Geosci. Instrum. Method. Data Syst., 4, 45–56, 2015
www.geosci-instrum-method-data-syst.net/4/45/2015/
doi:10.5194/gi-4-45-2015

Geoscientific Instrumentation Methods and Data Systems
Open Access

COSIMA data analysis using multivariate techniques

J. Silén[1], H. Cottin[2], M. Hilchenbach[3], J. Kissel[3], H. Lehto[4], S. Siljeström[5], and K. Varmuza[6]

[1]Finnish Meteorological Institute, Erik Palmenin aukio 1, PB 503, 00101 Helsinki, Finland
[2]Laboratoire Interuniversitaire des Systèmes Atmosphériques (LISA), UMR7583 – CNRS, Université Paris Est – Créteil (UPEC), Université Paris Diderot (UPD), 61 Avenue du Général de Gaulle, 94010 Créteil, France
[3]Max Planck Institute for Solar System Research Justus-von-Liebig-Weg 3, 37077 Göttingen, Germany
[4]Tuorla Observatory Dept of Physics and Astronomy University of Turku, 21500 Piikkiö, Finland
[5]Department of Chemistry, Materials and Surfaces, SP Technical Research Institute of Sweden, Borås, Sweden
[6]Vienna University of Technology, Department of Statistics and Probability Theory, Wiedner Hauptstrasse 7/107, 1040 Vienna, Austria

Correspondence to: J. Silén (johan.silen@fmi.fi)

Received: 3 May 2014 – Published in Geosci. Instrum. Method. Data Syst. Discuss.: 15 August 2014
Revised: 31 October 2014 – Accepted: 24 December 2014 – Published: 27 February 2015

Abstract. We describe how to use multivariate analysis of complex TOF-SIMS (time-of-flight secondary ion mass spectrometry) spectra by introducing the method of random projections. The technique allows us to do full clustering and classification of the measured mass spectra. In this paper we use the tool for classification purposes. The presentation describes calibration experiments of 19 minerals on Ag and Au substrates using positive mode ion spectra. The discrimination between individual minerals gives a cross-validation Cohen κ for classification of typically about 80 %. We intend to use the method as a fast tool to deduce a qualitative similarity of measurements.

1 Introduction

The TOF-SIMS (time-of-flight secondary ion mass spectrometry) instrument Cometary Secondary Ion Mass Analyser (COSIMA), on Rosetta provides high quality measurements of chemically complex cometary dust (Kissel et al., 2007). The instrument is designed to collect dust on target substrates. An ion beam with the diameter of about 50 µm bombards the collected dust by indium ions. The secondary ions created are counted. The instrument is able to detect and measure the mass of individual ions at an intermediate mass resolution $m/\delta m \sim 1400$ at mass 100. The top few monolayers of the dust collected are probed by indium ions. The instrument is sensitive enough to detect individual molecules and is capable to even provide the molecular structure.

From the knowledge of composition of the dust from cometary sources (Stephan, 2008), it seems that we will encounter both inorganic and organic substances in mixtures. The dust-collecting substrates exposed at comet 67P/Churyumov–Gerasimenko are designed to catch dust grains, impacting at low speed, with a high probability (Hornung, 2014). The dust grain size is an a priori unknown factor. Dust smaller than visible by the optical microscope, Cosiscope, will be detectable by the TOF-SIMS instrument. Therefore we need to establish methods by which we can identify a grain on a visually seemingly empty background. The optimum size of the dust grains to be measured would be the same size as the beam diameter.

The purpose of this presentation is to investigate how well we can identify a set of selected dust grains, from the substrate background, using a few different sets of data and applying multivariate statistical analysis enhanced with the relatively recent random projection method (Bingham and Mannila, 2001), and developed for our purpose (Varmuza et al., 2011).

Traditional methods of analyzing mass spectra rely on studying details of individual mass lines or groups of lines. This type of investigation will be made by us only as a subsequent second step. In the early phases of the mission, when a small number of spectra are available, we need to be able to

establish a link between the observed spectra and those measured earlier in the laboratory. This task has to be based on estimating the similarity between spectra and as it is a challenging task, we first present an overview of methods traditionally utilized for this purpose. Multivariate approaches in most cases require a reduction of dimensionality by selecting specific features. We avoid doing this selection manually either by using a peak list derived from a spectrum or by applying the method of random projections. This technique gives additional advantages as it reduces the computational complexity of the classification task. We demonstrate how to use the method and show that it is a useful tool.

1.1 Data analysis and classification

The lack of an overall applicable quantitative model for the ionization processes in SIMS, Gross (2004), suggests that we approach the data set in an exploratory manner. The interest in data mining techniques is the subject of intense research and the amount of published material is large (Dasu and Johnson, 2003). Finding features and patterns in a set of measurements, does not necessarily provide understanding (Keogh et al., 2004). We therefore regard the chemometrics methodology to be the only viable basis for our analysis attempts (Varmuza and Filzmoser, 2009). The amount of information hidden in any given spectrum is large. Therefore it is sometimes an advantage, and sometimes simply necessary, to reduce the dimensionality of the data before applying multivariate statistical tools (Varmuza et al., 2010).

From a general point of view, creating understanding implies reducing dimensionality of observations. Commonly, linear methods, like principal component analysis (PCA), have been used to find structure in a data set. For our purpose it is not clear how well this would be valid and applicable. A more general approach is provided by clustering methods, which can be shown to cope even with non-linear situations and embedded subspaces (Roweis and Saul, 2000). In some sense this understanding is accessible best by using some random sampling strategy (Candes, 2006).

Interestingly, it has been shown that the commonly used tools like the bilinear PCA and K means clustering methods, can be understood from a common base (Ding and He, 2004). There is no guarantee that the data we collect are in any sense linear. When measuring a mixture of two compounds A and B and expressing the result in some suitable abstract representation space, the results ought to be found along the line joining these two observations in the space containing both of them. This is an important issue when the dimensionality of the data is large. Each of our mass spectra contain some 10^5 numbers which can be reduced to a few dozen to hundreds of meaningful mass lines. This is still a large dimension where, in particular, clustering methods tend to become hard to use as the computational task becomes too costly.

When the number of parameters grows, the interpretation gets even harder, the curse of dimensionality. This problem has been investigated and quantitative bounds for the validity of estimates of principal components have been established (Nadler, 2008). This makes it possible to create a clustering representation in a space of sufficient dimension to make the cluster reliable, i.e. contain all required information about the cluster members. New observations are brought into this map, and estimates are made about how of much the map changes as the population distribution is modified. To learn from the established representation, one needs to make a non-trivial reverse mapping, where observations are related to the interpretation. This task has also recently been studied (Monnig et al., 2013).

Other multivariate statistical methods and techniques like data mining can be added to these tools (Dasu and Johnson, 2003; Giudici and Figini, 2009). For the present presentation we will not go further into this direction.

When operating the instrument, we will encounter mixtures (Stephan, 2008). Because quantitative absolute measurements are not possible, we need to address the question of precision and accuracy in the present investigation. Bayesian methods (Broemeling, 2009) are robust and additionally provide knowledge about our possible ignorance of the experimental setup. This method will later serve as a first step in the data-processing chain of events (Lehto, 2014).

1.2 Considerations when operating the instrument

During the operational phase of the ROSETTA mission, measurement strategy decisions have to be made on the fast track. The TOF-SIMS device is an extremely sensitive instrument providing measurements of ions from the dust grain surface. Contamination, emerging through out gassing from the electronics and structural elements on the space craft, or other unknown sources, is of great concern. Contamination tends to be organic compounds that may migrate and cover the very small size dust grains collected during the target exposure.

Repeating measurements at some time intervals does therefore not necessarily give the same results. Some aspects of this is covered in a separate article (Hilchenbach, 2014). Contamination is a hard problem, as there are indications that comets are surrounded by dust with a size distribution such that small grains are much more abundant than large ones (McDonnell et al., 1991; Kolokolova et al., 2004; Tuzzolino et al., 2004). This results in a "contamination-like" influence on the measurements which might show up as a real measured signal. Being indeed of cometary origin it would contaminate the entire surface being investigated. The consequence of this is that contamination always is present and will forever be an inseparable part of the measurement.

We do have information about cometary matter from a number of sources (Tsou et al., 2004; Stephan, 2008; Mumma and Charnley, 2011), but few data are collected in a non-destructive manner in situ.

Our approach is to construct maps of some laboratory spectra based on multivariate statistical techniques. Standard methods of principal component analysis are applied directly to peak lists derived from full mass spectra. This list represents a first large dimensionality reduction, which is sufficient for most purposes. It does not evaluate the influence of the ever-present signal noise residing between individual mass lines, or multiply ionized ions at fractional masses m/z.

As pointed out before, there are limits on dimensionality reduction, which must not be exceeded. The dimensionality reduction method, random projections (RPs), introduced a decade ago (Bingham and Mannila, 2001), can be shown to possess a number of universally attractive properties (Candes, 2006). The problem of outliers in a principal component analysis as described by Filzmoser et al. (2009) is eliminated by design in the method of random projections, which is in some sense optimal (Candes, 2006).

It may be even possible to construct maps for nonlinear cases, using this method (see Roweis and Saul (2000) or Monnig et al. (2013) for a good overview of the mapping and its inversion).

During the operational phase of the Rosetta mission, there is a need to make fast decisions on what to measure, based on possibly very limited information. In this context, a quick analysis where measured spectra can directly be related to any laboratory results or to previous in situ measurements is important. Support for this kind of reasoning is an additional goal for this work.

2 Method

Principal component analysis, PCA, is a well known bilinear method. It simply establishes an orthogonal coordinate system in which the measured data can be expressed in a clear manner. Each direction established is independent of each other. The principal components (PCs) each tell about what features, in order of importance, are separating our individual measurements.

When using the PCA method, it is important to centre the data. For very high dimensional data it is reasonable to apply the method in two steps where the random projection (RP) method is first used; on the result, the PCA method is applied. The properties of RP essentially whiten the data and removes biases. It is important to use a dimensionality large enough to correctly preserve the data structure under consideration.

In mathematical terms we can express these statements as follows. We may regard each measurement as a row vector $\boldsymbol{d}$. The dimensionality is determined by the number N of parameters measured. The complete set of data is then given by the matrix $\mathbf{D}$. It consists of M measurements, rows, each N dimensional vectors $\boldsymbol{d}$.

A linear dimensionality reduction then consists of a projection of $\mathbf{D}$ using a transformation matrix $\mathbf{R}$ giving a transformed data set $\mathbf{X}$ with dimension $N \times K$ as

$$\mathbf{X}_{M\times K} = \mathbf{D}_{M\times N} \cdot \mathbf{R}_{N\times K}.$$

Random projection is suitable to map even a very high dimensional data set into a reduced dimension which still retains the most important features of the original data but is small enough to be numerically tractable. The selection of transformation matrix is of critical importance. If we by ad hoc reasoning decide that only certain lines in the mass spectrum are important, we actually choose a very limiting simple form of projection. On the other hand, if we randomly select different features from our data set multiple times, we effectively ensure that we do not loose anything important. This can be proven mathematically in a strict manner (Candes, 2006). This reduction can be understood relatively easily by realizing that each projected dimension provides a view of the full original data set viewed from some particular direction. Therefore, a random projection into 10 dimensions provides 10 complete views of the entire data. For our task we select each element of $\mathbf{R}$ from a standard normal distribution $N(0, 1)$. Using an optional normalization $N(0, 1)/\sqrt{N}$ makes the columns of $\mathbf{R}$ approximately unit length and the average of $\mathbf{X}$ is approximately zero, fulfilling the statistical requirements imposed by PCA.

The results for PCA obtained from the reduced space, can be directly related to the original high dimensional data. Therefore an advanced kind of information extraction is utilized in a robust manner (Candes, 2006). We can actually compute the PCA components reliably by first projecting the data set by RP and then apply standard PCA.

When the number of samples, i.e. spectra, grows, the accuracy of the PCA method decreases. For this case, clustering approaches for classification can be shown to perform better than simple PCA (Varmuza and Filzmoser, 2009).

We describe a set of calibration measurements of 19 minerals measured by the instrument. A separate paper, (Krüger, 2014), will discuss the sample preparation and details of the manual spectral evaluation. In this paper we are interested in constructing a method which would accurately and rapidly relate a new measurement to previously measured spectra. It should ideally be able to indicate the presence of mineral mixtures or clearly indicate nonconformity of something yet not observed.

A typical measurement sequence consists of visually, using Cosiscope, identifying a grain or a substrate area to be analyzed (Fig. 1). This is covered by a 5×5 matrix of measurements (Fig. 2). In most cases some of the measurements will hit a mineral grain and in some cases the substrate only. From the 25 spectra collected, we can derive typical properties of individual spectra and their variability. A principal component analysis loading vectors also pinpoint the typical features identifying the spectra. In some cases the discriminating signatures can be very small. Often therefore heuristic arguments are used to look at selected subsets of the spectra. This selection process effectively increases the signal to

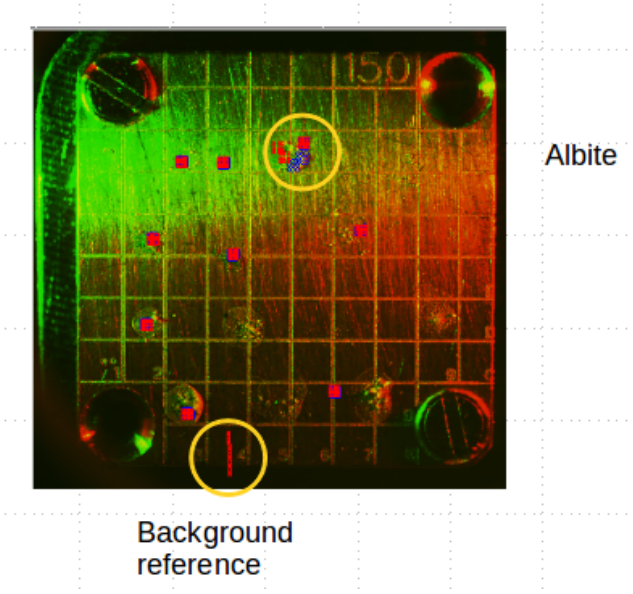

Figure 1. A substrate to collect dust. This one contains multiple mineral samples. In the middle at the top the albite mineral is deposited and a 5×5 matrix of closely spaced measurements is made. At the bottom is a straight vertical matrix of measurements on a clean area of the substrate which serves as a background reference.

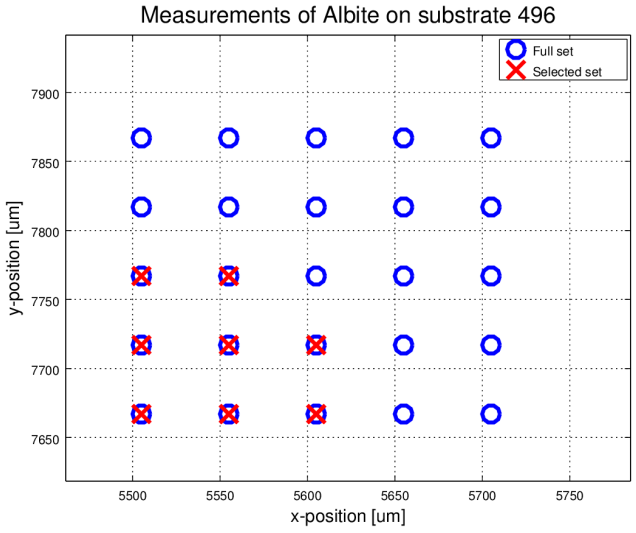

Figure 2. Picture showing the position of measurements at the mineral albite. The beam radius corresponds to the distance between the indicated positions. The marked squares indicate the hand picked positions on the edge of the mineral sample. As the sample is placed by wet deposition, it is likely that the other positions are contaminated by the same mineral substance in a possibly diluted form.

noise ratio. In this work we rather rely on using the full spectral data and a dimensionality reduction by random projection. This method being in some sense quite optimal, manages to reduce the dimensionality of the data studied. The resulting reduced data set is made small enough to be able to apply standard multivariate techniques.

We perform this analysis for two sets of data. First for a substrate with several deposited minerals of which we study 4, to establish what the reliability is of separating background and minerals. As an option we may use a set of 40 background additional spectra measured at the substrate in a position free from contamination.

For the second set we use all data as one uniform data set. To be able to compare spectra, we normalize spectra by taking the square root of the ion counts of individual bins or mass peaks and then normalize the result into unit vectors using an L2 norm, i.e. Euclidean norm. This corresponds to scaling spectral components by their standard deviations. Computations using the methods described in Nadler (2008), show that typical noise levels in signals are so small that several dozens of principal components are statistically significant and exceed the signal noise power. Therefore we may indeed utilize even relatively high order principal components if the particular component responds to a proposed question. As the principal components are statistically orthogonal, we can in some sense treat them as independent probes into the data set. This corresponds to the popular technique of manual feature selection like using specific mass line amplitudes or ratios between amplitudes.

It can be shown that computing PCA components from the data projected by a random projection gives the same results as doing it from the original data. The RP technique makes it possible to compute the clustering of even the complete set of calibration measurements. For the sake of demonstration we have used it only on a few thousand spectra investigating the properties of wet deposited minerals.

The probabilities for a correct identification are computed by the octave program (Eaton et al., 2009), function *xval* (Duda et al., 2001; Schlogl et al., 2007). It allows a wide variety of classification methods to be used. It also provides a statistically valid cross validation. Linear discriminant analysis of various flavours give comparable results.

During the interactive discussion of this paper, one of the reviewers pointed out several additional useful facts relevant to this presentation. The PCA method being a bilinear method, connects the data and the parameter spaces. The properties of this connection in the presence of signal noise has been described in in Paatero and Tapper (1993) and Paatero and Hopke (2003) and it may open up some additional research opportunities.

3 Data sources

For this study, we use two separate sets of data.

The first set is peak lists derived from complete positive ion mode spectra using a heuristic method providing a separation of mass peaks into inorganic and organic components. This minimal peak-list data set is produced on board the

spacecraft by the instrument software and is transmitted to ground regardless of telemetry constraints.

The instrument autonomously estimates mass line intensities for integer masses up to 300 Da. For higher values, an additional 30 numbers describe integrals of mass line intervals. For the mass range 301 to 400, intervals of 10 masses are used. The next 401 to 600 gives intervals of 20 followed by intervals of 50 up to mass 1000. The last two ranges are masses from 1001–1200 and finally above 1200. For each interval the total count of ions is given.

A further useful feature is introduced by a heuristic division of masses into organic and inorganic components based on an empirical rule commonly used in mass spectrometry (F. Krüger, personal communication, 1992). The method is based on the fact that hydrogen is a common element in organic compounds that has a mass slightly above 1 Da. This causes the organic ions to generally have a total mass slightly larger than the closest integer mass while the inorganic compound tends to stay below this value.

The second set of data is complete positive ion mode mass spectra, each containing 2^{17} numbers of ion counts in bins representing ion flight times.

This selection of data is motivated by the fact that it is generated autonomously in the instrument and transmitted to ground before the full spectrum. In addition a refined analysis of the full spectrum could later support this approach in more quantitative manner (Lehto, 2014).

Finally it is worth mentioning that a random projection could be used on the full spectrum and serve as a compressor or whitening step followed by the methods used in this presentation. The advantage of this approach is that a cross validation of the clustering results is possible between the full spectra and the limited mass line driven representations.

Database

A database provides fast and easy access to the data. Three different instrument models are accessible, namely the laboratory reference model, the flight model at the comet and a second laboratory reference model in France. More than a hundred thousand spectra are available. A wide variety of substances, both organic and inorganic, have been measured in positive and negative ion modes.

For the purpose of this study we use data from the database covering the substrates and minerals as shown in Table 1.

The samples studied in this presentation are listed in Table 2. Calibration experiments validating the measurements of organics are also described separately in Le Roy (2014).

4 Constructing maps

We use multivariate analysis to construct a relation between observations and represent them projected on a map. As an example of constructing a map, we take a subset of measured peak-list data shown in Fig. 1 visually on a substrate. A 5×5 matrix of measurements has been made and they are in the circle as shown in the figure. The layout of the TOF-SIMS measurements, for this particular measurement matrix, albite, in detail shown magnified in Fig. 2. The events marked by $\times$:s are spectra assigned manually as mineral calibration samples. The rest are regarded as "background". The minerals are deposited as a droplet of suspension; thus, it is probable that the sample is not sharply distributed, as anticipated, but rather as diffuse, spread-out contamination. This might be analog to the dust collection at the comet where a size distribution of dust is encountered and statistically distributed over the collector substrate. For convenience, the average of eight peak-list spectra for albite is shown in Fig. 3.

Table 1. Layout of minerals on substrates. In particular we use substrate 496 to validate classification of four minerals and background.

SUB	Minerals deposited			
4AF	Calcite	Dolomite	Sphalerite	
4B0	Corundum	Ilmenite	Magnetite	Richterite
41D	Clinopyroxene			
41E	Orthopyroxene			
41F	Olivine			
48B	Plagioclase			
420	Forsterite			
421	Sulfide			
422	Smectite			
496	Albite	Fayalite	Hyperstene	Orthoclase

There are other minerals deposited locally on the target as well, measured using a similar matrix scan. At the bottom of the substrate is a vertical set of 40 spectra in a region that should be devoid of any mineral samples (see Fig. 1).

The final set of measurements in positive ion mode consist of 25 spectra each for albite, fayalite, hypersthene and orthoclase to which a background reference of 40 spectra is added. We use the peak lists of this set of data consisting of 300 mass lines each of inorganic and organic ions. We make no feature identification of any kind. Our data therefore consists of a 140×600 matrix, 140 ($= 25 \times 4 + 40$) spectra each with 600 mass lines. From this we compute the principal components. The score plot for the first two PCs is shown in Fig. 4. As is evident from the 3-D presentation of the first three PCs in Fig. 5, the clustering is not confined to any simple plane. Each of the four minerals nicely separates into independent clusters, despite their pairwise chemical similarity (see Table 2).

It is instructive to look at the cross validation of a classification analysis of the minerals without using the background as a reference. Taking our four minerals, 4×25 spectra, we postulate that they consist of two groups. The first group seems to originate from the visually observable mineral and the second group seem to come from a background of the

Table 2. Table showing the names, composition and substrates used for this investigation. Minerals are deposited on blank Ag or Au substrates either by suspension (S), by pressing (P) or heterogeneous (H), see last column. ID refers to substrate ID fond in the COSIMA data base. Fayalite is a heterogenous suspension. For calcite and dolomite, carbon was not measured. For richterite and smectite, hydrogen was not measured; for magnetite, oxygen was measured as FeO. Details on the measured compositions and target preparations are discussed in Krüger (2014).

Mineral	Formula	Measured	Target	ID	Dep.
Albite	$NaAlSi_2\ O_2$	$NaAlSi_3O_8$	Ag blk	496	S
Calcite	$CaCO_3$	Ca_3O_3	Ag blk	4AF	S
Corundum	Al_2O_3	Al_2O_3	Ag blk	4B0	S
Clinopyroxene	$CaMgSi_2O_6$	$Mg_{0.9}$, $Fe_{0.1}$, $Al_{0.1}Ca_{0.7}Si_{1.8}O_6$	Au blk	41D	P
Dolomite	$(Ca, Mg)(CO_3)_2$	$(Ca_3, Mg_{2.5}, Fe_{0.5})O_6$	Ag blk	4AF	S
Fayalite	Fe_2SiO_4	$Fe_{1.9}Si_{1.0}O_4$	Ag blk	496	S, H
Hyperstene	$(Mg, Fe)_2Si_2O_6$	$(Mg_{0.91}, Al_{0.06}, Fe_{0.36})Si_{1.29}O_3$	Ag blk	496	S
Ilmenite	$FeTiO_3$	$(Fe_{0.8}Mg_{0.2})TiO_3$	Ag blk	4B0	S
Magnetite	Fe_3O_4	$Fe_{2.5}O_4$ (O measured as FeO)	Ag blk	4B0	S
Nepheline	$(Na, K)AlSiO_4$	$(Na_{0.6}, Ca_{0.3})$, $Al_1Si_1O_4$	Ag blk	497	S
Olivine	Mg_2SiO_4	$(Mg_{1.8}, Fe_{0.2})Si_{1.0}O_4$	Au blk	41F	P
Orthopyroxene	$(Mg, Fe)_2Si_2O_6$	$(Mg_{0.9}, Fe_{0.1}, Al_{0.1})Si_{1.9}O_6$	Au blk	41E	P
Orthoclase	$KAlSi_3O$	$(Na_{0.3}, K_{0.6}, Al_{1.0})Si_3O_8$	Au blk	496	S
Plagioclase	$(Na, Ca)(Si, Al)_4O_8$	$(Na_{0.5}, Ca_{0.5})(Si_{2.5}, Al_{1.5})O_8$	Ag blk	48B	P
Richterite	$Na[CaNaMg_5][(OH)_2[Si_8O_{22}]]$	$Na_{0.9}Al_{0.3}Ca_{1.6}(Mg_{4.7}, Fe_{0.4})[Si_{8.2}O_{24}]$	Ag blk	4B0	S
Smectite	$Ca_{0.25}(Mg, Fe)_3((Si, Al)_4O_{10})(OH)_2nH_2O$	$Ca_{0.2}(Mg_{0.1}, Fe_{2.7})[(Si_{4.3}, Al_{0.2})O_{12}]$	Au blk	422	P
Forsterite	Mg_2SiO_4	Mg_2SiO_4	Au blk	41F	P
Sphalerite	$(Zn, Fe)S$	ZnS	Ag blk	4AF	S
Sulfide	Fe_2S_2	FeS_2	Au blk	421	P
Background			Ag blk	496	

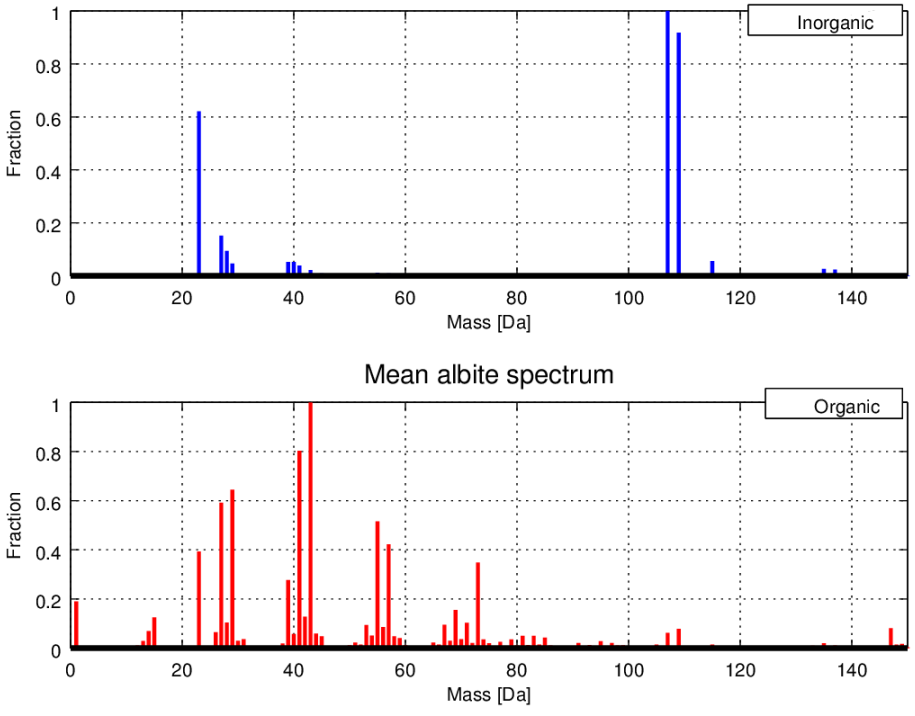

Figure 3. Example of peak lists. This is the average of peak lists for albite spectra. Organic and inorganic components are separated.

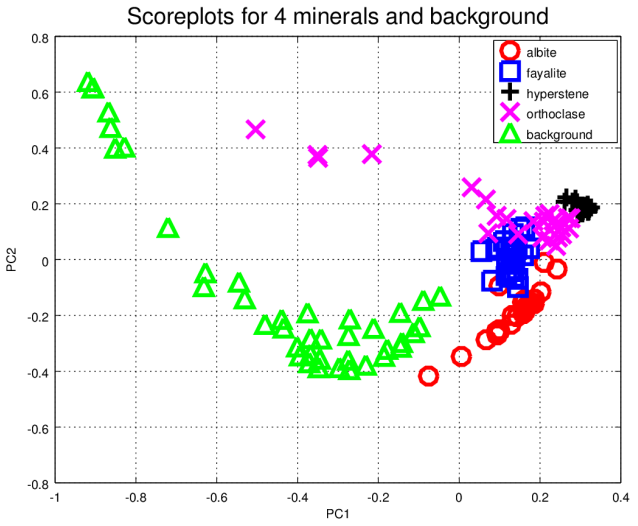

Figure 4. Score plot using the two first PCA components. The events represent the four minerals and the background for substrate SUB496. The scores from peak lists and RP results are essentially identical.

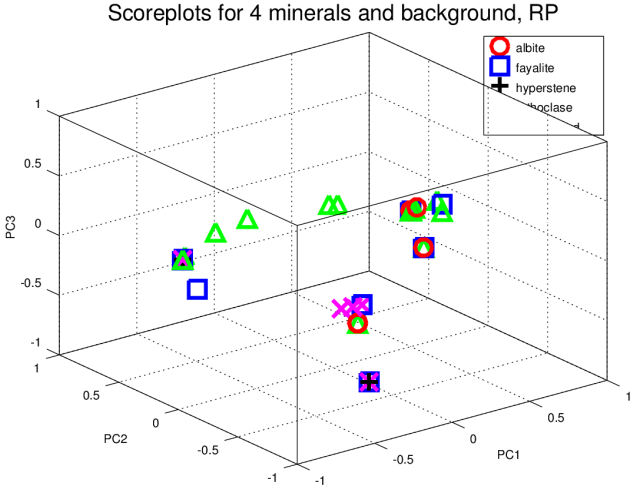

Figure 5. Score plot using the three first PCA components. The events represent the four minerals and the background for substrate 496.

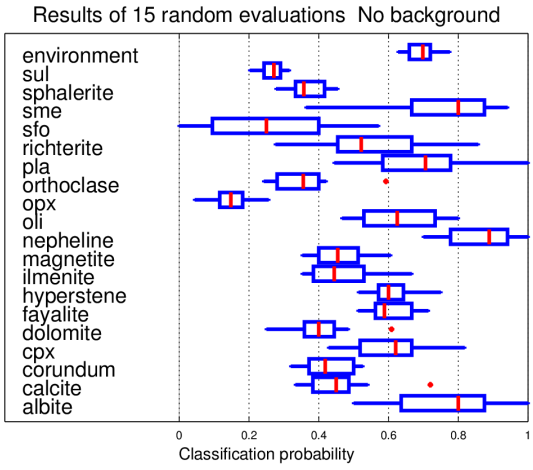

Figure 6. Cross validation accuracies showing the Cohen κ value computed for the data projected to a 100 dimensional space using random projections. For this case, background is regarded as those visually identified spectra surrounding measured minerals, marked as "environment".

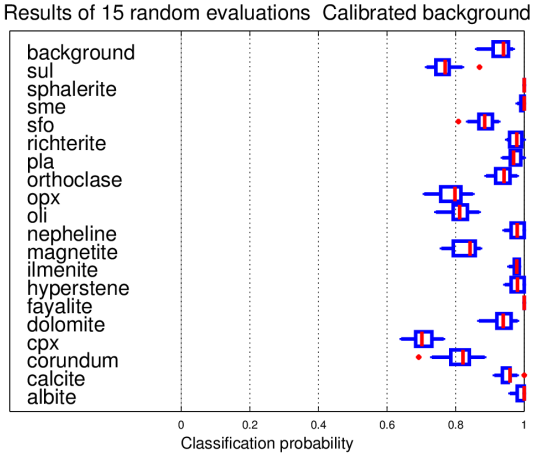

Figure 7. Cross validation accuracies showing the Cohen κ value computed for the data projected to a 100 dimensional space using random projections. For this case, a separate reference background measurement is made. When included the Cohen $\kappa \sim 0.88 \pm 0.04$.

"not contaminated" substrate. For this scenario applied to the complete set of 19 minerals, we arrive at a picture (Fig. 6). The Cohen $\kappa \sim 0.4 \pm 0.03$ clearly indicates that we do not obtain a correct interpretation of the ingoing data set (Cohen, 1960). (An observation is assigned a value between 0 and 1, indicating the probability of correct identification.) The Cohen κ coefficient shows a far too low probability for a correct classification. When including the background reference measurements made separately, the results are much improved as can be seen in Fig. 7.

Now κ becomes 0.88 ± 0.03 and the multivariate approach is valid. Binary mixtures are a little easier to classify and only one or two spectra out of a hundred are assigned an incorrect group. The background effectively serves as an additional constraint, which directly improves the overall performance.

A very similar result, as far as peak lists are concerned, can be achieved at much lower computational effort by first computing a random projection of the spectrum (either the full spectrum with 10^5 bins or the peak lists with some 600 peaks) and then continue with PCA. Because the RP reduces the dimensionality greatly, the PCA becomes much faster. The speed depends primarily on the dimensionality. The reduction of it therefore speeds up the repetitive computations required for the cross validation. The process is difficult when

full spectra are considered because of the high dimensionality.

The selection of random number sequences used in the RP has a small and insignificant influence on the results, as long as the same projection matrix is used for the entire data set studied. The transformation matrix establishes an approximately orthonormal coordinate system in which the data is represented. The clustering can be directly computed from the RP projected data set but alternatively also be made by other methods, which we do not discuss here any further.

We finally estimate the number of valid signal carrying components using the method of Nadler (2008). For the described $4+1$ classification of 140 spectra we get ~ 60 PC components larger than the noise. This indicates that we need to use this number of degrees of freedom to catch all phenomena hidden in the data. Some information can already be extracted at lower dimensionalities. A dimensionality of ~ 10 is already sufficient to get the Cohen κ estimates to within one % of what the full data shows. This can easily be understood by realizing the speed of decline of the eigenvalues in the principal component estimate.

Because the RP is a linear method, it is natural that it is compatible with principal component analysis. It can be shown that the results of a PCA of a RP data set is approximately the same as the result computed directly (Seitola et al., 2014). The results are very similar as long as the dimensionality of the RP is sufficient to contain the information of the original data.

Classification

Establishing classification in a lower dimension is much less costly than directly doing it in a data volume of high dimensionality. Classification can in many cases be achieved using only a small number of PC components. In the general case we need to include a number large enough to represent all aspects of the data.

The RP technique makes it possible to compute a full cross validation of the classification of a large data set. We can easily in a few minutes establish the classification of the data set consisting of peak-list spectra from 19 minerals and 40 background spectra. We process a total of about 880 peak lists. The data matrix is 880×600 numbers. It is important to note that for this relatively large number of spectra and the likewise many parameters (peaks) the PCA in general is quite noisy. Therefore it is essential to ask how many components are actually meaningful and exceeding a noise background.

Applying Nadler (2008) to our data set again shows that we have about 90 meaningful components when computed from the peak-list data directly from the set of over 800 events. Therefore the maximum dimensionality needed for the RP would be of that same order of magnitude. The number of components required scales weakly with the number of spectra and parameters in the data. Analysing the previously discussed smaller data set of 140 spectra for four minerals and a background set, would require about 50 components and the same subset when projected to 100-dimension reduces the set to about 25. A small variation in the numbers is present and depends on details in the behaviour of the least significant components responsible for information at the 10^{-3} level of the total variance.

Using the RP for the purpose of performing the full cross validation of the data set classification is rewarding. The RP is able to pick out the discriminating features in a very beneficial manner. Using different projection matrices makes it possible to apply the methods multiple times and establish proper error estimates (sensitivity analysis).

5 Results

We have developed a technique of classifying measured TOF-SIMS mass spectra from the COSIMA instrument using multivariate statistical analysis of data projected by random projections into a lower dimensional space. A manual visual inspection and identification of spectra originating from mineral or background does not correctly identify the class memberships, while the task is performed well by the methods presented above. This fact demonstrates the usefulness of our approach.

The Cohen κ probability of correct classification varies from useless to reasonable (0.2–0.8) when using manual visual identification to derive class memberships as compared to almost complete success when using an additional reference background measurement. These results show the importance of measuring several reference spectra prior to exposing targets for dust. It is found that even peak lists are sufficient to determine class membership to a high degree of confidence.

We have found that using the random projection method of reducing the data dimensionality, classification results are obtained at a cross validation level very similar to what can be achieved using the full spectrum data set. The variation in scores is smaller when utilizing the random projection method, than when using full spectra.

The results of classification and cross validating four minerals and a background, is shown in Table 3. The classification is in this case computed from peak-list data, on which we optionally have applied random projections to validate the computational method. The variation in classification errors when changing the random projection dimensionality is of the order of 3 % as estimated by the Cohen κ. The variability in the κ estimate between individual cross validation runs when keeping the dimensionality fixed is of the order of 0.5 %.

The random projection method also makes it possible to compute the classification directly from full size spectra (see Figs. 8 and 9). From a spectrum truncated to 100 000 bins, we reduce the dimensionality for instance to 300 to catch even subtle features. This data set consists of 113 spectra. We

Table 3. Cross validation results for four minerals and background residing on the same substrate SUB-496. Classification accuracy shown at bottom computed from the average of several cross validations. Two albite spectra are incorrectly classified as fayalite and one orthoclase spectrum as albite, shown for one particular realization.

	Alb	Fay	Hyp	Ort	Bak
Albite	23	2	0	0	0
Fayalite	0	25	0	0	0
Hypersten	0	0	25	0	0
Orthoclase	1	0	0	24	0
Background	0	0	0	0	40
sACC	0.939	0.962	1.00	0.980	1.00

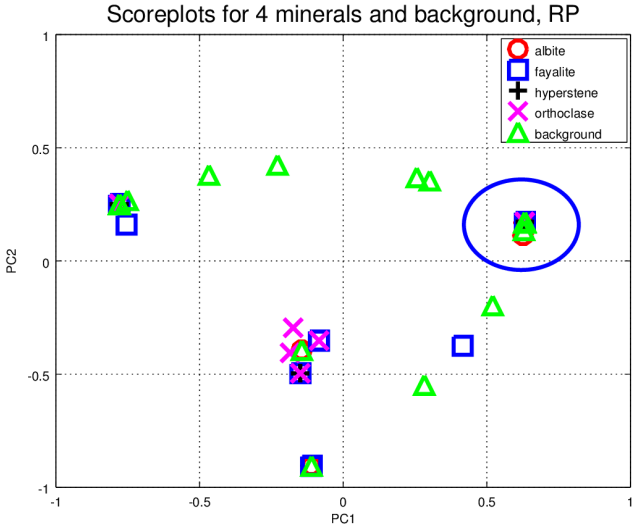

Figure 8. Score plot using the two first PCA components from RP projection of full spectra. The events represent the four minerals and the background for substrate 496. The circled region is magnified in Fig. 9.

compute the principal components from the 300-dimensional data and from the full spectra by limiting the number of components. As the PCA method is sensitive to outliers and large dimensions, the expected match between the two results, shows resemblance but not a complete match. The full spectra provide a much larger spread in scores than the RP result does. Also more details and repeated features are retained in the random projection results. An example is shown in Fig. 8 which should be compared to Figs. 4 and 5 above. The improved quality of the PCA from the projected full spectra is demonstrated by the enlarged rightmost cluster of points in Fig. 8 shown in Fig. 9.

A similar result can be obtained for the measurements treated as pairs of a binary mixture of mineral on background. The results for that case are very good with few errors in classification. A more demanding case is to classify all 19 minerals with the background. When applying the ran-

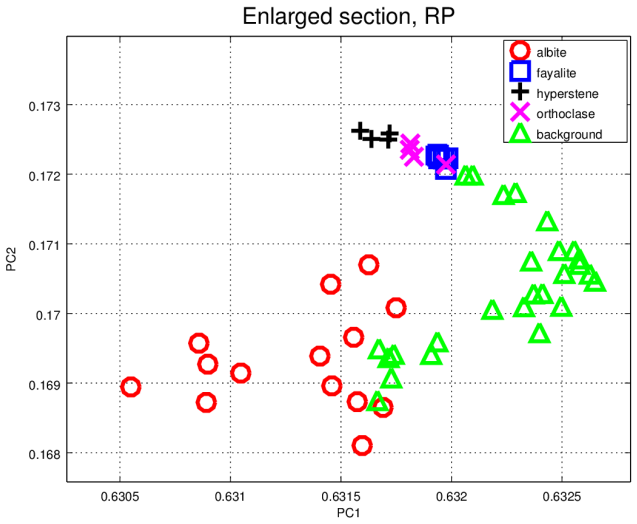

Figure 9. Score plot using the two first PCA components from RP projection of full spectra. The region shown is zooming into the cluster at far right in Fig. 8.

dom projection method to the data, the task is solved quickly and cleanly. Comparing the classification solution from the direct method operating on the full set of data and the case using the RP gives similar results. The Cohen κ is similar to within a few % and the classification results essentially identical. The compute time is one to several orders of magnitude lower with the RP providing results in only a few minutes. Furthermore the RP seems to yield a computationally more stable solution.

The classification and cross validation for the full data set using the RP method for computations is shown in Table 4. The number of PC values exceeding the noise level is about 90 for the peak-list data set. Decreasing the RP dimension to correspond to the data set in Table 4 only slightly changes the κ, decreasing it by less than 1 %. The corresponding score plot for the first two principal components is shown in Fig. 10. It represents the heart of the instrument performance. The shape is due to the fact that we view a point set (spectra) distributed on a unit sphere. We assume that we will be able to distinguish between the measured minerals quite well.

6 Conclusions

We have successfully applied the method of random projections to the classification of a set of TOF-SIMS spectra complete and alternatively represented by peak lists derived from full spectra. The reduced data set has been used directly or by first applying a random projection. For both cases we obtain very similar results. The computational cost and usefulness of the latter being superior.

Table 4. Cross validation results for all 19 minerals and background.

	alb	cal	cor	cpx	dol	fay	hyp	ilm	mag	nep	oli	opx	ort	pla	ric	sfo	sme	sph	sul	bak
alb	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
cal	0	22	0	0	1	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0
cor	2	0	12	0	0	0	0	10	1	0	0	0	0	0	0	0	0	0	0	0
cpx	0	0	0	18	0	0	0	0	1	0	2	13	0	0	1	0	0	1	2	0
dol	0	0	0	0	20	0	0	0	4	0	0	0	1	0	0	0	0	0	0	0
fay	0	0	0	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0
hyp	0	0	0	0	0	0	25	0	0	0	0	0	0	0	0	0	0	0	0	0
ilm	0	0	6	0	0	0	0	55	2	0	0	0	0	0	2	0	0	0	0	0
mag	0	0	6	0	0	0	0	0	19	0	0	0	0	0	0	0	0	0	0	0
nep	3	0	0	0	0	0	0	0	0	22	0	0	0	0	0	0	0	0	0	0
oli	0	0	0	4	0	0	0	0	0	0	60	1	0	0	0	0	0	0	0	0
opx	0	0	1	1	0	1	1	0	0	0	9	43	0	0	0	1	3	0	1	0
ort	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0	0	0	0	0	0
pla	0	0	0	0	0	0	0	0	0	0	0	0	2	26	0	0	0	0	0	0
ric	0	0	1	0	2	0	0	0	9	0	0	0	0	0	51	0	0	0	0	2
sfo	0	1	0	0	0	0	1	1	0	0	0	2	0	0	1	36	0	0	0	0
sme	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	40	0	0	0
spa	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	0	0
sul	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	0	47	0
bak	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25

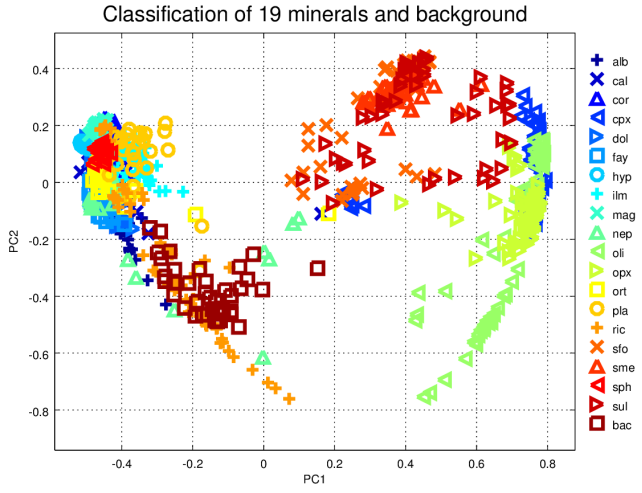

Figure 10. Classification of 19 minerals and background of a substrate. The PC components show only the most pronounced features. The classification matrix is shown in Table 4. Each sample is represented by a marker. The colours and markers represent individual classes. The elements are the same as in Table 4. More quantitative details are seen only when rotating a 3-D image of suitably chosen principal components.

The RP compressed data is an unbiased and a very robust tool to acquire information that is hidden in the data. This technique opens up a very powerful and general way of finding global taxonomy of observations.

The instrument provides us a peak list at very low telemetry cost. This data is shown to be sufficient for several tasks of discriminating between minerals. This makes the method well suited to make assessments on measurement strategies during the scientific phase of the mission.

Future work will include deriving peak lists using Bayesian methods to establish statistically strict quantitative information about spectra. This reduced set can then be analyzed by the method described here to achieve even better results.

Acknowledgements. COSIMA was built by a consortium led by the Max-Planck-Institut für Extraterrestrische Physik, Garching, Germany in collaboration with Laboratoire de Physique et Chimie de l'Environnement, Orléans, France, Institut d'Astrophysique Spatiale, CNRS/INSU and Université Paris Sud, Orsay, France, Finnish Meteorological Institute, Helsinki, Finland, Universität Wuppertal, Wuppertal, Germany, von Hoerner und Sulger GmbH, Schwetzingen, Germany, Universität der Bundeswehr, Neubiberg, Germany, Institut für Physik, Forschungszentrum Seibersdorf, Seibersdorf, Austria, Institut für Weltraumforschung, Österreichische Akademie der Wissenschaften, Graz, Austria and is lead by the Max-Planck-Institut für Sonnensystemforschung, Göttingen, Germany. The support of the national funding agencies of Germany (DLR), France (CNES), Austria and Finland and the ESA Technical Directorate is gratefully acknowledged. We thank the Rosetta Science Ground Segment at ESAC, the Rosetta Mission Operations Centre at ESOC and the Rosetta Project at ESTEC for their outstanding work enabling the science return of the Rosetta Mission. S. Siljeström acknowledges funding from the Swedish National Space Board (Contract 121/11).

Edited by: M. Díaz-Michelena

References

Bingham, E. and Mannila, H.: Random projection in dimensionality reduction: applications to image and text data, in: Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2001), New York, 245–250, 2001.

Broemeling, L. D.: Bayesian Methods for Measures of Agreement, Chapman & Hall/CRC Boca Raton, 2009.

Candes, E.: Compressive Sampling, Proceedings of the International Congress of Mathematicians, Madrid, Spain, 2006.

Cohen, J.: A coefficient of agreement for nominal scales, Educ. Psychol. Meas., 20, 37–46, 1960.

Dasu, T. and Johnson, T.: Exploratory Data Mining and Data Cleaning, Wiley, Hoboken, 2003.

Ding, C. and He, X.: K-means clustering via principal component analysis, in: Proceedings of the 21st International Conference on Machine Learning, Banff, Alberta, Canada, 2004, Proceeding ICML'04 Proceedings of the twenty-first international conference on Machine learning ACM New York, NY, 29–38, doi:10.1145/1015330.1015408, 2004.

Duda, R., Hart, P., and Stork, D.: Pattern Classification, 2nd Edn., John Wiley and Sons, 2012.

Eaton, J. W., Bateman, D., and Hauberg, S.: GNU Octave Version 3.0.1 Manual: a High-level Interactive Language for Numerical Computations, CreateSpace Independent Publishing Platform, Boston, 2009.

Filzmoser, P., Hron, K., and Reinmann, C.: Principal component analysis for compositional data with outliers, Environmetrics, 20, 621–632, 2009.

Giudici, P. and Figini, S.: Applied Data Mining for Business and Industry, John Wiley & Sons, Chichester, West Sussex, UK, 2009.

Gross, J. H.: Mass Spectrometry, Springer, Berlin, Heidelberg, 2004.

Hilchenbach, M. et al.: Contamination Impact on COSIMA Measurements, in preparation, 2014.

Hornung, K., Kissel, J., Fisher, H., Mellado, E. M., Kulikov, O., Hilchenbach, M., Krüeger, H., Engrand, C., Rossi, M., Langevin, Y., Krueger, F. R.: Collecting Cometary Dust Particles on Metal Blacks with The COSIMA Instrument Onboard ROSETTA, Planet. Space Sci., 103, 309–317, 2014.

Keogh, E., Lonardi, S., and Ratanamahatana, C. A.: Towards parameter-free data mining, in: KDD '04, 22–25 August 2004, Seattle, WA, USA, 2004.

Kissel, J., Altwegg, K., Clark, B. C., Colangeli, L., Cottin, H., Czempiel, S., Eibl, J., Engrand, C., Fehringer, H. M., Feuerbacher, B., Fomenkova, M., Glasmachers, A., Greenberg, J. M., Grun, E., Haerendel, G., Henkel, H., Hilchenbach, M., von Hoerner, H., Hofner, H., Hornung, K., Jessberger, E. K., Koch, A., Kruger, H., Langevin, Y., Parigger, P., Raulin, F., Rudenauer, F., Ryno, J., Schmid, E. R., Schulz, R., Silen, J., Steiger, W., Stephan, T., Thirkell, L., Thomas, R., Torkar, K., Utterback, N. G., Varmuza, K., Wanczek, K. P., Werther, W., and Zscheeg, H.: Cosima high resolution time-of-flight secondary ion mass spectrometer for the analysis of cometary dust particles onboard rosetta, Space Sci. Rev., 128, 823–867, 2007.

Kolokolova, L., Hanner, M., Levasseur-Regourd, A.-C., and Gustafson, B. Å. S.: Physical properties of cometary dust from light scattering and thermal emission, in: Comets II, edited by: Festou, M. C., Keller, H. U., and Weaver, H. A., University of Arizona Press, Tucson, 2004.

Krüger, H., Stephan, T., Engrand, C., Briois, C., Siljeström, S., Merouane, S., Baklouti, D., Fischer, H., Fray, N., Hornung, K., Lehto, H., Orthous-Daunay, F.-R., Rynö, J., Schulz, R., Silen, J., Thirkell, L., Trieloff, M., and Hilchenbach, M.: COSIMA-Rosetta calibration for in-situ characterization of 67P/Churyumov–Gerasimenko cometary inorganic compounds, Planet. Space Sci., submitted, 2014.

Lehto, H. J., Zaprudin, B., Lehto, K. M., Lönnberg, T., Silén, J., Rynö, J., Krüger, H., Hilchenbach, M., and Kissel, J.: Analysis of COSIMA spectra: Bayesian approach, Geosci. Instrum. Method. Data Syst. Discuss., 4, 563–588, doi:10.5194/gid-4-563-2014, 2014.

Le Roy, L., Bardyn, A., Briois, C., Cottin, H., Fray, N., Thirkell, L., and Hilchenbach, M.: COSIMA calibration for the detection and characterisation of the cometary solid organic matter, Planet. Space Sci., 105, 1–25, 2014.

McDonnell, J., Lamy, P., and Pankiewicz, G.: Physical Properties of Cometary Dust, Comets in the Post-Halley Era, Kluwer Academic Publishers, Printed in the Netherlands, 1991.

Monnig, N. D., Fomberg, B., and Meyer, F. G.: Inverting Nonlinear Dimensionality Reduction with Scale-Free Radial Basis Function Interpolation, arXiv:1305.0258v2, 1, 2013.

Mumma, M. and Charnley, S.: The chemical composition of comets – emerging taxonomies and natal heritage, Annu. Rev. Astron. Astr., 49, 471–524, 2011.

Nadler, B.: Finite sample approximation results for principal component analysis: a matrix perturbation approach, Ann. Stat., 36, 2791–2817, 2008.

Paatero, P. and Hopke, K.: Discaring or downweighting high-noise variables in factor analytic models, Anal. Chim. Acta, 490, 277–289, 2003.

Paatero, P. and Tapper, U.: Analysis of Different Modes of Factor Analysis as Least Squares Fit Problems, Chemometr. Intell. Lab. Syst., 18, 183–194, 1993.

Roweis, S. T. and Saul, K. L.: Nonlinear dimensionality reduction by locally linear embedding, Science, 290, 2323–2326, 2000.

Schlogl, A., Kronegg, J., Huggins, J. E., and Mason, S. G.: Evaluation criteria in BCI research, in: chapter 19 of: Toward Brain-Computer Interfacing, MIT Press, Cambridge, Mass., 327–342, 2007.

Seitola, T., Mikkola, V., Silen, J., and Jarvinen, H.: Random projections in reducing the dimensionality of climate simulation data, Tellus A, 66, 25274, doi:10.3402/tellusa.v66.25274, 2014.

Stephan, T.: Assessing the elemental composition of comet 81P/Wild 2 by analyzing dust collected by Stardust, Space Sci. Rev., 138, 247–258, 2008.

Tsou, P., Brownlee, D. E., Anderson, J. D., Bhaskaran, S., Cheuvront, A. R., Clark, B. C., Duxbury, T., Economou, T., Green, S. F., Hanner, M. S., Hörz, F., Kissel, J., McDonnell, J. A. M., Newburn, R. L., Ryan, R. E., Sandford, S. A., Sekanina, Z., Tuzzolino, A. J., Vellinga, J. M., and Zolensky, M.: Stardust encounters comet 81P//Wild 2, J. Geophys. Res., 109, 1–8, 2004.

Tuzzolino, A., Economou, T., Clark, B., Tsou, P., Brownlee, D., Green, S., McDonnell, J., McBride, N., and Colwell, M.: Dust measurements in the coma of comet 81P/Wild 2 by the dust flux monitor instrument, Science, 304, 1776–1780, 2004.

Varmuza, K. and Filzmoser, P.: Introduction to Multivariate Statistical Analysis in Chemometrics, CRC Press, Boca Raton, FL, 2009.

Varmuza, K., Filzmoser, P., and Liebmann, B.: Random projection experiments with chemometric data, J. Chemometr., 24, 209–217, 2010.

Varmuza, K., Engrand, C., Filzmoser, P., Hilchenbach, M., Kissel, J., Krueger, H., Silen, J., and Trieloff, M.: Random projection for dimensionality reduction – applied to time-of-flight secondary ion mass spectrometry data, Anal. Chim. Acta, 705, 48–55, 2011.