Data Analysis in Combinatorial Experiments: Applying Supervised Principal Component Technique to Investigate the Relationship Between ToF-SIMS Spectra and the Composition Distribution of Ternary Metallic Alloy Thin Films

Rossana Dell'Anna[a]*, Paolo Lazzeri[a], Roberto Canteri[a], Christian J. Long[b], Jason Hattrick-Simpers[b], Ichiro Takeuchi[b], Mariano Anderle[a]

[a] Fondazione Bruno Kessler – irst, Via Sommarive 18, 38050 Povo (Trento), Italy, E-mail: dellanna@itc.it
[b] Department of Materials Science and Engineering, University of Maryland, College Park, MD 20742, USA

Keywords: Composition spreads experiments, Feature selection, Regression models, Supervised principal components, ToF-SIMS spectra

Received: January 30, 2007; Accepted: March 30, 2007

DOI: 10.1002/qsar.200740008

Abstract
We apply a semi-supervised technique called Supervised Principal Component (SPC) to explore the relationship between the composition of a thin film combinatorial library and the peaks of Time-Of-Flight Secondary Ion Mass Spectrometry (ToF-SIMS) spectra acquired from the library. SPC is first used to select a subset of the available multivariate features (the peak intensities of the ToF-SIMS spectra) based on their association with the outcome variable (the elemental concentration of the thin film samples). Next, using only the selected features, SPC creates optimal linear models which map the ToF-SIMS data onto the composition data. The models for the first two of the considered elemental concentrations use only eight of the 55 available ToF-SIMS peaks, making interpretation of the model much simpler than for a model which uses all 55 available peaks. The percentage of explained variance (R^2) in concentration data is in both cases about 0.80. These results are obtained during the model validation phase, performed on test data, which are exclusively used for this purpose. The model for the third considered element did not produce significant results due to the poor variability of the dataset. This work illustrates for the first time that using a multivariate analysis technique, one can establish the relationship between ToF-SIMS measurements and stoichiometric data in a combinatorial experiment. More generally, the described feature selection approach provides an example of how combinatorial experiments can be useful for accelerating the understanding of the chemical–physical behaviors under investigation.

1 Introduction

Combinatorial experiments for the discovery and optimization of new materials [1] generate data that are usually multivariate, as arrays of variables (features) and multiple structural and/or functional outputs are typically associated with a library of compounds. A crucial aim of combinatorial experiments is to exploit these data for developing reliable predictive models which are capable of identifying materials possessing desirable physical properties. By recognizing new patterns in data, data mining strategies could provide new hypothesis on composition-structure–property relationships to be further validated by new experiments or theoretical explanations.

One area of critical importance in data mining is feature selection [2]. Feature selection is crucial in data mining because it helps to filter out both redundant and irrelevant information from a multivariate dataset. The decrease in dimensionality of a multivariate dataset obtained by feature selection not only reduces the computational expense of the algorithm, but can also reduce the risk of designing models which are over-fitted to data. By determining relevant modeling variables, a new insight into the mechanism governing the considered physical behavior is possible. Interpretability, scalability, and, possibly the accuracy of the

Abbreviations: PC, principal component; **PCR,** principal component regression; **SPC,** supervised principal component; **ToF-SIMS,** time-of-flight secondary ion mass spectrometry; **WDS,** wavelength dispersive spectroscopy resulting models are also improved. Therefore, feature selection crucially contributes to the knowledge discovery process.

In this work, we apply a technique called Supervised Principal Component (SPC) [3, 4] to a regression problem, namely investigating the relationship between the stoichiometry mapping from a Ni–Ti–Cu composition spread library, which was synthesized by using the combinatorial approach, and a set of peaks from Time-Of-Flight Secondary Ion Mass Spectrometry (ToF-SIMS) spectra. Generally speaking, the SPC method predicts a continuous response variable Y from a matrix $\mathbf{X}$ of r features measured on each of N samples. By a cross-validation procedure, SPC selects only those variables with the strongest computed correlation with Y and sorts them based on their influence on the model. For our case, the continuous response variable is the elemental concentration of a set of thin film samples and the features of the multivariate analysis are the peak intensities of ToF-SIMS spectra acquired from the same thin film samples.

ToF-SIMS provides information related to the chemical composition of the sample and, to some extent, its molecular structure [5]. The technique is characterized by a reasonably high throughput (tens of spectra *per* hour), good sensitivity, and provides intrinsically spatially resolved information. Analytical routines can be automated. ToF-SIMS can thus be used for the fast characterization of large combinatorial libraries. Although the ToF-SIMS technique is most commonly used for the analyses of molecular solids such as polymers, its utility is tested here for the characterization of ternary metallic alloy compounds.

The final goal of our investigation is to ascertain whether the ToF-SIMS data relate to any crystallographic materials phases in a metal alloy [6]. In fact, while it is known from the literature that ToF-SIMS can be used for speciation and phase identification of inorganic ionic compounds (typically catalysts [7]), a corresponding approach for metallic alloys has not been reported to the best of our knowledge. In this framework, we initially focus on the stoichiometric data.

2 Materials and Methods

2.1 Material Library Synthesis and Processing

Natural thin film composition spreads of the Ni–Ti–Cu system were deposited using an ultra high-vacuum three gun magnetron cosputtering system with a base pressure of 10^{-9} Torr (10^{-7} Pa) on 3-inch (76.2 mm) diameter (100) oriented Si wafers. The films were deposited at room temperature followed by an *in situ* annealing for 2 h in high vacuum at temperatures in the range of 853–970 K. A physical shadow mask was placed on the substrate during deposition in order to produce a grid of lines on the wafer for which there was no deposition. This grid divided the composition spread wafer into 535 individual 1.75 mm × 1.75 mm squares, with each square possessing a slightly different composition. Figure 1a shows a schematic of the synthesis of a ternary composition spread which covers the relevant part of the phase diagram. Further details of the synthesis procedure can be found in references [8, 9]. After the deposition, the composition of each square was immediately determined *via* Wavelength Dispersive Spectroscopy (WDS) in atomic percent. After the WDS measurement, ToF-SIMS measurements were made on 86 of the 535 samples. The composition distribution of samples for which both WDS and ToF-SIMS measurements were performed appears in Figure 1b. The results of the summary statistics separately calculated on the Ni, Ti, and Cu concentration values of Figure 1b are reported in Table 1, which clearly shows the poor variance of Cu concentration data with respect to the variances of Ni and Ti concentration distributions. The effects of these differences on the predictive power of the calculated SPC models are discussed in Section 4.

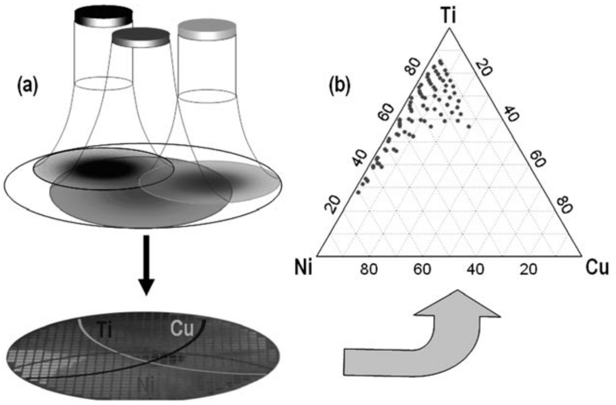

Figure 1. a) A schematic of the deposition of a thin film composition library made using a 3-gun cosputtering system. The three targets used in the deposition were Ni, Ti, and Cu. b) The region of the ternary diagram spanned by the 86 thin film samples of the considered dataset.

Table 1. Results of the summary statistics separately calculated on the Ni, Ti, and Cu concentration data (atomic percent) of Figure 1b.

	Ni	Ti	Cu
Minimum	10.9	27.8	1.1
First quartile	17.4	52.9	2.5
Mean	31.2	61.8	6.9
Third quartile	42.1	73.6	9.4
Maximum	70.4	85.2	28.8
SD	17.0	14.7	6.1

2.2 ToF-SIMS Spectra

ToF-SIMS analyses were performed in positive ions detection mode, using Ar^{+} (10 keV) as a primary beam. Prior

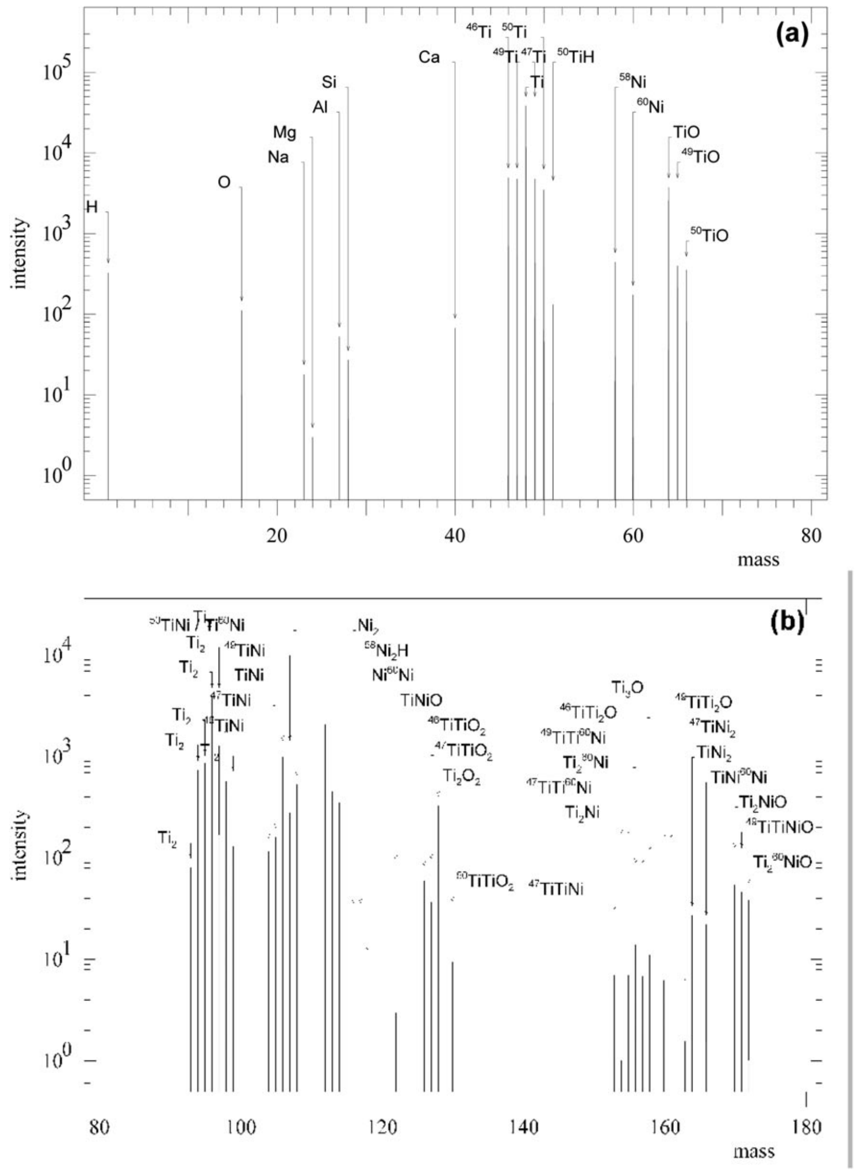

Figure 2. a) and (b): A typical TOF-SIMS spectrum from the Ni–Ti–Cu library. It is divided in two successive mass ranges for readability reasons.

to the analyses, the sample was pre-sputtered (using Ar^+) to remove any organic contaminants on the surface of the film. Care was taken in order to insure the minimization of any induced damage. The mass resolution was about 5000 at $m/z = {}^{48}Ti$. This allowed the identification of the largest part of the peaks. For each sample, a total of 55 peaks were present in the spectrum. The intensity of each peak was evaluated over an interval kept constant for all the spectra. A total of 86 spectra were recorded from the library, covering a wide range of elemental concentrations (see Table 1). Four samples were used as internal references to evaluate any drift in the ToF-SIMS spectra within the timeframe required to complete the analyses. Figures 2a and 2b show an example of a positive mass spectrum obtained from the Ni–Ti–Cu library. All of the spectra obtained are similar to that reported in Figures 2a and 2b with the exception that there is a variation in the relative peak intensity from sample to sample. Along with the pure elements related to the matrix species (Ti and Ni mainly), molecular species such as $TiNi_x(O_y)$ are present to a large extent. Residual inorganic contaminants (Na, Mg, Al, and Ca) can also be observed in the lower mass region. It must be noted that none of the spectra show any peaks related to Cu, despite the fact that according to the WDS measurements, the composition of the sampled region varies from 1 to 30%. We believe the absence of Cu peaks is related to the sample surface oxidation state, as the oxygen affinity is different for Cu, Ni, and Ti elements.

It has been observed that copper is present in the surface oxide of Cu-rich Ti–Cu and Ni–Cu alloys whereas in the surface oxide of Ti-rich or Ni-rich alloys (as in the present case) Cu is depleted [10]. As ToF-SIMS analyzes only the outermost (1–2) atomic layers of the sample, and since we only removed a surface layer a few nanometers thick during the pre-sputtering, it is reasonable to think that the sampled region was still low in Cu concentration. The presence of numerous oxide peaks is also consistent with the fact that only the top atomic layers are sampled.

2.3 Supervised Principal Components

SPC is a method for predicting a quantitative response variable Y from a set of features $X_1 \ldots X_r$ measured on N samples. Therefore, each feature X_j, $j=1,\ldots,r$, is a vector of N components. SPC is a generalization of Principal Component Regression (PCR). In a traditional PCR analysis, the first few Principal Components (PCs) are orthogonal linear combinations of the features, which account for the largest variances in the dataset. However, these new coordinates may not be highly correlated with the outcome variable of interest. SPC therefore computes the standardized regression coefficients that measure the correlation of each original feature to the response variable Y. It then selects only those features whose regression coefficient value is greater than a threshold value, which is determined by a cross-validation procedure. A principal component analysis is then carried out exploiting only the selected features and the first PCs (up to three) are used in a linear regression model to predict the response variable Y. Figure 3 shows the flow diagram for the SPC analysis. An exhaustive discussion of the method can be found in Ref. [3]. We used SPCs provided by the superpc package of R (version 2.3.0) [11, 12].

For the analysis reported in this paper, the considered features ($X_1 \ldots X_r$) are the 55 peaks of the ToF-SIMS spectra acquired from $N=86$ ternary alloys in the combinatorial library. Each spectrum was normalized to its total counts and, following the procedure outlined in Ref. [3], the features were centered to have a mean value of zero. Each compound is characterized by unique values of Ni, Ti, and Cu atomic concentrations. In our analysis, each elemental concentration represents a different outcome variable Y. In order to obtain an estimate of the generalization error of the model, the dataset was randomly partitioned into a training set of 43 alloys and a test set of the remaining 43 alloys. The SPC method uses cross-validation to estimate the best subset of features for the linear regression through estimation of the best threshold value. We used three-fold cross-validation. The training set is therefore divided into three mutually exclusive subsets of approximately equal size. Each model is trained on 2/3 of the training data, and then SPC computes the log-likelihood ratio test statistic on the remaining 1/3. The results are averaged over three runs. We repeated this process ten times choosing the best provided value. The independent test set is separately used for final validation, by considering both the log-likelihood ratio statistic and the p value provided by the significance test of the regression coefficient. When similar results were obtained for different threshold values, we always considered the threshold selecting the minimum set of variables. We considered only the first SPC, since all of the corresponding models were always highly significant.

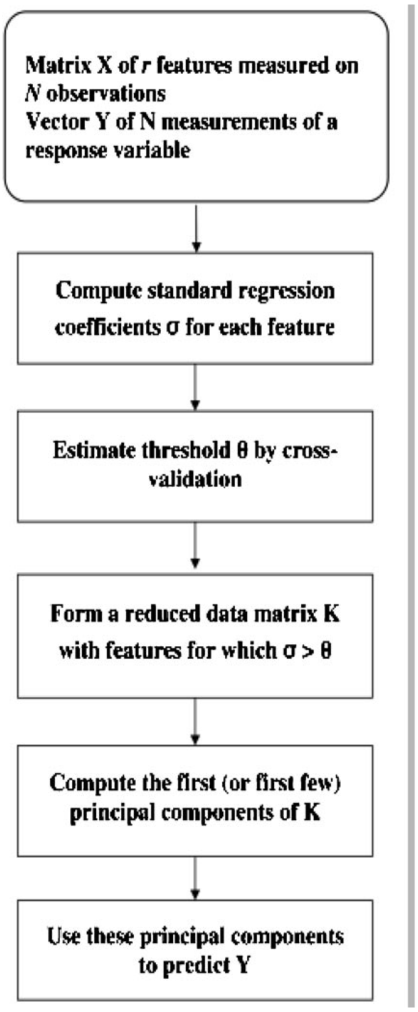
Matrix X of r features measured on N observations
Vector Y of N measurements of a response variable
Compute standard regression coefficients σ for each feature
Estimate threshold θ by cross-validation
Form a reduced data matrix K with features for which σ > θ
Compute the first (or first few) principal components of K
Use these principal components to predict Y

Figure 3. The flowchart for the SPC analysis.

3 Results

The procedure described in Section 2.3 was applied to build separate linear regression models matching the ToF-SIMS features to each of the Ni, Ti, and Cu concentrations. The results of the feature selection for Ni concentration are given in Table 2. The procedure selected eight of the 55 ToF-SIMS peaks, which are reported in order of decreasing importance score. The importance score is de-

Table 2. Results of SPC analysis for the modeling of Ni concentration. The linear regression statistics are the regression coefficient, *m*, with accompanying *p* value of the *t*-test for evaluating if *m* value is significantly different from zero. The *p* value represents the probability of finding a slope that is as large as or larger than the observed slope, under the null hypothesis that the true slope is 0. The adjusted R^2 value denotes the sample size dependent percentage of variation in the dependent variable (the Ni concentration) accounted for by the independent predictor variables (the selected ToF-SIMS peaks). The model is built using only eight of the available 55 ToF-SIMS peaks. They are here given sorted by decreasing absolute value of importance score.

Nickel concentration model

Linear Fit statistics	Selected features	Importance score
	$Ni^{60}Ni$	131750
	$TiNi_2$	130813
$m = 0.78$	$TiNi^{60}Ni$	119254
$p < 2 \times 10^{-16}$	$TiNiO$	114697
$R^2 = 0.81$	$Ti_2{}^{60}NiO$	95116
	$^{46}TiTiO_2$	83278
	$^{49}TiTiNiO$	53398
	$^{58}Ni_2H$	20160

fined in Ref. [3] as the inner product between each feature X_j and the first SPC u_1:

$$\text{imp}_j = \langle X_j, u_1 \rangle, j = 1, \ldots, r$$

The features with larger absolute values of importance contribute most to the model of *Y*. Table 2 also reports the results of the application of the linear regression model on the test data. In particular, it shows the regression coefficient, *m*, and the *p* value of the associated *t*-test for the null hypothesis of a zero slope. The value of the adjusted R^2, *i.e.*, the percentage of variance reduction, is also given as a measure of the goodness of fit.

Similarly, Table 3 displays the eight ToF-SIMS peaks selected by the SPC method for use in the model of Ti concentration. They are reported in order of decreasing importance score, together with the results of the performed linear regression on the test set. Figures 4 and 5 illustrate predicted Ni and Ti concentration values, calculated on the test set by the cross-validated regression models, *versus* correspondingly measured concentrations. The fitted lines are also reported.

For a further validation of these results, we checked the performance of the SPC procedure on the same dataset but with randomized assignments of the Ni and Ti concentration values to ToF-SIMS spectra. We also carried out another check by randomly assigning in each spectrum the measured peak intensities to the mass list. As expected, in all these cases no significant models were found, both for Ni concentrations and Ti concentrations prediction.

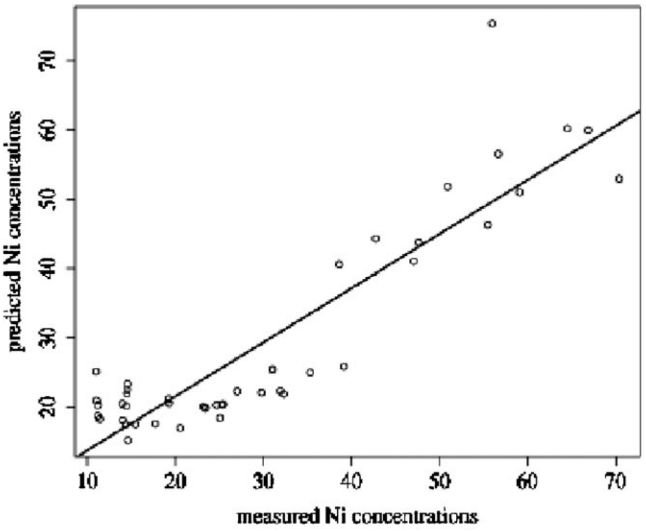

Figure 4. Scatter plot of predicted Ni concentration values *versus* measured concentrations with fitted line.

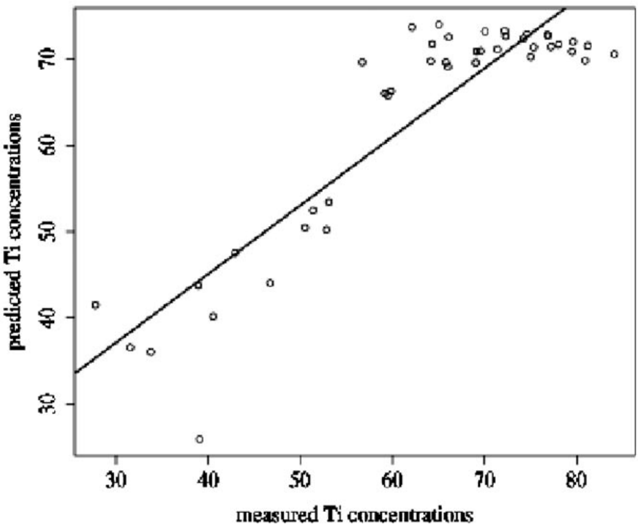

Figure 5. Scatter plot of predicted Ti concentration values *versus* measured concentrations with fitted line.

In contrast to what was obtained for Ni and Ti concentrations, the linear models found when considering Cu concentration as the outcome variable do not sufficiently explain the variance in Y data. As discussed in Section 2.2, no signals related to Cu are present in the TOF-SIMS spectra, even from the samples with the highest Cu content. However, the Cu concentration can be immediately calculated using only Ni and Ti values, as the sum of these three concentration values is always 100%. Therefore, since we separately obtained significant models for Ni and Ti concentrations, it is reasonable to look for linear regression models matching ToF-SIMS features to Cu concentrations, even though the ToF-SIMS spectra do not indicate

Table 3. Results of SPC analysis for the modeling of Ti concentration. See Table 2 for description.

Titanium concentration model		
Linear fit statistics	Selected features	Importance score
	$Ni^{60}Ni$	124342
	$TiNi_2$	122887
$m=0.79$	TiNiO	109865
$p=2.5\times10^{-16}$	$Ti_2{}^{60}NiO$	91007
$R^2=0.80$	$^{46}TiTiO_2$	80554
	O	75995
	$^{49}TiTiNiO$	50494
	$^{58}Ni_2H$	18633

that Cu is present. The negative result obtained is discussed in the next Section.

4 Discussion

We applied the SPC method to find the optimal reduced sets of ToF-SIMS spectra peaks for building linear regression models able to fit the composition information of a Ni–Ti–Cu combinatorial library. The fit results for Ni and Ti concentrations are separately reported in Tables 2 and 3. In both cases, the regression coefficient m is significantly different from zero and the percentage of variance in the concentration explained by the model is definitely high. SPC builds the models by selecting in both cases eight of the 55 ToF-SIMS peaks, which are reported in the same tables. The significant decrease in dimensionality of the multivariate dataset we obtained by feature selection reduces the computational expense of the algorithm. Therefore, the model scalability is improved. It also reduces the risk of designing models which are over-fitted to data. Most of all, by determining relevant modeling variables, the interpretability of these resulting models is facilitated, as we will discuss hereinafter. Therefore, the feature selection approach we applied surely helps to follow a more effective knowledge discovery process.

A significant result of our analysis is that for both Ni and Ti concentration cases the optimal predictive model is obtained by only considering the heaviest molecular species (see Figures 2a and 2b), except for the presence of the oxygen in Table 3. Noting that ToF-SIMS measurements provide a fragmentation pattern consisting of both elemental and molecular species, we can conclude that the model built with the selected features indicates a reduced importance of the elemental species for the modeling of concentration. This is interesting because these signals are those usually singularly considered in univariate analysis performed for the same purpose. In addition, the two models are consistent, as the selected features for modeling the Ni concentration and those for Ti concentration are nearly identical, with the only exception being that the $TiNi^{60}Ni$ peak in the Ni concentration model is replaced by the O peak in the Ti concentration model. It is interesting to note that both lists contain molecular fragments which do not contain the element whose concentration is being modeled. For example, the list of selected features for Ni contains $^{46}TiTiO_2$ and the list for Ti contains $Ni^{60}Ni$ and $^{58}Ni_2H$. This is not surprising. The high value of importance of these features is a result of the absence of a Cu signal in the ToF-SIMS spectra and the fact that the sum of the concentrations of Ni, Ti, and Cu must add up to 100%. These constraints imply that if a feature containing Ni is strongly correlated to Ni concentration, it will also be correlated to Ti concentration, and *vice versa*. Since the SPC method uncovers groups of peaks which coexpress the elemental concentration considered [3], it includes both features which are strongly correlated and features which are indirectly strongly correlated to the concentration considered.

The feature selection algorithm of the SPC method correctly discards Al, Ca, and alkali peaks, the occurrence of which is related to contaminants. A consistent part of selected features is represented by oxides, which of course reflect the surface oxidation state. In this perspective, the O peak of Table 3, selected for prediction of Ti concentration values, tells us that it comes essentially from the samples and not from impurities. In addition, it should be mainly related to Ti oxidation. This is a simple example of how feature selection in combinatorial data analysis can contribute to the knowledge discovery process.

As written in Section 3, the linear models provided by SPC do not satisfyingly explain the variance in Cu concentration data (best adjusted $R^2=0.42$), even though the obtained slope coefficients are significantly different from zero ($m=0.43$, $p=1.5\times10^{-6}$). Therefore, in order to test if another method was more effective than SPC in finding correlations between the Cu concentrations and the ToF-SIMS data, we also decided to apply the Partial Least Squares (PLS) technique to the same problem. We carried out the PLS calculation on the same training and test sets used for SPC analysis. PLS is a supervised technique that builds predictive models by constructing weighted linear combinations of the input variables that have the maximal covariance with the output (see for instance Ref. [13]). One important difference between PLS and SPC is that PLS uses all of the dataset features while SPC uses only a subset of the features. We found that even the PLS model is definitely not able to account for the variance in Cu concentration in terms of all ToF-SIMS peak intensities (best adjusted $R^2=0.51$, $m=0.53$, $p=5.5\times10^{-8}$). In contrast, PLS models mapping Tof-SIMS spectral data onto Ni and Ti concentration are quite significant. Thus, the PLS results confirm that the unsatisfactory results in modeling the Cu concentration using the SPC method are mainly due to the poor variability of the training set. As illustrated in Table 1, the range of concentration values for Cu is quite narrow compared to the range of concentration values for Ni and Ti. This leads us to the conclusion that there is not enough diversity in the examples provided during the learning phase to attain an accurate model for Cu concentration. Therefore, the poor content of Cu (up to 30%) causes the absence of Cu signal in ToF-SIMS spectra, while the poor variability of Cu content (variability of Cu values in the 1–30% range) causes the Cu concentration model to be not significant. The relationship between the Cu concentration and ToF-SIMS data should therefore be explored on a different dataset. In contrast to the model for Cu concentration, the SPC model accuracy (evaluated by three-fold cross-validation on the training set and further assessed on the test set) for Ni and Ti concentration regression was very good. In any case, a further assessment on a larger, independent dataset should be carried out to confirm the reliability of the model.

To better depict SPC performances and merit, a more exhaustive comparison with other regression approaches (including PLS) should be accomplished. This investigation is beyond the scope of this paper and was already carried out by the developers of SPC [3]. Here we only mention that we compared SPC with both PLS and PCR technique.

In our comparison of SPC and PLS results, the best models provided by PLS to map ToF-SIMS spectral data onto Ni and Ti concentration data were built from more than three latent variables and performed better than SPC models (six latent variables for Ni concentration model, $R^2=0.91$; four latent variables for Ti-concentration model, $R^2=0.88$). However, the goal of this article is not to promote the SPC method as an alternative to PLS approach, but rather to stress the possibility of building a predictive model which exploits a minimum set of features. In this sense, SPC surely helps to follow a more effective knowledge discovery process with respect to PLS technique. We already described the effects of the SPC feature selection approach, here we want to stress that the good SPC performance is guaranteed in our cases by using models built on only eight instead of 55 features. SPC also provides the importance score of these predictor variables. On the contrary, PLS does not perform an initial thresholding of features. PLS is a good regression method, but its drawback is that a PLS model could be difficult to interpret, as each latent variable involves a linear combination of all features. Figure 6 is illustrative of this concept. As an example, we consider the best cross-validated predictive model ($R^2=0.88$) for Ti concentration values that we obtained by PLS.

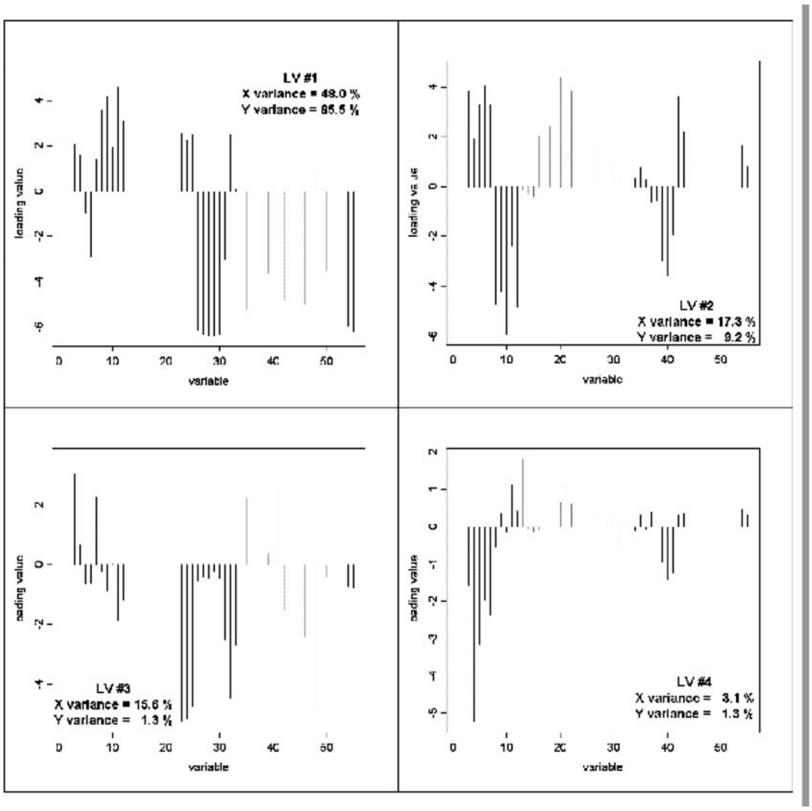

Figure 6. The four loading plots of the most accurate PLS model for Ti concentration prediction. Each plot is associated to a different latent variable (LV) and shows the relationships among the predictor variables producing that new coordinate. The X-axis reports the 55 considered features. For readability reasons they are numbered following their position in the ToF-SIMS spectrum (see Figures 2a and 2b). Each plot reports the corresponding LV and the percentages of the simultaneously explained variances of X and Y.

It exploits four latent variables. Therefore, the predictor variables interpretation proceeds through the parallel analysis of the four loading plots reported in Figure 6, with particular attention for the first plots, which most contribute to the overall percentage of captured covariance. This is surely a difficult task that probably can not be completely accomplished. Therefore, when there is the need to reduce the number of features in order to gain insight into the mechanism governing the behavior of some physical property, the usefulness of PLS is limited and SPC can instead provide interesting results.

In PCR technique [14], instead of regressing the system property (*e.g.* concentrations in our case) on the original measured variables (ToF-SIMS peak intensities in our case), the property is regressed on the PC scores of the measured variables, which are orthogonal and, therefore, inherently uncorrelated. We decided to build models based on the first PC to directly compare them with SPC results and therefore analyze the strength of our feature selection approach. We obtained for Ni concentration data $R^2=0.81$ and for Ti-concentration data $R^2=0.76$. Even in this case the Cu concentration model was not predictive enough ($R^2=0.21$). The accuracy of SPC model for Ni data was similar and that for Ti data was slightly better. Therefore, this comparison confirms that the feature selection approach performed by SPC analysis was effective in filtering out both redundant and irrelevant information, as SPC models involve eight instead of the 55 features exploited in PCR models. In addition, the higher number of features of PCR models causes even in this case their difficult interpretation.

5 Conclusions

In this work, we applied the SPC method in order to map the relationship between the composition information of a Ni–Ti–Cu thin film library and the ToF-SIMS spectra acquired from the library. The possibility of performing univariate analysis relating the intensity of a single peak of ToF-SIMS spectrum to stoichiometric data is well understood. However, to the best of our knowledge this is the first study in which the correlation is assessed by a multivariate approach analyzing ToF-SIMS spectra taken from a combinatorial library. In the context of combinatorial data analysis, the selection of a reduced number of predictor variables allows us to build more understandable models, thereby accelerating the knowledge discovery process. In the future, we plan to apply the same SPC approach in order to predict other physical properties.

Acknowledgements

This study has been granted by the Autonomous Province of Trento, through the FP-PAT Micro-Combi triennial (2004–2007) Project. This work was supported in part by the IMI Program of the National Science Foundation under Award Number DMR-0603644 and NIH EB005997.

References

[1] I. Takeuchi, X. D. Xiang (Eds.), *Combinatorial Materials Synthesis*, Dekker, New York **2003** (ISBN: 0-8247-4119-6).

[2] S. Weiss, C. Kulikowski, *Computer Systems That Learn*, Morgan Kaufmann Publishers, San Hatteo (CA) USA **1991**, pp. 72–76 (ISBN: 1558600655).

[3] E. Bair, T. Hastie, D. Paul, R. Tibshirani, *J. Am. Stat. Assoc.* **2006**, *101*, 119–137.

[4] E. Bair, R. Tibshirani, *PLoS Biol.* **2004**, *2*, 511–522.

[5] J. C. Vickerman, D. Briggs (Eds.), *ToF-SIMS Surface Analysis by Mass Spectrometry*, IM (Chichester) and Surface Spectra (Manchester) Publications, UK **2001** (ISBN 1901019 039).

[6] I. Takeuchi, C. J. Long, O. O. Famodu, M. Murakami, J. Hattrick-Simpers, G. W. Rubloff, M. Stukowski, K. Rajan, *Rev. Sci. Instrum.* **2005**, *76*, 062223.

[7] F. De Smet, M. Devillers, C. Poleunis, P. Bertrand, *J. Chem. Soc., Faraday Trans.* **1998**, *94*, 941–947.

[8] I. Takeuchi, O. O. Famodu, J. C. Read, M. Aronova, K. S. Chang, C. Craciunescu, S. E. Lofland, M. Wuttig, F. C. Wellstood, L. Knauss, A. Orozoco, *Nat. Mater.* **2003**, *2*, 180–184.

[9] O. O. Famodu, J. Hattrick-Simpers, M. Aronova, K. S. Chang, M. Murakami, M. Wuttig, T. Okazaki, Y. Furuya, I. Takeuchi, *Mater. Trans., JIM* **2004**, *45*, 173.

[10] A. Dhawan, S. Roychowdhury, P. K. De, S. K. Sharma, *Bull. Mater. Sci.* **2003**, *26*, 609–612.

[11] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria **2006** (ISBN: 3-900051-07-0), http://www.R-project.org.

[12] E. Bair, R. Tibshirani, superpc: Supervised principal components, R package version 1.03, **2005**, http://www-stat.stanford.edu/~tibs/superpc.

[13] P. Geladi, B. R. Kowalski, *Anal. Chim. Acta* **1986**, *185*, 1–17.

[14] H. Martens, T. Næs, *Multivariate Calibration*, John Wiley & Sons, Chichester, UK **1989** (ISBN:0471930474).