A curious method for data from curious experiments:

Random projection with TOF-SIMS data from comet-relevant samples

K. Varmuza

Vienna University of Technology, Institute of Chemical Engineering, Austria; kvarmuza@email.tuwien.ac.at

Random projection (RP) [1] is a rather new method in chemometrics for dimensionality reduction [2]. In RP high dimensional data $\boldsymbol{X}(n \times m)$ are transformed to a score matrix

$$\boldsymbol{U}(n \times a) = \boldsymbol{X}(n \times m) \,.\, \boldsymbol{B}(m \times a) \qquad a << m$$

and the a loading/projection vectors in $\boldsymbol{B}$ given by appropriate random numbers. RP is a simple and fast technique, and may be an alternative to classical methods, especially for data sets with large n and large m, or for specific applications with limited computer resources such as space experiments.

The European Space Agency project *Rosetta* [3] will bring instruments near a comet for in situ measurements (launch was 2004, arriving at the comet is scheduled for 2014). One of the instruments is *Cosima*, a time-of-flight secondary ion mass spectrometer (TOF-SIMS) for the analysis of comet dust particles [4].

An equivalent laboratory instrument has been used for measurements of a target on which a grain of the mineral clinopyroxene was deposited; similar minerals are expected in cometary grains. In one experiment, a set of 49 spectra has been scanned at locations of a quadratic grid at distances of 100 μm horizontally and vertically. After some data reduction each spectrum consists of m = 8118 variables (the number of ions in m time bins), thus $\boldsymbol{X}$ has size 49 × 8118. A reduction to a = 5 RP projection scores gave a matrix $\boldsymbol{U}(49 \times 5)$. PCA of $\boldsymbol{X}$ and $\boldsymbol{U}$ show very similar results (Figure 1), and an excellent separation of the spectra from the target material and from the mineral.

The dimensionality reduction by RP is also promising for an automatic selection of relevant spectra (Figure 2).

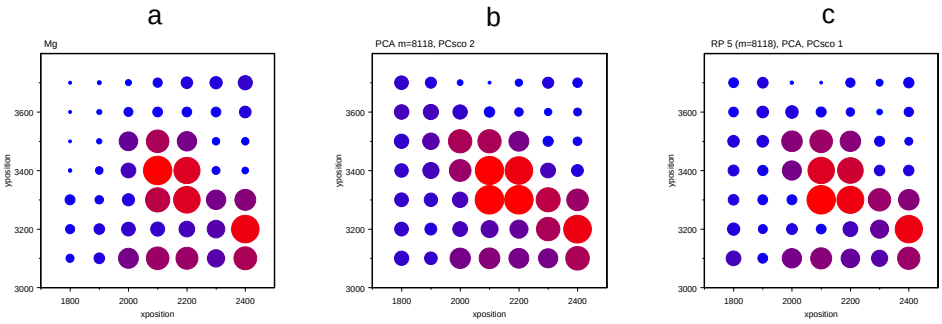

Figure 1. At the 49 locations of the 7 × 7 grid different measures are displayed by size and color of the symbol. The measures are:
(a) number of Mg^+-ions (considered as indicator of mineral clinopyroxene);
(b) PC2 score from a PCA with all 8118 variables (PC1 score does not show useful information);
(c) PC1 score from a PCA with only 5 RP projections.

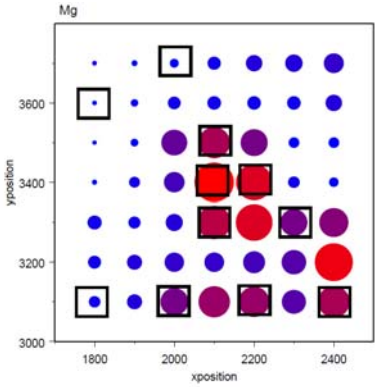

Figure 2. The measurement of TOF-SIMS spectra at many locations of a target may be limited by computational restrictions. The spectra (in this work 49) are measured sequentially. Assume it may be not possible to store all full spectra in memory or on hard disk. Then it would be useful to decide after each newly registered spectrum whether the new spectrum is a new spectrum type and should be stored or not. Furthermore, assume that memory size does not allow to store many (more than one) spectrum with full data. For such a situation a strategy for a sequential spectra selection was tested. Size and color of the circles are proportional to the Mg^+-ions reflecting the presence of the mineral. 5 RP projection scores have been used for the calculation of the Euclidean distances (spectra dissimilarities) between a tested spectrum and the spectra already stored. From 49 spectra a set of 11 "more or less unique" spectra (marked by squares) have been selected.

[1] Achlioptas D.: *J. Comp. Sys. Sci.* **66**, 671-687 (2003)
[2] Varmuza K., Filzmoser P., Liebmann B.: *J. Chemometrics*, **24**, 209-217 (2010)
[3] http://en.wikipedia.org/wiki/Rosetta_(spacecraft)
[4] Kissel J., et al.: In Schulz R., et al., Rosetta - ESA's mission to the origin of the solar system, p. 201-242, Springer, New York (2009)