Random Projection

for

Dimensionality Reduction

Applied to TOF-SIMS Data

Kurt Varmuza[1*], Peter Filzmoser[2],
Martin Hilchenbach[3], Jochen Kissel[3], Harald Krüger[3],
Johan Silén[4]

[1] **Vienna University of Technology, Institute of Chemical Engineering**
Laboratory for Chemometrics, Getreidemarkt 9/166, A-1060 Vienna, Austria
kvarmuza@email.tuwien.ac.at, www.lcm.tuwien.ac.at

[2] **Vienna University of Technology, Institute of Statistics and Probability Theory**
Wiedner Hauptstrasse 8-10, A-1040 Vienna, Austria
P.Filzmoser@tuwien.ac.at, www.statistik.tuwien.ac.at/public/filz

[3] **Max Planck Institute for Solar System Research**
D-37191 Katlenburg-Lindau, Germany
hilchenbach@mps.mpg.de, cometkissel@arcor.de, krueger@mps.mpg.de,
www.mps.mpg.de/en/

[4] **Finnish Meteorological Institute**
Erik Palmenin aukio 1, FIN-00560 Helsinki, Finland
johan.silen@fmi.fi, www.ilmatieteenlaitos.fi/research_space/space_29.html

* Presenting author

Poster Presentation:
12th Int. Conference on Chemometrics in Analytical Chemistry, CAC 2010
October 18–21, 2010, Antwerp, Belgium

Introduction

Random projection (RP) is a linear method for a projection from a high-dimensional space into a low-dimensional space, using projection vectors (loading vectors) with random numbers as vector components.

RP is based on the fact that high-dimensional vectors with randomly chosen vector components are very frequently "almost orthogonal". Some RP methods apply orthogonalization of the projection vectors [1-3]. RP makes a statistical unbiased sampling of the high dimensional space into a tractable low dimensional one.

RP projection uses loading vectors which are independent from the data. RP is very simple and fast in computation, and is appropriate for large data sets. Successful applications have been reported for clustering and classification of textual documents and image data [4-5]; recently, RP was introduced into chemoinformatics and chemometrics [6].

RP is especially useful in situations with severe hardware restrictions or with huge data sets. Thus, this study was partly motivated by planned TOF-SIMS measurements of dust particles near a comet in 2014 (ESA mission Rosetta [7], Cosima instrument [8]).

We report on preliminary experiments using RP

mainly for an automatic selection of relevant spectra in TOF-SIMS scanning experiments without the need for storing many full spectra.

Method

Generation of Random Projection Vectors

The components of projection vectors in RP are random numbers from a distribution with a mean of zero, e.g.

- normally distributed numbers from N(0, 1),
- uniformly distributed numbers from U[-1, +1] (this work)
- fixed values randomly selected from {-1, 0, +1}

- The projection vectors are normalization to unit length.
- For saving data memory, the components of RP vectors can be generated element-wise applying a user-controlled random number generator. Thus, each projection vector can be defined by only 3 numbers: a *seed*, the *dimensionality*, and a *normalization factor*. RP vectors can be defined before any data are available.

Software

All computations were performed within R [9].
The free R-library "chemometrics" [10] has been used.

TOF-SIMS

The laboratory TOF-SIMS instrument (time-of-flight secondary ion mass spectrometer) *Cosima RS* has been used, which is equivalent to the instrument [8] that is installed on Rosetta [7].

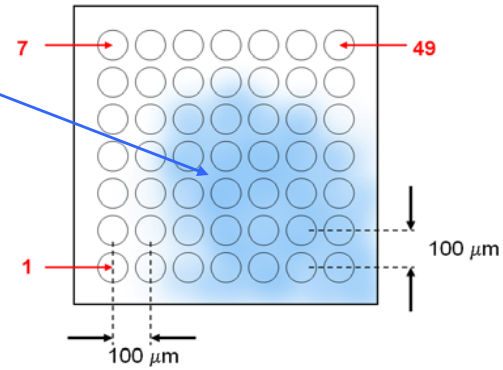
Ion gun: $^{115}In^{+}$; 8 keV; 5 ns pulses with 1 kHz
Sample measurement area: ca 50 μm diameter
TOF: 1 kV accelerating voltage for secondary ions; ion reflector;
$m/\Delta m$ = 1500; *m/z* 1 - 4000; ca 60,000 time bins per spectrum

Results

TOF-SIMS Experiment

A mineral grain (*clinopyroxene*, Al,Mg silicate), diameter ca 500 μm, has been deposited on a silver target.

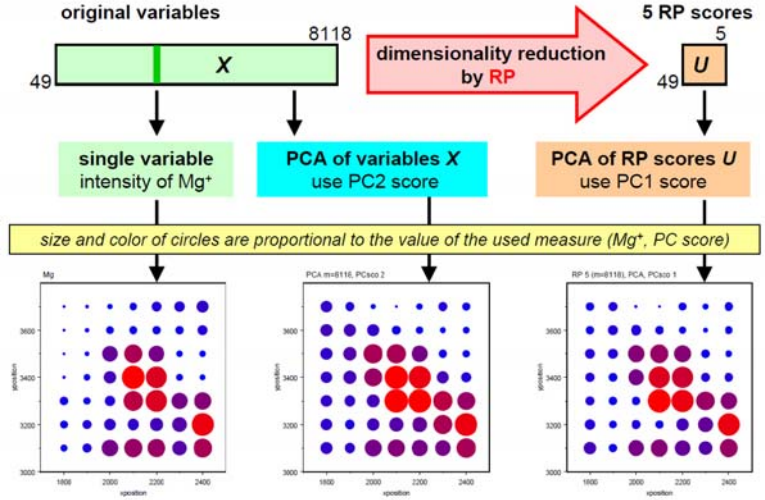
The blue area indicates the position of the mineral grain and contaminations of the mineral at the target



n = 49 mass spectra have been measured at grid positions 7×7.

After some preprocessing a mass spectrum (*m/z* 1 ... 113) consists of m = 8118 variables (= no. of detected ions in 4 ns intervals (bins, mass intervals).

PCA with original variables and with RP scores



- In this demo experiment is known that Mg^{+} ions give a characteristic signal.
- RP with only 5 projection scores preserves information target/mineral.

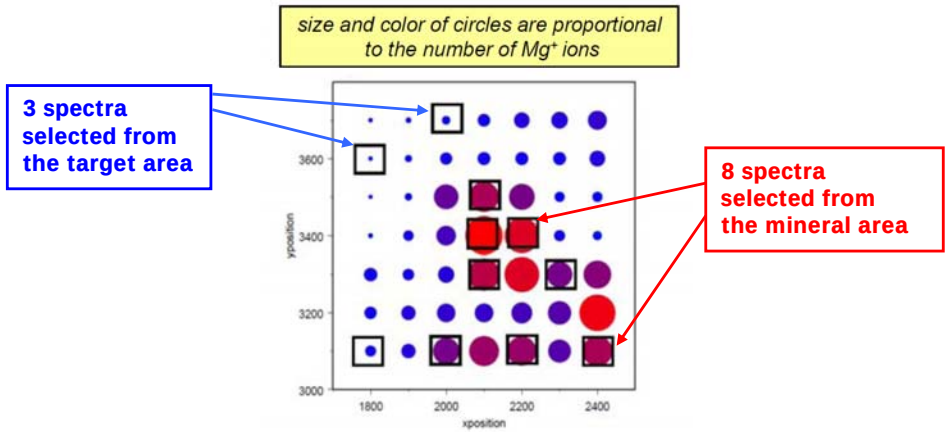
Results

Sequential selection of characteristic spectra in a pre-defined sampling area - using RP scores

(0) Define 5 - 10 RP loading vectors to be used.
(1) Choose randomly a position within the sampling area.
Measure a mass spectrum at this position.
(2) Transform the spectrum into RP scores.
(3) Store the first measured spectrum (as RP scores) as a selected spectrum.
(4) Choose randomly a position not yet used within the sampling area.
Measure a mass spectrum at this position.
(5) Transform the spectrum into RP scores.
(6) Compare the actual spectrum (RP scores) with all already selected spectra (RP sores) by e.g. the Euclidean distance or another dissimilarity criterion.
(7) If all calculated dissimilarity criteria are above a defined ***threshold***, then the actual spectrum is considered to be characteristic and is stored (as RP scores).
(8) Continue with step (4) until a pre-defined number of measuring positions has been reached.

The ***threshold*** is critical; it influences the number of selected spectra.

Result



Conclusions

Dimensionality reduction of TOF-SIMS spectra by RP (from ca 8000 dimensions to 5) preserved information for the recognition of a mineral grain at a silver target.

A series of high-dimensional vectors can be transformed sequentially into a few RP scores fast, and without the need of storing high-dimensional loading vectors. Individual loadings can be computed reproducibly "on the fly".

- Random Projection (RP) appears as a useful **"niche approach" for dimensionality reduction.**
- RP is especially suitable for very large data sets (with the data structure unknown in advance), and/or severe restrictions for hardware and software, and/or severe restrictions for computing time.

[1] S.S. Vempala. The random projection method. *Series in Discrete Mathematics and Theoretical Computer Science*, vol. 65, American Mathematical Society, Providence, RI (2004).

[2] S. Dasgupta. Experiments with random projection. In *Proceedings of the Sixteenth Conference* on *Uncertainty in Artificial Intelligence*, p. 143-151. Morgan Kaufmann Publishers Inc., San Francisco, CA (2000).

[3] X.Z. Fern, C.E. Brodley. Random projection for high dimensional data clustering: A cluster ensemble approach. In *Machine Learning - International Workshop* (2003).

[4] E. Bingham, H. Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2001).

[5] N. Goel, G. Bebis, A. Nefian. Face recognition experiments with random projection. In *Proceedings SPIE* (2005).

[6] K. Varmuza, P. Filzmoser, B. Liebmann: *J. Chemometrics,* 24, 209-217 (2010).

[7] R. Schulz et al. (eds.): *Rosetta: ESA's mission to the origin of the solar system*, Springer, New York (2009).

[8] J. Kissel J., et al.: *Space Science Reviews,* 128, 823-867 (2007).

[9] R. A language and environment for statistical computing. R Development Core Team, Vienna, Austria, 2008. www.r-project.org.

[10] K. Varmuza, P. Filzmoser. *Introduction to multivariate statistical analysis in chemometrics*, CRC Press, Boca Raton, FL (2009). Info: www.lcm.tuwien.ac.at/cm-book.pdf.

Acknowledgment: Austrian Science Fund (FWF), project P22029-N13

Abstract of Poster

Random projection for dimensionality reduction - applied to TOF-SIMS data

Kurt Varmuza[1], Peter Filzmoser[2], Martin Hilchenbach[3], Jochen Kissel[3], Harald Krüger[3] and Johan Silén[4]

[1] Vienna University of Technology, Institute of Chemical Engineering, A-1060 Vienna, Austria
[2] Vienna University of Technology, Institute of Statistics and Probability Theory, A-1040 Vienna, Austria
[3] Max Planck Institute for Solar System Research, D-37191 Katlenburg-Lindau, Germany
[4] Finnish Meteorological Institute, FIN-00560 Helsinki, Finland

Random projection (RP) is a method for dimensionality reduction and is rather new in chemometrics. In RP, high dimensional data $\boldsymbol{X}(n \times m)$ are transformed to a score matrix $\boldsymbol{U}(n \times a) = \boldsymbol{X}(n \times m).\boldsymbol{B}(m \times a)$ with $a<<m$. The a loading/projection vectors in $\boldsymbol{B}$ are defined by appropriate random numbers; $\boldsymbol{B}$ need not to be stored but can be easily generated element-wise. RP is simple and fast, and may be an alternative to classical methods for data with large n and m, or for applications with limited computer resources.

The ESA project *Rosetta* will bring instruments near a comet; launch was 2004, arrival and entering an orbit around 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 cometary dust particles. *Cosima* will collect cometary dust on metal targets, will investigate the exposed targets optically and will measure mass spectra. The available memory for storing full spectra is very limited.

An equivalent laboratory instrument has been used for measurements of mineral grains as expected in cometary material. In one of the experiments a clinopyroxene particle with a diameter of ca 300 μm was deposited on a silver target. A set of 49 spectra has been scanned at locations of a quadratic grid at distances of 100 μm horizontally and vertically. After data reduction each spectrum consists of m = 8118 variables which are the number of ions in 4 ns time bins, thus $\boldsymbol{X}$ has size 49 × 8118. PCA with RP score matrices $\boldsymbol{U}$ containing only 5 or 10 dimensions gave very similar scatter plots as PCA with $\boldsymbol{X}$; spectra from the target and from the mineral were well separated in the PCA score plots from $\boldsymbol{U}$. Furthermore, RP is promising for an automatic selection of relevant spectra measured sequentially, without the need of storing full spectra.

Abstract of Cited Paper

J. Chemometrics **24**, 209-217 (2010)

Varmuza K., Filzmoser P., Liebmann B.:

Random projection experiments with chemometric data

Random projection (RP) is a linear method for the projection of high-dimensional data onto a lower dimensional space. RP uses projection vectors (loading vectors) that consist of random numbers taken from a symmetric distribution with zero mean; many successful applications have been reported for high-dimensional data sets.

The basic ideas of RP are presented, and tested with artificial data, data from chemoinformatics and from chemometrics. RP's potential in dimensionality reduction is investigated by a subsequent cluster analysis, classification or calibration, and is compared to PCA as a reference method. RP allowed drastic reduction in data size and computing time, while preserving the performance quality.

Successful applications are shown in structure similarity searches (53 478 chemical structures characterized by 1233 binary substructure descriptors) and in classification of mutagenicity (6506 chemical structures characterized by 1455 molecular descriptors). Only in calibration tasks with low-dimensional data as in many chemical applications, RP showed limited performance.

For special applications in chemometrics with very large data sets and/or severe restrictions for hardware and software resources, RP is a promising method.