Chemometrics in Crystallography: Cluster and Partial Least Square Regression analyses from a crystallographic perspective

Gwilherm Nénert

PANalyticl B.V. Netherlands; gwilherm.nenert@panalytical.com

Cluster analysis[1] and Partial Least Square Regression (PLSR)[2] analyses are rather common chemometrics method applied in various fields where large amounts of data need to be treated. While those methods have been developed in social sciences, they became popular analytical techniques in spectroscopy. Recent years have shown that these methods could be applied also successfully in crystallography[3,4,5]. The aim of this contribution is to give a simple introduction to cluster analysis and PLSR methods and treat few examples aiming to give a glance of their possibilities. Examples are treated with the HighScore suite[6] which implemented the necessary algorithms to carry out such analyses besides the usual capabilities of crystallographic softwares.

Cluster analysis

Cluster analysis greatly simplifies the analysis of large amounts of data. It automatically sorts closely related scans of an experiment into separate clusters and marks the most representative scan of each cluster as well as outlying patterns. From diffraction perspective, it can be useful for non-ambient experiments, for synthesis experiments of e.g. zeolites, to find polymorphs and solvates in drug development, etc.

Cluster analysis is basically a three step process, but contains an optional fourth step too:

1. **Comparison** of all scans in a document with each other. The result is a correlation matrix representing the similarity of any given pair of scans.
2. **Agglomerative hierarchical cluster analysis** puts the scans in different classes defined by their similarity. The output of this step is displayed as a **dendrogram**, where each scans starts at the left side as a separate cluster, and these clusters amalgamate in a stepwise fashion until they are all united.
3. The **number of clusters** is estimated by the KGS test or by the largest relative step on the dissimilarity scale. Also the most representative scan within each cluster is determined.
4. PCA (**P**rincipal **C**omponents **A**nalysis) can be carried out as a separate and independent method to visualize and to judge the quality of the clustering. The correlation matrix of step 1 is used as input.

An example of clustering is for instance the case of young child looking at a picture. He or she can quickly label the objects in a picture as buildings, vehicles, people, animals, etc.

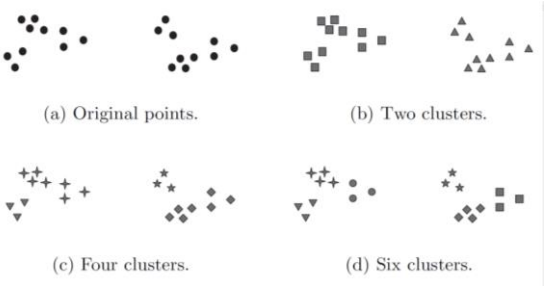

Figure 1: Different way of clustering the same set of points.

In many cases, the definition of a cluster is something not so obvious. Let's take the example of Figure 1 where we have 20 points distributed in space. Figures 1a, 1b and 1c show 3 different ways to cluster those points. The shape of the markers indicate cluster membership. Figures 1b and 1d divide the points into two and six clusters, respectively. However the apparent division of each of the two larger clusters into 3 subclusters may simply be an artefact of the human visual system. It is not unreasonable either to consider that those 20 points can be divided into 4 clusters such as in Figure 1c. This is illustrating that the definition of a cluster is imprecise and that the best definition depends on the nature of the data and the desired results.

There are numerous ways in which clusters can be formed. We can distinguish various types of clustering: hierarchical versus partitional, exclusive versus overlapping versus fuzzy, and complete versus partial. Hierarchical clustering is one of the most straightforward methods. It can be either agglomerative or divisive. Agglomerative hierarchical clustering begins with every case being a cluster itself. At successive steps, similar clusters are merged. The algorithm ends with all small clusters in one single cluster. Divisive clustering starts with everybody in one cluster and ends up with everyone in individual clusters. Obviously, neither the first step nor the last step is a worthwhile solution with either method.

To form clusters using a hierarchical cluster analysis, you must select several criteria. One criterion is used to determine the similarity or distance between cases, another criterion for determining which clusters are merged at successive steps and finally the number of clusters you need to represent your data.

One important note here is that there is no right or wrong solution when you do a cluster analysis about the number of clusters that you use. It depends on what you are going to do with those clusters. To find a good cluster solution, you must look at the characteristics of the clusters at successive steps and decide when you have an interpretable solution or a solution that has a reasonable number of fairly homogeneous clusters.

Consequently, the next question that one asks is how similar or different are the cases, meaning how close or not are the various clusters. For this purpose, one measure the distance between two objects. This distance is a measure of how similar or different are these two objects, whether those objects are cases or clusters. To visualize the representation of the distance at which clusters are combined, you can look at the Figure 2 which is called a dendrogram.

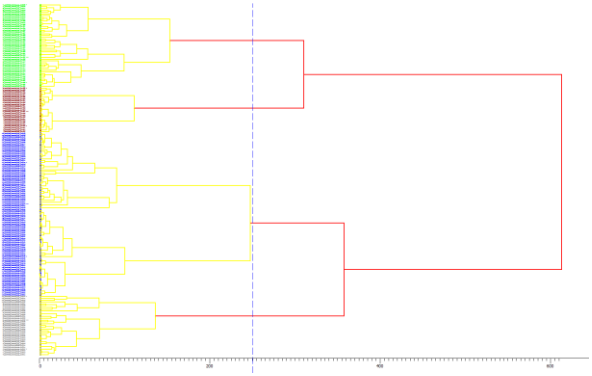

Figure 2: Dendrogram of the temperature variation of the compound $Mn_3V_2O_8$.

Example: Magnetic phase transition in $Mn_3V_2O_8$

Using the high intensity two-axis powder neutron diffractometer D20 at the Institut Laue Langevin, one can collect a large range of datasets within few hours. Here the temperature dependence of $Mn_3V_2O_8$ has been investigated giving rise to 196 patterns showing the various magnetic phase transitions present in this sample as function of temperature. An illustration of the results is shown in Figure 3. While visually, one can clearly notice 3 different regimes suggesting 3 different phases, in practice one would need to go through the 196 patterns to determine phase transitions temperature and identify the most characteristic patterns within this list. An alternative approach which does not require to treat the data is to carry out a cluster analysis on the whole dataset. For such cluster analysis, the clustering is done on the peaks present in the patterns and the comparison is done using the position and the intensity of those peaks.

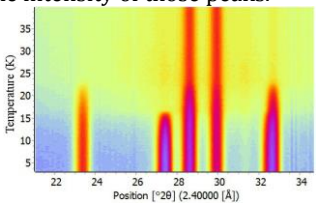

Figure 3: Isoline plot of the temperature dependence of $Mn_3V_2O_8$.

The cluster analysis suggests the existence of 4 main clusters which are illustrated in the Figure 4. In this figure, we have coded also the temperature variation. The higher the temperature, the larger are the spheres in Figure 4.

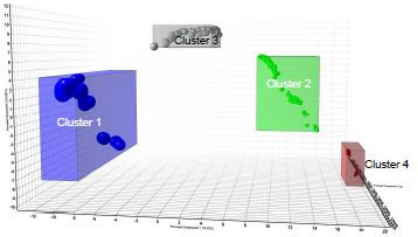

Figure 4: Principal component analysis of the temperature behavior of $Mn_3V_2O_8$

From Figure 4, one can notice that a possible clustering of the temperature behaviour of $Mn_3V_2O_8$ results in 4 clusters which can be classified also as function of temperature. The

cluster 1 corresponds to the high temperature phase of $Mn_3V_2O_8$ (paramagnetic phase) while the cluster 4 corresponds to the LT1 low temperature phase[7].

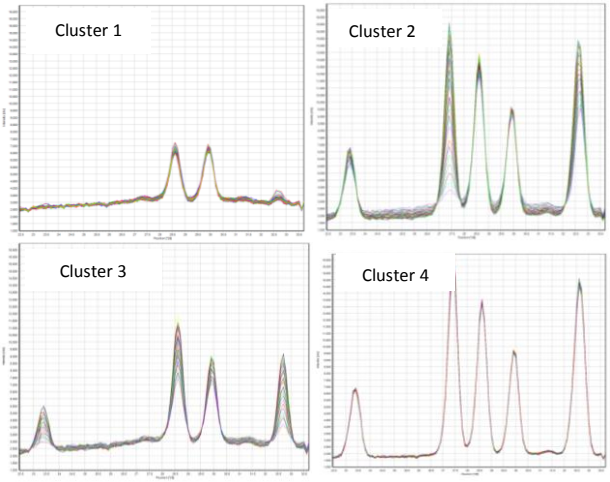

Figure 5: Illustration of the 4 clusters of the temperature behavior of $Mn_3V_2O_8$.for the clustering shown in the dendrogram of Figure 2.

For the assignment of the other 2 clusters, one can make a comparison using the prototype scan of each cluster. The prototype scan is the scan the most representative of each cluster. This type of information is given once the analysis is done using HighScore. The comparison between the 4 prototype scans is shown in Figure 6.

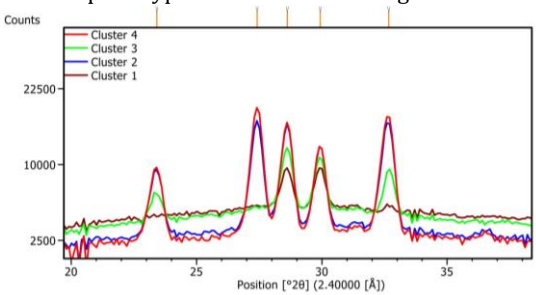

Figure 6: Comparison between the 4 prototype scans of the 4 clusters for $Mn_3V_2O_8$.

From Figure 6, one can notice that clusters 2 & 3 differs by the presence of one reflection at about 27° characteristic of the HT1 phase[7]. Consequently, cluster 3 can assigned to the HT1 phase[7]. The change from cluster 2 to 3 seems actually marginal and thus could be merged together. In Figure 4, the temperature dependence is also plotted and thus we can look determine the transition temperature between the HT1 phase and the paramagnetic phase which is found to be around 23 K in good agreement with the literature. The same holds for the transition from LT1 to HT1 which corresponds to the transition from cluster 2 to 3. So using a cluster analysis on this large set of diffraction data allow us within few minutes to determine the transition temperature between the various phases and identify the most representative datasets which are representative of each phases. This is a significant gain in time and really helpful in further processing of the data.

Partial Least Square Regression

The goal of PLSR and associated methods is to predict or analyze a set of dependent variables from a set of independent variables. PLSR is particularly useful when one needs to predict a set of dependent variables from a very large set of independent variables. Initially this method was used in social sciences then later on in chemometrics. In the PLSR method, you have I observations which are described by K dependent variables which are stored into a matrix named **Y** which is defined as **Y** = *I* x *K*. The values of J predictors (properties of interest e.g. acidity, temperature, Fe^{2+} content, etc) measured on these I observations are collected in the matrix **X** which is defined by **X** = *I* x *J*. The goal of the PLSR method is to predict **Y** knowing **X**. Using simple linear regressions, in a fully defined system (= N unknowns for N equations), it is possible to solve it an easy way. The advantage of PLSR is that it can analyze data which are strongly correlated (in mathematical definition, strongly collinear), noisy and incomplete.

To solve such sets of equations, there are several approaches. One possibility is to eliminate some predictors (properties of interest) but this is usually not what is desirable. The other possibility is to carry out a principal component analysis of the **X** matrix and then use these principal components of **X** (i. e. eigenvectors) as regressors of **Y**. In order words, in the analogy with a linear regression with an equation of the type y = af(x) + bg(x) +ch(x) where f(x), g(x) and h(x) are known functions of x, you are determining the coefficients a, b, and c. The eigenvectors of the **X** matrix are the a, b, c coefficients from which you can calculate any value of **Y** for a given value of **X**.

However, the determination of these eigenvectors is only an approximation in the case of the PLSR method due to the strong correlation, noisiness and incompleteness of the data. Thus this method gives rise to a prediction of the **Y** property for a given **X** composition for instance. The advantage of this method which has been popularized in chemometrics is the capability of prediction giving rise to possibilities to be explored.

Example: Quantification of Fe^{2+} in iron ores using diffraction data

Iron sinter is an important feedstock material for the steel industry. Due to increased quality requirements and the need to reduce energy consumption and CO_2 emissions, the phase composition and chemistry of iron sinter requires faster and innovative analysis methods. The Fe^{2+} content in iron sinter is a major factor in making iron and steel. Traditionally wet chemistry is used to determine the Fe^{2+} content but this is time consuming (several hours) compared to carry out powder diffraction experiment. For determining

The main minerals in iron sinter are hematite ($Fe_2^{3+}O_3$), magnetite ($Fe_2^{3+}Fe^{2+}O_4$), wuestite ($Fe^{2+}O$), larnite (Ca_2SiO_4), silico ferrites of calcium and aluminium (SFCA, occurring in different modifications, SFCA and SFCA-I) and a glass phase (amorphous). Figure 7 illustrates an example for a full pattern Rietveld quantification of one iron sinter sample. Besides the crystalline phases, the amorphous content can be also determined.

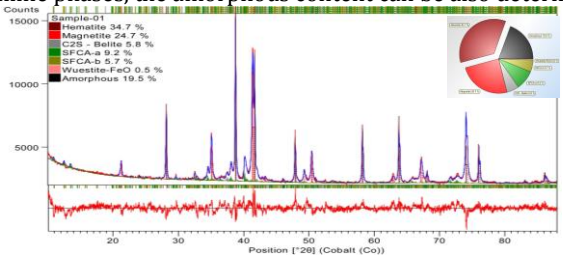

Figure 7: Example of Rietveld quantification of one iron sinter sample. Measured scan, calculated pattern, difference plot and quantitative phase composition.

In order to determine the content of Fe^{2+} from the raw diffraction data, we measured 48 iron sinter samples with known amount of Fe^{2+}. So here I = 48 while J = 1 and is given by the amount of Fe^{2+}. The definition of K is much more complicated as it contains many parameters: the various phases being in presence, their amount, their chemistry, etc. Using the PLSR algorithm implemented in the HighScore suite6, we derived a prediction model for the amount of Fe^{2+} in iron sinters. Using this prediction model, we can compare the results obtained by wet chemistry and the PLSR method. This comparison is shown in Figure 8 for 35 samples. It clearly shows that the powder diffraction in combination with PLSR is a robust and fast solution to determine the amount of Fe^{2+}.

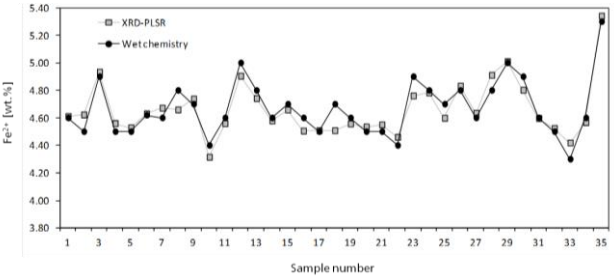

Figure 8: Comparison of wet chemistry and PLSR results for Fe^{2+} in iron ore sinter[4].

Conclusion and outlook

Cluster and PLSR methods applied to crystallography are still methods in their youth although already commonly used in other sciences. One of their major advantages which maybe the reason also for their insufficient use is their ability to treat a large amount of data. While these methods start to growth to understand complex industrial process, their use in academic research is still in its infancy but should be a powerful tool to design new materials with given properties.

References:

[1]Hartigan, J. A. (1975). *Clustering Algorithms (Probability & Mathematical Statistics)*. John Wiley & Sons.

[2]Wold, H. (1966). *Estimation of principal components and related models by iterative least squares*. In P.R. Krishnaiaah (Ed.). Multivariate Analysis. (pp.391-420) New York: Academic Press

[3] Kostov K. S., Moffat K., (2011) *Biophysical Journal* **100**, 440–449

[4]König U; Degen T., Norberg, N. Powder Diffraction (2014), **29** (S1), S78-S83

[5]Stegk T. A., Mgbemere H., Herber R.-P., Janssen R., Schneider G. A., (2009) *Journal of the European Ceramic Society* **29** 1721–1727

[6]Degen, T. Sadki, M., Bron, E., König, U., Nénert, G. (2014) *Powder Diffraction* **29** (S2), S13-18 and references therein.

[7]Clemens O., Rohrer J., Nénert G., (2016) *Dalton Trans.*, **45**, 156-171