometimes require on-the-fly model construction in an unattended fashion. In this case it is not possible to use standard techniques, as they can be completely fooled by the presence of outliers and this can lead to poor results which would in fact be very difficult to detect.*'

Thus, detection of outlying observations at an early stage can provide insight to the data and perhaps will allow defects in the production process to be traced. Moreover, wrong decisions, costing time and money, can be avoided by using a robust statistical method.

# Past issues?

## *Tony Greenfield urges you to reach out to those who need us*

Many years ago, at a scientific meeting, a speaker asserted that he did not need statistically designed experiments because he was certain of his theory. I jumped up to remind the died-in-the-wool physical modellers that their beloved immutable universal truths were just a previous generation's statistical fit to some experimental observations. This stirred the argument.

I was flattered later to be told that Francis Bacon had made a similar comment in 1620: 'The sciences we now possess are merely systems for the nice ordering and setting forth of things already invented.'

This he wrote in his *Novum Organum: or Indications Reporting the Interpretation of Nature*, in which he railed at Aristotle's ideas. Aristotle (384 to 322 BC) had a complete theory of science and a simple method for discovering scientific truths. Gather a group of clever people, he said, and encourage them to argue. If they are clever enough then the truth must emerge. Bacon poured scorn on this, saying that the Greek philosophers were like spiders, spinning webs from their own substance. He added: 'It was well observed by Heraclitus that men look for sciences in their own lesser worlds, and not in the greater or common world.'

A physicist, on a project for development of super-conductors, told me he had seven compositional variables and four process variables. I suggested I might help by designing an efficient experiment. 'It's not your field,' he said. 'I have to study the physics and chemistry of each variable separately to understand the problem clearly and to reach a useful result. It will take years.' He is a clever man but he lives in his lesser world.

The world is full of people who need to be able to use statistical methods in their work. That is why applied statistics is such an interesting and challenging career: it presents the statistician with so many different problems. In particular, the staff of manufacturing industries need to be skilled in the use of statistical tools. The managers, the engineers and the physical scientists can benefit greatly from an integration of statistical methods into their work. Sadly, most of them are not aware of these benefits. Statisticians are frustrated by that ignorance. We know that we have so much to offer that could have a major impact, through the manufacturing industries, on our national economies. That is why we must address the question of communicating statistics. We need a change of culture to bring about a greater acceptance by non-statisticians of statistical methods. But it is our culture, not theirs, which must first be changed. Instead of publishing so much to ourselves, through our own conferences and journals, we should reach out to those who need us, to convince them that they really do need us, so that they will come banging on our doors demanding our help.

We, in ENBIS, aim to become a web-based society. Winfried Theiss has worked long, hard, and skilfully to develop an excellent website. But, while that is of prime importance, there is a danger in focusing entirely on that. The danger is that, although the website is open to all, only members will visit it. If we restrict our publications to that and to academic journals, we shall create our lesser world. That is why we must thank Tom Wilkie, editor in chief of Europa Science, for providing us with four pages in every issue of *Scientific Computing World*. Perhaps you can suggest more ways to reach those who need us. Non-members will never see the website without these pointers.

How successful has this arrangement been? A recent message to me was: 'The magazine works by far the best. I downloaded all the issues and it is indeed impressive.'

So please do that. Download all the issues from www.enbis.org. Read them and consider how you will contribute. Do not allow ENBIS to become a lesser world. Make it a major planet. Remember the essential purpose of technical journalism: to communicate technical information to other people in such a way that they can use it. And remember the words of Isaac Asimov: '*I'm on fire to explain. I don't indulge in scholarly depth.*'