Alma Mater Studiorum – Università di Bologna

DOTTORATO DI RICERCA IN SCIENZE CHIMICHE

Ciclo XXVIII°

Settore Concorsuale di afferenza: 03/A1 - Chimica Analitica

Settore Scientifico disciplinare: CHIM/12 - Chimica dell'Ambiente e dei Beni Culturali

DIRECT QUANTITATIVE ANALYSIS OF SOLID SAMPLES: CHEMOMETRICS AND SHRINKAGE METHODS APPLIED TO SPECTROSCOPIC DATA

Presentata da:

Dott. Francesco de Laurentiis

Coordinatore Dottorato

Prof. Aldo Roda

Relatore

Prof.ssa Dora Melucci

Co-Relatore

Prof. Clinio Locatelli

Esame finale anno 2017

Index

Abstract

Multivariate analysis has rapidly developed in the past few years. This rise is due to advances in intelligent instruments and laboratory automation as well the possibility of using powerful computers and user-friendly software. In the field of analytical chemistry, the capability of newer, mostly multicomponent or multielement analytical methods produces so many data, that only the use of mathematical and statistical techniques can provide a suitable interpretation. The term Chemometrics, introduced for the first time by the Swedish scientist Swante Wold in the early 70's, concerns the implementation of multivariate principles to chemical data with the goal both to describe observed phenomena (and relationships among involved variables) and to create useful models for prediction purposes.

The aim of the present work is to develop multivariate methods for processing experimental data obtained through non-destructive techniques, in which it is possible to investigate samples without altering them, and keeping the sample available for further analysis. For qualitative investigation, such "direct" analytical procedures like infrared spectroscopy are available; however, the univariate approach is not exaustive in case of very complex matrices. The quantitative approach is still an open issue, due to the strong matrix effect hindering the creation of univariate calibration methods in interpolation mode. Multivariate analysis may be the solution.
This thesis is organized as follows:

In Section 1 the general problem of high-dimensional data is introduced, reviewing the basic principles of Principal Components and their implementation for descriptive and predictive purposes. In the last part of this section the core of the present work is discussed: advanced algorithms aimed to perform standard addition method in multivariate analysis.

Section 2 is dedicated to theory of the employed analytical technique: the basis of infrared spectroscopy, focusing particular attention to reflectance technique and issues related to data handling.

Section 3 describes the typologies of the analysed samples (marine sediments), the reason of interest of one of their specific components (biogenic silica), and shortcomings related to traditional methods of analysis.

In Section 4 experimental data, their computational treatment and a final discussion of results compared with other reference methods are presented.

Section 1

1.1 A current topic: facing with big dataset

Advances in data collection and storage capabilities during the past decades have led to an information overload in most sciences. Researchers working in domains as diverse as engineering, astronomy, biology, remote sensing, economics, and consumer transactions, face larger and larger observations and simulations on a daily basis. In many cases such observations are related with data where many variables are involved. *Multivariate statistics* include all statistical techniques for analyzing two or more variables of interest: it essentially concerns the statistical process of simultaneously analyzing multiple independent variables (predictors) with multiple dependent variables (outcome or responses) using matrix algebra. While these analyses have been a part of statistics since the early 1900's, the development of mainframe and microcomputers and subsequent analytical software has made the once tedious calculations fairly simple and very fast. In the field of chemistry the processes requiring investigation have become increasingly complex. Consequently, the relevance of analytical chemistry has rapidly increased. In the past the acquisition of data was a limiting step in the analytical process, but the introduction of new instrumental methods has improved the production of analytical information. Such abundance of data provided the possibility of more detailed and quantitative description of the observed phenomena. With the development of computer science and technology it became easier for analytical chemists to apply computational and advanced statistical methods in their work. This link between statistic and chemical knowledge led to the birth of a new discipline called Chemometrics.

In analytical chemistry, a typical case of application of multivariate techniques are spectroscopic and chromatographic measurements. As a matter of fact, such methods provide analytical data on many components of a single sample. For example, when we record a spectrum, each wavelength can be considered a single variable and the measured sample is described by the absorbances of the whole examined wavelength range. In other words, we can say that most instrumental measurements are inherently multivariate, since many variables can be related to a single sample. In the recent years, a number of methods have been developed for big-data processing, with the aim either of exploring relationships and structures or of confirming hypotheses. The methods rely

strongly on graphical presentation of the results. Among the most well known and analysed methods in this tradition, there are principal component analysis (PCA) for interpreting large data matrices, partial least squares (PLS) regression and principal component regression (PCR) for relating different data sets to each other, and shrinkage methods (e.g. LASSO, Ridge Regression) aimed to select significant variables.

1.2 Data presentation: dataset and graphs

Datasets The first step in facing a *huge* amount of data is to organize them in structures that can be subjected to mathematical elaborations. The key concept underlying the classification of multivariate methods is the data matrix. A conceptual illustration is shown in Table 1.1. It can be noticed that the table (also called dataset) consists of a set of objects (the *m* rows) and a set of measurements on those objects (the *n* columns). Cell entries represent the value x_{ij} of i-th object on j-th variable. The objects are any kind of entity measurable characteristics (numeric or not). The variables are characteristics of the objects and serve to define the objects in any specific study.

	Variables				
Objects	*1*	*2*	...	...	*n*
1	x_{11}	x_{12}	...		x_{1n}
2	x_{21}	x_{22}	...	...	x_{2n}
...					
...					
m	x_{m1}	*xm2*			x_{mn}

Table 1.1

Graphs When only two variables are measured, the information they bring can be graphically represented: a plot can be created where the coordinates of the points are the values measured for the two variables. The point can also be defined by a vector, called a data vector, drawn to it from the origin: in this case we have a two-dimensional vector. Objects which have similar properties will have similar data vectors, that is they

will lie close to each other in the space defined by the variables. Such a group is called a cluster. This way of presentation is very valuable in practice, since the human eye is superior to anything else in detecting structures and relationships. A graphical representation is less easy for three variables and no longer possible for four or more: it is here that computer analysis is particularly valuable in finding patterns and relationships. Matrix algebra is needed in order to fully describe the methods of multivariate analysis.

1.3 Dimension reduction: introduction to PCA

Modern datasets, in contrast with smaller more traditional datasets that have been widely studied in the past, present new challenges in data analysis. Traditional statistical methods break down, partly because of the increase in the number of observations, and also because of the increase in the number of variables associated with each observation. High-dimensional datasets present many mathematical issues as well as some opportunities, which give rise to new theoretical developments. One of the problems with high-dimensional datasets is that, in many cases, not all the measured variables are "important" for understanding the underlying phenomena of interest. In fact, quite frequently there is some correlation between the variables, and so some of the information is redundant. The use of redundant, irrelevant, and noisy variables tends to compromise the performance of many statistical tools, leading to unreliable inferences and costly data collection. This is particularly true when calibration methods are applied. Moreover, a huge volume of data may make it difficult to see patterns and relationships. For example, a spectrum would normally be characterized by several hundred intensity measurements and in this case it is impossible to represent graphically the samples in order to highlight similarities and differences.

Dimension Reduction refers to the process of converting a set of data having vast dimensions into data with lesser dimensions, ensuring that it concisely conveys similar information. There are many reasons to reduce the dimension; some of them are listed below.

- It helps in data compressing and reducing the required storage space.
- It fastens the time required for performing same computations. Less dimensions leads to less computing, also less dimensions can allow usage of algorithms unfit for a large number of dimensions.

- In some cases, for example in performing regression with PLS method, collinearity between variables can affect model performance. By removing redundant features the problem can be overcome.
- Reducing the dimensions of data to 2D or 3D may allow to plot and visualize it. Patterns can be more clearly observed.
- It is also helpful in noise removal, and as result we can improve the performance of models.

From a mathematical point of view, the problem of dimension reduction can be stated in the following terms: given the p-dimensional random variable $x = (x_1; \ldots ; x_p)$, find a lowest dimensional representation of it, $s = (s_1; \ldots ; s_k)$ with $k<p$, that captures the content in the original data, according to some criterion (for example imposing orthogonality). The components of s are sometimes called the latent components. Different fields use different names for the p multivariate vectors: the term “variable" is mostly used in statistics, while “feature" and “attribute" are alternatives commonly used in the computer science and machine learning literature.

There are many methods to perform dimension reduction. For example, in the method called Factor Analysis if some variables are highly correlated, these variables can be grouped by their correlations, *i.e.* all variables in a particular group can be highly correlated among themselves but have low correlation with variables of other group(s). Here each group represents a single underlying construct or factor. These factors will supposedly be small in number as compared to large number of original variables. In summary, the purpose of factor analysis is to discover simple patterns in the pattern of relationships among the variables. In particular, it seeks to investigate whether the observed variables can be explained largely or entirely in terms of a much smaller number of variables called factors. Other techniques (Backward Feature Elimination, Decision Trees, Low Variance) can be used in order to lower the dimension; however, Principal Component Analysis is doubtless the oldest and best known of the techniques of multivariate analysis.

1.4 Principal Component Analysis

Principal Component Analysis (PCA) is probably the most popular multivariate statistical technique and it is used by almost all scientific disciplines. It is also likely to be the oldest multivariate technique. In fact, its origin can be traced back to

Pearson(1901) but its modern instantiation was formalized by Hotelling (1933) who also introduced the term *principal component*. PCA analyses a data table where observations are described by several dependent variables, which are, in general, inter-correlated. The method is aimed to extract the relevant information from the data table and to express this information as a set of new orthogonal variables called principal components (PCs). The goals of PCA can be summarized in the following points:

- extract the most important information from the data table
- compress the size of the dataset by keeping only this significant variables
- simplify the description of the data set
- analyse the structure of the observations and the variables.

In order to achieve these goals, PCA algorithm performs a rotation of original axes and transforms original variables a into a set of new, latent variables, which are linear combination of the original ones. PCA also provides graphical representation showing the pattern of similarity of the observations and the variables by displaying them as points in graphs. Mathematically, PCA can be obtained by the eigen-decomposition of covariance or correlation (positive semidefinite) matrices and/or by the singular value decomposition (SVD) of rectangular matrices. The method is used in numerous areas of application and has been the source of inspiration for much of the development which has taken place in multivariate analysis. In fact, even though PCA is essentially a descriptive technique, some regression methods like Principal Component Regression (PCR) and Partial Least Square (PLS) rely on the same concepts. PCA is based on identifying the most important directions of variability in a multivariate data space (data matrix) and present to the user the results in graphical plots on the computer screen.

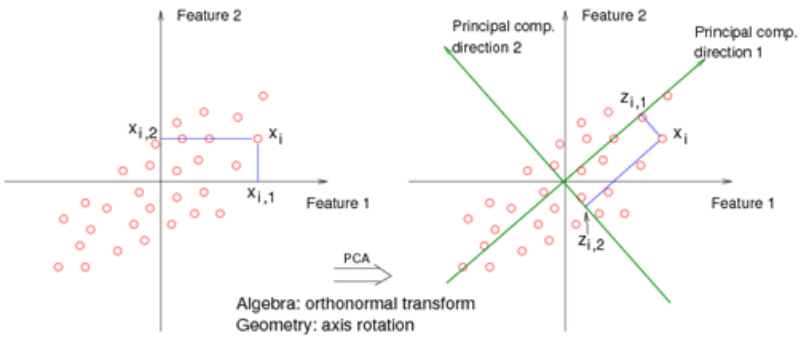


Fig.1.1

This way of presentation is very valuable in practice, since thanks to graphical representation it is possible to assemble objects (*i.e.* samples) in clusters, detect correlations among original variables and, finally, relationship between samples and original variables. However, it should be stressed that since only the directions of main variability in the data matrix are given attention in the PCA, the more subtle sources of variability may pass unnoticed. Therefore, multivariate treatments of this kind of data should usually be accompanied by more detailed studies, by for instance ANOVA methods.

1.5 PCA mathematical feature

Before implementing the PCA algorithm, in some cases it can be necessary to perform a mathematical pre-treatment on variables. The pre-processing part can make the difference between a useful model and no model at all, and it is advisable when the variables are measured with different units.

One term for pre-processing is called autoscaling, which is the combination of mean centering and standardization. Standardization is one type of scaling where each value is scaled by 1/STD

$$a_{ij}^{\dagger} = \frac{a_{ij} - \bar{a}_j}{s_j}$$

Mathematically, PCA is defined as a orthogonal linear transformation and assumes all basis vectors are an orthonomal matrix. PCA is concerned with finding the variances and coefficients of a dataset by finding the eigenvalues and eigenvectors. The PCA is computed by determining the eigenvectors and eigenvalues of the covariance matrix. The covariance matrix is used to measure how much the variables vary from the mean with respect to each other. The covariance of two random variables is their tendency to vary together. For two variables X and Y, this can be explicitly written out as:

$$\mathrm{cov}(x, y) = \sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{N}$$

with $\bar{x}$ = mean(X) and $\bar{y}$ = mean(Y), where N is the dimension of the dataset. The covariance matrix is a matrix A with elements $A_{i,j}$ = cov(i,j).

In the covariance matrix, the exact value is not as important as its sign (*i.e.* positive or negative). If the value is positive, it indicates that both dimensions increase, meaning that as the value of dimension X increased, so did the dimension Y. If the value is negative, then as one dimension increases, the other de-creases. In this case, the dimensions end up with opposite values. In the final case, where the covariance is zero, the two dimensions are independent of each other.

If we consider a dataset of N objects and K variables

$$X = \begin{pmatrix} x_{11} & x_{12} & . & . & x_{1K} \\ x_{21} & . & & & \\ & & . & & \\ & & . & & \\ & & . & & \\ & & . & & \\ & & . & & \\ x_{N1} & & . & . & x_{NK} \end{pmatrix}$$

the covariance matrix can be expressed by the product

$$\mathrm{Cov}(X)=X^{T}X$$

where X^T is the transposed matrix.

Due of the commutative attribute, the covariance between x_{ij} and x_{ji} is equal to the covariance between x_{ji} and x_{ij}. So the covariance matrix is symmetric.

From linear algebra we know that a symmetric matrix can be decomposed: the spectral decomposition (or Jordan decomposition) links the structure of a matrix to the eigenvalues and the eigenvectors.

$M = SJS^{-1}$ where M and J are called similar matrices

Therefore the eigenvectors and eigenvalues of the covariance matrix can be calculated

The computation of eigenvalues and eigenvectors is an important issue in the analysis of matrices. For a given Σ (covariance matrix)

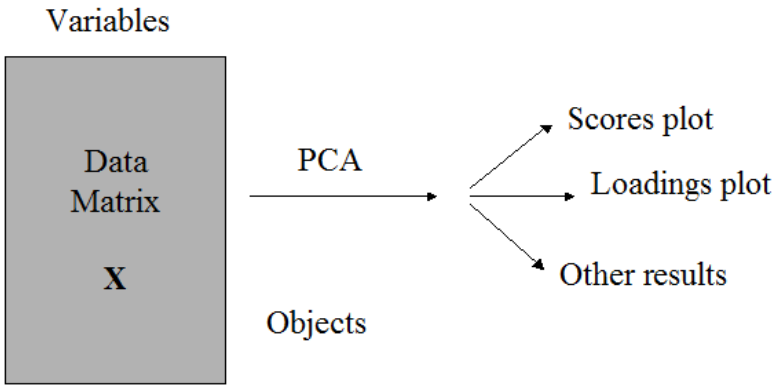
$\Sigma=U\Lambda U^{T}$

where Λ = diag (1; … ; *p*) is the diagonal matrix of the ordered eigenvalues 1,…,*p*, and U is a *pxp* orthogonal matrix containing the eigenvectors.

Once the eigenvectors and the eigenvalues are calculated, the eigenvalues are sorted in descending order. This gives us the components in order of explained variance. The eigenvector with the highest eigenvalue expresses the direction with the highest variance. The second component is computed under the constraint of being orthogonal to the first component and to have the largest possible variability. The other components are computed likewise. Finally, in order to get the coordinates in the new space (PCA space) the original dataset is projected in the eigenvectors directions. Therefore, principal components are calculated by multiplying each row of the eigenvectors with the sorted eigenvalues.

$$T = XU$$

where X is the original dataset



The values of the new variables for the observations are called factor *scores*, and these factors scores can be interpreted geometrically as the projections of the observations onto the principal-components space. PCA provides several results and numerical outcomes, but probably the most useful are the graphs of *scores* and *loadings*. The scores are the coordinates with respect to PCs, while loadings are normalized coefficient accounting for the relevance of original variables with respect to PCs. In such plots, samples and original variables are depicted in the new "latent" space where is possible detect similarities and correlations.

1.6 Regression

Calibration is fundamental to achieve consistency of measurement. Often calibration involves establishing the relationship between an instrument response and one or more reference variables. In statistics, the term regression is used to describe a group of methods that summarize the degree of association between one variable (or set of variables) and another variable (or set of variables). The simplest and most common type of association between two variables is the linear relationship $y=a+b\cdot x$. However, not all physical or chemical relationships can be adequately described using the simple linear model and more complex functions, such as quadratic and higher order polynomial equations, may be required to fit the experimental data. For this reason there are a number of regression models such as logarithmic, exponential and power. The most common statistical method used to perform a regression is least-squares regression, which works by finding the "best curve" through the data that minimizes the sums of squares of the residuals. The important term here is the "best curve", not the method by which this is achieved. Even though there are a number of least-squares regression models, linear regression is anyway one of the most frequently used statistical methods in calibration. Once the relationship between the input value and the response value (assumed to be represented by a straight line) is established, the calibration model can be used in reverse, that is to predict a value from an instrument response. The principal aim in undertaking regression analysis is to develop a suitable mathematical model for descriptive or predictive purposes.

The model can be used

1) to confirm some idea or theory regarding the relationship between variables

2) to predict some general, continuous response function from discrete and possibly relatively few measurements.

The most common application of regression analysis in analytical laboratories is undoubtedly curve-fitting, where the creation of calibration lines from data is obtained from instrumental methods of analysis. In many cases, more than one variable may be measured. For example, multiwavelength calibration procedures are finding increasing applications in analytical spectrometry and multivariate regression analysis forms the basis for many chemometric methods reported in the literature.

1.7 Univariate regression

In its univariate form, a linear calibration model is $y=a+b\cdot x+e$ where y_i represents the response (for example an instrumental signal like an absorbance) of the *ith* calibration sample, x_i denotes the corresponding descriptive variable (in analytical chemistry often a concentration), *a* and *b* are the constant coefficients characteristic of the regression line (intercept and slope), and e_i signifies the error associated with the *i*-th calibration sample, assumed to be normally distributed random, *N* (0,1). A single instrument response, *e.g.* absorbance at a single wavelength, is measured for each calibration sample: the vecto rs **x**, **y** and **e** will contain the corresponding values . In vectorial terms

y=a+b**x**+**e**

where **y** , **x** , and **e** are *n* dimensional vectors for *n* calibration samples

Values in **y** and **x** are used to estimate the model parameter *b* by the least squares procedure. The method is based on minimizing the Root (Residual) Sum of Square (RSS).

In matricial terms, **b** (vector of coefficients) is computed by:

$$\hat{b} = (X^T X)^{-1} X^T y \qquad \text{Eq.1.1}$$

where a column of ones is added to vector **x** to take into account the intercept term

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & \\ . & \\ 1 & x_n \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix}$$

In Eq. 1.1 the symbol $\hat{b}$, called "b-hat," is used to highlight its role as an estimate of b. The resulting calibration model is used to predict the analyte concentration for an unknown sample

$$\hat{y} = x_{unk}\hat{b} + a$$

where x_{unk} represents the response for the unknown sample measured at the calibrated wavelength. This kind of calibration is called univariate calibration because there are just one descriptor and one response variable.

1.8 Multivariate case

Univariate calibration is specific to situations where the instrument response depends only on the target analyte concentration. In multivariate systems, responses depend on the target analyte in addition to other chemical or physical variables and, hence, model parameters must take into account these interfering and/or multiple effects. Assuming a linear relation for the *i*-th calibration sample, the model can be written as

$y=b_0+x_{i1}b_1+x_{i2}b_2+\ldots+x_{ij}b_j+e_i$

$$\begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{1\,1} & x_{12} & \ldots & x_{1n} \\ 1 & x_{12} & x_{22} & \ldots & \\ \ldots & & \ldots & & \\ \ldots & & \ldots & & \\ 1 & x_{1m} & x_{12} & \ldots & x_{1n} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \ldots \\ b_n \end{bmatrix}$$

the equation becomes

y=**Xb**+**e** Eq. 1.2

the same equation of univariate is used in computing the vector of coefficients.

To obtain an estimate of the regression vector **b** by use of Eq1.2, *i.e.* to ensure that the inverse $(\mathbf{X}\mathrm{T}\mathbf{X})^{-1}$ exists, the determinant of $(\mathbf{X}^{\mathrm{T}}\mathbf{X})$ must be different from zero. When in the X matrix there is complete collinearity, $(\mathbf{X}^{\mathrm{T}}\mathbf{X})$ is singular, its determinant is zero, so that it is not invertible. Even if only partial collinearity is present, the determinant will be very small and estimated coefficient will not be accurate. The issue can be easily understood if we consider that in the least square method partial derivative are computed. This means that regression coefficients provide an estimate of the effect of a one unit change in an independent variable, holding the other variables constant. If in the given data set one variable x_1 is highly correlated with another independent variable,

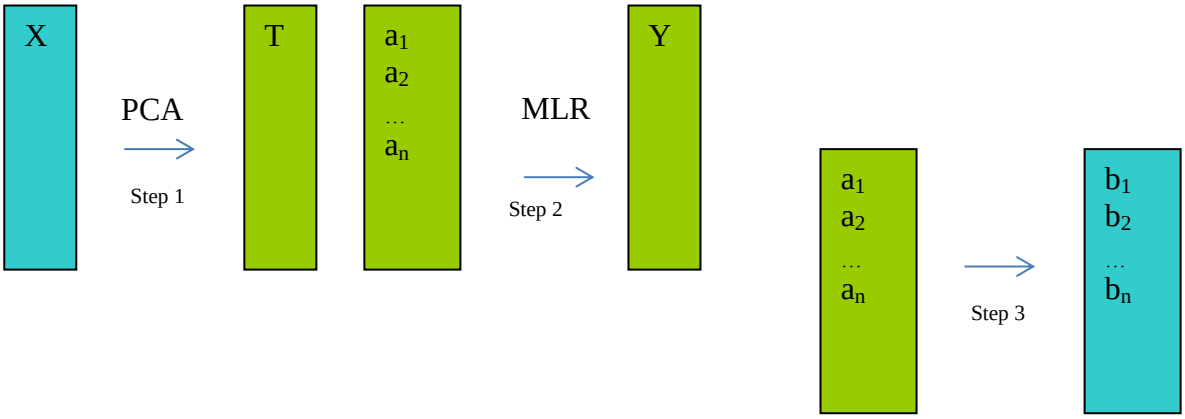
x_2, then we have a set of observations for which x_1 and x_2 have a particular linear stochastic relationship. We don't have a set of observations for which all changes in x_1 are independent from x_2. This means that we have an imprecise estimate of the effect of independent changes in x_1 because its collinear variables contain the same information. In addition, the standard errors of the coefficients tend to be large, and small changes to the input data can lead to large changes in the model, even resulting in changes of sign of parameter estimates. This is the collinearity problem in regression, and estimates can be seriously degraded. Thus, selection of specific descriptive variables to be included in the model is critical to the performance of the model. This aspect will be further discussed in the shrinkage section.

Furthermore, we can have more than just one response: in this case, the model begin $\mathbf{Y} = \mathbf{X}\,\mathbf{b} + \mathbf{E}$, where $\mathbf{Y}$ is $n \times q$ matrix for q responses: solution for the regression matrix is still obtained by Eq.1.2, with $\mathbf{Y}$ replacing $\mathbf{y}$.

1.9 PCR

The key concept on the basis of calculation of principal components, *i.e.* of generating orthogonal linear combinations of variables in order to extract maximum information from a dataset, is also useful in treatment of regression. As mentioned above. it is often the case with multiple regression analysis involving large numbers of independent variables that there exists extensive collinearity or correlation between these variables. Collinearity adds redundancy to the regression model since it happens that are more variables included in the model, for adequate predictive performance. To avoid collinearity, the regression coefficients should be orthogonal (*i.e.* independent each other). Among the methods available to the analytical chemist for regression analysis with protection against the problems caused by correlation between variables, principal components regression (PCR) is the most commonly employed. Besides bypassing multicollinearity, the method results in estimation and prediction better than ordinary least squares when successfully used. With this method, the original k descriptive variables are transformed into a new set of orthogonal or uncorrelated variables called principal components of the correlation matrix. This transformation ranks the new orthogonal variables, ordering them basing on their significance, and the procedure then allows eliminating some of the principal components to realize a reduction in variance. After elimination of the least important principal components, a multiple regression

analysis of the response variable against the reduced set of principal components is performed using ordinary least squares estimation (ordinary least square, OLS, or multivariate linear regression, MLR). Because the principal components are orthogonal, they are pair-wise independent and hence OLS can be used. Once the regression coefficients for the reduced set of orthogonal variables have been calculated, they are mathematically transformed into a new set of coefficients that correspond to the original or initial correlated set of variables. These new coefficients are principal component estimators .



1.10 PLS

In principle, MLR can be used with very numerous factors. However, if the number of factors gets too large (for example, greater than the number of observations), surely we will get a model that fits the sampled data perfectly but that will fail to well predict new data. This phenomenon is called over-fitting. In such cases, although there are many manifest factors, there may be only a few underlying or latent factors that account for most of the variation in the response. The general idea of PLS is to try to extract these latent factors, accounting for as much of the manifest factor variation as possible while modeling the responses well. For this reason, the acronym PLS besides Partial Least Regression has also been taken to mean Projection to Latent Structure.

Partial least squares (PLS) was first developed by H. Wold in the field of econometrics in the late 1960s. Two different methods are available, called PLS1 and PLS2. In PLS1,

separate calibration models are built for each column in **Y**. With PLS2, one calibration model is built for all columns of **Y** simultaneously.

The basic principle in PLS is that modeling the response (*y*) information, is as important as modeling the descriptors (or *x*) information. In performing PCR the PCs are calculated only on the X matrix, and do not take account of the Y matrix. In PLS, components are obtained using together *x* and *y* data. In other words, PLS maximizes the covariance between the two set of variables instead of the variance of the *x* variables as happens in PCR. PLS finds a variable that maximizes *xy*, or the product of the independent variables data with the responses. In physical terms, PLS assumes that there are errors in both blocks which are of equal weight. This is reasonable: in spectrophotometric measurement, for example, the concentrations used in calibration are subject to error (*e.g.* dilution and weighing) just as much as the spectra. MLR and PCR as commonly applied in chemistry assume that all the errors are in the measured data and that the concentrations in the calibration set are unaffected by errors.

1.11 PLS matricial algorithm

In PLS, the matrix of descriptors **X** is decomposed in a similar mode to principal component analysis, generating a matrix of scores, **T**, and loadings or factors, **P**. The same decomposition is performed for the response matrix **Y**, producing a matrix of scores, **U**, and loadings, **Q**.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathrm{T}} + \mathbf{F}$$

The goal of PLS is to model all the variables belonging to **X** and **Y** so that the residuals for the **X** block, **E**, and the residuals for the **Y** block, **F**, tend to zero. An inner relationship is also created that relates the scores of the **X** block to the scores of the **Y** block

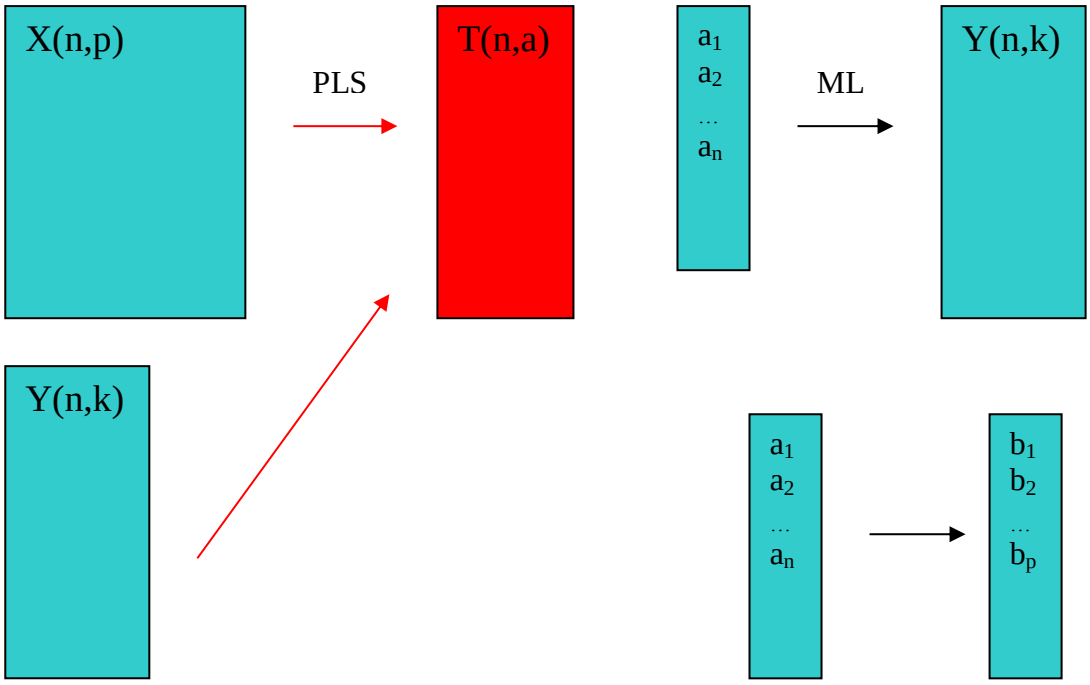
$$\mathbf{U} = \mathbf{T}\mathbf{W}$$

The above mentioned model is improved by establishing the so-called inner relationship. Because latent vectors at the beginning are calculated for both blocks independently, they will have only a weak relation to each other. The inner relation is improved by exchanging the scores, **T** and **U**, in an iterative calculation. By this iterative process information from one block is used to adjust the orientation of the

latent vectors in the other block, and *vice versa*. Once the complete model is calculated, the equations can be combined to give a matrix of regression vectors, one for each component in **Y**:

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}$$

The complementary use of information from **X** and **Y** makes PLS algorithm more problematic than PCR. However, PLS can allow to develop better regression vectors, *i.e.*, more parsimonious with respect to the bias/variance trade-off. Some authors also report that PLS can sometime provide more acceptable solutions when compared to PCR. Other authors have reported that PLS has a greater tendency to overfit noisy **Y** data. It is often reported in the literature that PLS is preferred because it uses fewer factors than PCR and, hence, provides a more parsimonious model.



Similarly to other techniques, the quality of the models can be determined by the size of the errors, so usually normally the sum of squares of **E** and **F** is calculated. The number of significant PLS components can be estimated according to the size of these errors, often using cross-validation, the bootstrap or test sets, although it can be done on the training set (often called autoprediction)

1.12 High-dimensional regression shortcomings and possible solutions

It is known that the aim of variable selection is to select the optimal variables subset that can improve the prediction performance and make the calibration reliable when carried out prior to a multivariate calibration method like PLS and PCR. Some applications have also validated the positive effect of variable selection in the case of high-complexity samples. Additionally, a large body of literature has been devoted to the method of variable selection in a different way.

A basic topic in the present study concerns treatment of spectroscopic data. Spectral measurement are a typical case of high dimensional problem, because undoubtedly the number of variables exceeds the number of observations. Moreover, spectroscopic variables are highly correlated. From a statistic point of view, high-dimensional regression problems are changeling because they cannot be solved by classical estimation procedures like the method of ordinary least squares. As discussed previously, the standard procedures require $\mathbf{X}^T\mathbf{X}$ to be nonsingular, otherwise $\mathbf{X}^T\mathbf{X}$ cannot be inverted and the parameters cannot be uniquely estimated. This obviously is not possible when variables>observations, as the covariate matrix does not have full column rank.

In addition to the problem arising from multicollinearity, OLS method is affected by further issues. In linear regression the goal is to build a prediction model with two properties

1. Low bias: it must be able to minimize the error in prediction for data belonging to training set
2. Low variance: it must be able to minimize the error in prediction for data not belonging to the training set

Classical OLS usually generates models with low bias but high variance hence a trade-off is supposed to be considered.

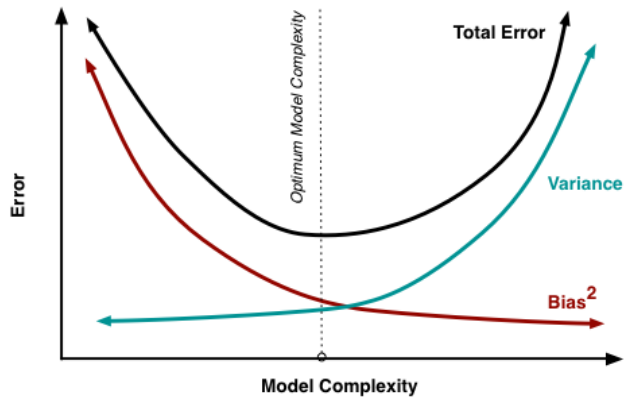


Fig.1.2

Thus some kind of strategies or alternative methods to classical regression are needed. There are a number of both simple and more advanced methods available, on the basis of occurrence. The most intuitive approach is maybe the preselection, that is to simply pick out a smaller subset of the covariates (≤ number of original variables) based on a certain relevant criterion and fit the model to these covariates only. This can, however, result dangerous because it might exclude relevant variables. Another approach is to use methods like principal components regression or partial least squares. These methods provide a small number of linear combinations of the original explanatory variables, and these "latent" variables are used instead of the original ones. This may be reasonable for prediction purposes, but models are often difficult to interpret (Hastie et al., 2009) because a problem arises: relating one or more descriptors to a dependent variable which we call the response.

1.13 Alternative regression techniques

Several methods known as Shrinkage or Penalization techniques have been proposed to improve prediction accuracy and interpretation of OLS. For example, ridge regression (Hoerl and Kennard,1970), also know as Tikhonov regularization, minimizes the residual sum of squares subject to a bound on the L_2-norm of the coefficients. In linear algebra the norm is a function that that assigns a strictly positive length or size to each vector: the L_2 or Euclidean norm for a *n*-dimensional vector is

$$\| x \| := \sqrt{x_1^2 + x_2^2 + ..x_n^2}$$

In ridge regression estimate β is the value of β that minimizes the function

$$\sum_i (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

Where solutions are indexed by tuning parameter λ . For every choice of λ, we have a ridge estimate of the coefficients of the regression equation: Y = Xβ(λ)+ε.

As a continuous shrinkage method, ridge regression achieves its better prediction performance through a bias–variance trade-off. Ridge regression often achieves better prediction accuracy by shrinking the OLS coefficients, particularly in the highly correlated predictor situation. However, ridge regression cannot produce a parsimonious model, because it always keeps all the predictors in the model. Best subset selection in contrast produces a so-called sparse model, but it is extremely variable because of its inherent discreteness A promising technique called the LASSO (Least Absolute Shrinkage and Selection) was proposed by Tibshirani (1996). The LASSO is a penalized least squares method imposing an L_1-penalty on the regression coefficients. The L_1 (Taxicab or Manhattan norm) norm is

$$\| x \|_1 := \sum_{i=1}^{n} | x_i |$$

The L_1 penalty appeared to have advantageous properties that could be exploited with great benefit in high-dimensional regression problems, and it is in the $p >> n$ (where p is number of variables and n is number of objects) problems that the LASSO methods have really proven their superiority compared to other existing methods.

Owing to the nature of the L_1-penalty, the LASSO does both continuous shrinkage and automatic variable selection simultaneously.

Tibshirani (1996) compared the prediction performance of the LASSO and ridge regression and found that none of them uniformly dominates the other. However, as variable selection becomes increasingly important in modern data analysis, the LASSO is much more appealing owing to its sparse representation. Although the LASSO has shown success in many situations, it has some limitations.

(a) In the $p>n$ case, the LASSO selects at most n variables before it saturates, because of the nature of the convex optimization problem. This seems to be a limiting feature for a

variable selection method. Moreover, the LASSO is not well defined unless the bound on the L_1-norm of the coefficients is smaller than a certain value.

(b) If there is a group of variables among which the pairwise correlations are very high, then the LASSO tends to select only one variable from the group and does not care which one is selected.

(c) For usual $n>p$ situations, if there are high correlations between predictors, it has been empirically observed that the prediction performance of the LASSO is dominated by ridge regression (Tibshirani, 1996).

1.14 Classical calibration and inverse calibration issues: the aim of the present work

Thanks to mathematical and statistical procedures, in multivariate calibration it is possible to relate many measured variables (like wavelengths) to properties of interest (like concentration values). We can create a predictive model, and then use that model to predict the same properties from the measured variables of unknown objects (samples). Based on the features and approaches of calibration, multivariate calibration can be divided into two classes: classical calibration (including multiple linear regression, MLR) and inverse calibration (including method like principal component regression, PCR, and partial least square, PLS) .

In the former the signal (*e.g.* spectral data) is the dependent variable and concentration is the descriptor. In the inverse case, concentration is expressed as function of the signal.

In the case of classical calibration, creating a model requires knowledge of pure spectra of all components. The spectra of mixture samples (the response values) are a linear combination of the pure spectra of all components (the independent variables). Based on the Beer–Lambert law or the similar in analytical chemistry, the classical calibration model can directly make predictions using least squares without any further validation like cross-validation, which is also called 'solid modeling'. Based on the available information of the pure spectra of interferential species, Lorber introduced the concept of the Net Analyte Signal (NAS) with the help of orthogonal projection, which can be further used as a practical tool to obtain the fundamental analytical figures of merit such as selectivity, signal-to-noise ratio and limit of detection for classical calibration.

In order to overcome the defect of collinearity of the descriptors matrix, the methods PCR and PLS were introduced into multivariate calibration, which made inverse

calibration very appealing and convenient, since it aimed directly at building the relationship between measured spectra of the samples and the concentrations (or even properties) of the analytes of interest.

Aside from the overcoming of the problem of collinearity, there are several advantages of chemical modeling by latent variables. Firstly, the original variables are replaced by a few latent variables or principal components of larger variance, which consist of linear combinations of the original variables. Only this fact makes the dimensional reduction of datasets of many variables possible. Secondly, the variable selection among the latent variables becomes much easier, since the latent variables are orthogonal with each other. Thus, the cross-validation technique can be easily introduced into PCR and PLS modeling to avoid the overfitting problem. It can also be proven that the measurement noise can be reduced if the number of selected principal components are just equal to the number of coexisting chemical species in a mixture system, which makes the resolution of a completely unknown mixture system (black analytical systems) possible. Despite all of these significant advantages, the use of inverse calibration is not applicable in some situations. As a matter of fact, in carrying out Standard Addition Method to a multivariate system we get from PLS algorithm a function where the unknown variable must be calculated by extrapolation. This means that we need the signal of the matrix (the sample without analyte). Specially when environmental samples are analysed this is not possible, hence an alternative strategy is required.

1.15 Net Analyte Signal

The essential issue in quantitative analytical chemistry is that of assessing within some acceptable approximation the wanted chemical or physical property for a sample of interest. Most of such approximations are made using theoretically or empirically calibration functions. Although these functions may be interpreted without problems for univariate analytical measurements such as pH or single-wavelength absorbance, highly multivariate calibration functions, like those used in spectroscopy, can be much less easy. In his work of 1986, Lorber provided an smart interpretation to these multivariate calibration functions with his concept of the Net Analyte Signal (NAS) vector in linear additive systems. Geometrically, the NAS vector represents the portion of the pure analyte signal vector that resides in a space orthogonal to the pure-component signals of all interfering species in a linear additive system (Fig. 1.3)

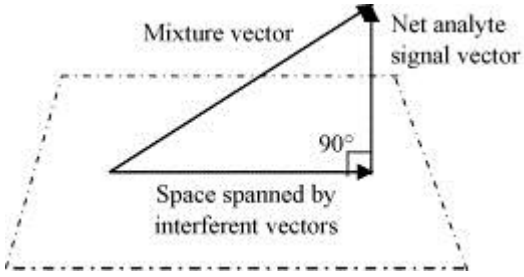


Fig. 1.3

Since the NAS vector indicates a direction only affected by changes in the analyte concentration, it can be used as fully selective property predictor. Sanchez and Kowalski (1988), and later Booksh and Kowalski (1994), employed the NAS concept for a larger discussion of analytical calibration theory and figures of merit in linear multivariate calibration. From Lorber's original work, they derived expressions for calculating the NAS vector(s), based on the pure-component signals of the chemical components in the system, both in classical and inverse calibration models. In the following years, several modifications of such calculations were proposed. The relation of the NAS concept to analytical figures became the basis of a considerable body of literature on multivariate measures of selectivity, sensitivity, limit of detection, and multivariate signal-to-noise ratios.

1.16 NAS algorithm and related issues

In multivariate inverse calibration a regression model is searched to predict the response *y* of size Ix1 from the multivariate measurements in X (IxJ). The regression vector **b** (Jx1) is found to minimize the residuals **e** (Ix1) in the equation

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$$

and additional constraints are often imposed. For example, in principal component regression (PCR), X is constrained to be in the subspace of the first principal components, whereas in partial least squares regression (PLS) different criteria are used. The NAS vector is a suitable tool that allows to perform calculations in a similar way as in univariate regression. Although net analyte signals have originally been thought in relation to spectral data and mainly in settings where Beer's law is assumed to be valid,

the principle is both general and applicable, and useful for all multivariate calibration models.

The basic idea of NAS method is to separate the contributions in the calibration data matrix X into one originating solely from the analyte of interest (called X_k to indicate that the *k*-th analyte is the analyte of interest) and another from other sources of variability such as interfering species (X_{-k})

$$X=X_k+ X_{-k}$$

X_k denotes the unique part of the analyte signal and X_{-k} is the matrix describing the signal orthogonal to that. From a matrix spanning the space of the interfering (X_{-k}), the net analyte signal vector of a sample *i* is calculated by

$$x^*_k=[I- X_{-k} X^+_{-k}] x_i$$

where x_i is the spectrum of the *i*-th sample, I is the JxJ identity matrix and $^+$ denotes a pseudoinverse. The net analyte signal is usually taken as the norm of $x^*_{k;i}$ and can be used similarly to a univariate signal in 'pseudo-univariate' linear regression.

The matrix $[I – (X^+_{\_k})X_{-k}]$ projects the calibration spectra onto the space orthogonal to that spanned by the spectra of all analytes except the sought *k*-th analyte. Thus, in order to find the NAS vector of a certain analyte, it is necessary to find this projection matrix: calculation of $[I – (X^+_{\_k})X_{-k}]$ involves finding the matrix describing the interfering spectra, X_{-k}. This makes the calculation of X_{-k} the key step of the procedure. There are several ways to estimate this matrix. Lorber et al. (1997) suggested a method that uses PCR or PLS. First, the calibration matrix X is rebuilt using a set of significant components obtained by PCR or PLS, yielding X_{reb}. Then a rank annihilation step in the*n*-dimensional space is used for finding the part of the original matrix spanned by the interfering species:

$$X_{-k} = X_{reb} - \alpha \hat{c}_k x^T$$

where $\hat{c}_k$ is the projection of the vector of responses c_k (Ix1) onto the *n*-dimensional subspace and is given by

$$\hat{c}_k = X_{reb} X^+_{reb} c_k$$

The vector x is a linear combination of the rows of X, which is chosen to include a contribution from the spectrum of the *k*-th analyte. Any reasonable spectrum can be used for this purpose, though it is recommended to use a spectrum that contains maximal information on the analyte. The scalar α can be calculated as

$$\alpha = \frac{1}{x^T} X^+ \hat{c}_k$$

Other approaches were proposed by Xu and Schechter (1997) who introduced another approach where c_k is used to define X_{-k}. The calibration matrix X is scaled by dividing each spectral vector of matrix X by the corresponding c_k-value such that each spectral vector contains the same contribution of the analyte

$$x_{i,sc} = \frac{x_i}{c_i}$$

In the next step the average of the scaled vectors is calculated and subtracted from all the scaled vectors. This gives a mean-centering pre-treatment of the scaled matrix, thus removing the constant contribution of the analyte:

$$x_{-k,i} = x_{i,sc} - 1^T \bar{x}_{sc}$$

where 1 is a J vector of unitary values. Combining x_{-ki} for all samples provides an estimate of X_{-k}. A similar approach is described by Goicoechea and Olivieri (1999) The mean calibration spectrum is obtained as

$$\bar{x} = \frac{1}{I} \sum_{i=1}^{I} x_i$$

the contribution of the analyte is then subtracted from the data matrix

$$X_{-k} = X - \frac{c_k \bar{x}^T}{\bar{c}}$$

where $\bar{c}$ denotes the mean calibration concentration of the analyte. Goicoechea and Olivieri (2001) proposed to define X_{-k} as the projection of X orthogonal to c_k as illustrated in Equation (10).

$$X_{-k} = [I - c_k (c_k^T c_k)^{-1} c_k^T] X$$

Faber [8] put forward an idea which does not require the calculation of X_{-k}. By this method. the NAS vector is calculated from the regression vector as

$$x_{k,i}^{*} = b(b^{T}b)^{-1}c_{i}$$

The rationale for this development was to circumvent the computational burden of some of the prior methods, and it was argued that the new method was an alternative that gave similar results than the older methods.

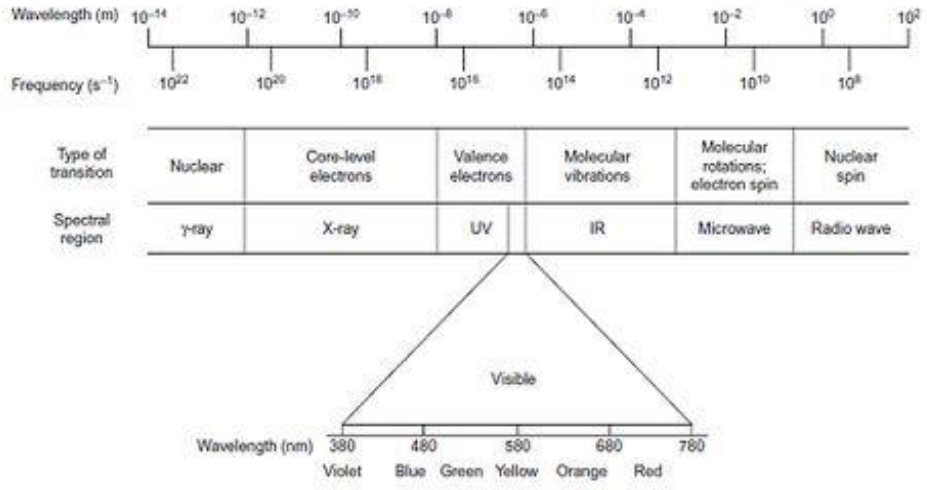
Section 2

2.1 Analytical Infrared Spectroscopy

Spectroscopy is the analysis of the interaction between matter and any portion of the electromagnetic spectrum. Traditionally, the first application of spectroscopy to analytical chemistry involved the visible spectrum of light, but X-ray, IR and UV spectroscopy also soon became valuable analytical techniques. Spectroscopy may involve any interaction between light and matter, including absorption, emission, scattering etc. Absorption processes, including those of vibration and rotation of molecules associated with infrared spectroscopy, can be represented in terms of quantized discrete energy levels E_0, E_1, E_2, etc. Each atom or molecule in a system must exist in one or other of these levels. Whenever a molecule is shined by radiation of proper energy, a quantum of energy (or photon) is either emitted or absorbed. In each case, the energy of the quantum of radiation must exactly fit the energy gap E_1-E_0 or E_2-E_1, etc. The energy of the quantum is related to the frequency by the following relation:

$$\Delta E = h\nu \qquad \text{Eq. 2.1}$$

For each processes of interaction a specific portion of the spectrum is involved



An unit which is widely used in infrared spectroscopy is the wavenumber, ν, in cm^{-1}. This is the number of waves in a length of one centimeter and is given by the following relationship

$\nu = 1/\lambda = \nu/c$

An increase in energy is equal to an increase in wavenumber.

The infrared range of the electromagnetic spectrum covers the wavelength region from 0.8 μm to 1 mm, which conforms to the wavenumber range 12500-10 cm^{-1}. It is split in the Near-IR (NIR), the Mid-IR (MIR) and the Far-IR (FIR) region (Fig. 2.1). It is neighbor to the visible region on one side and the microwave region on the other.

Region of IR	Wave length (μm)	Wave number (cm^{-1})
Near Ir (Overtone region)	0.8-2.5	12,500-4000
Mid Ir (Vibration-rotation region)	2.5-50	4000-200
Far Ir (Rotation region)	50-1000	200-10
Most used	2.5-25	4000-400

Fig. 2.1

Infrared spectroscopy is a very important non-destructive technique for gaining structural information and identifying the chemical bonds in unknown compounds. This information is important for qualitative as well as quantitative determination of the chemical compounds and is used in various scientific areas either in research activity or in the private/applicative sectors.

Applications fields have subjects such as:

- · Food analysis: additives, preservatives, colorants
- · Environmental analysis: water, atmospheric particles, gases
- · Forensic science: paints, textiles, cosmetics,

- · Semiconductor analysis
- · Pharmaceuticals.
- · Multilayer compounds: polymers, paintings, films
- · Geological samples: inclusions in stones

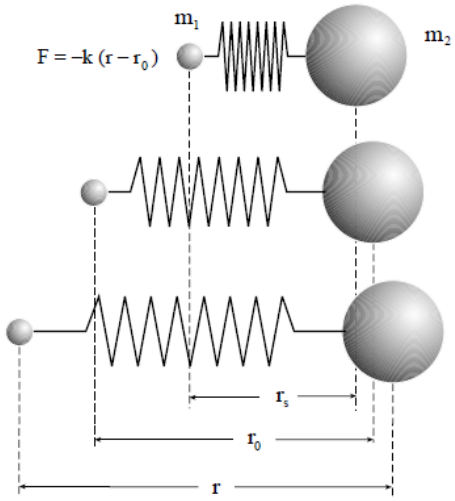
One of the great advantages of infrared spectroscopy is that it is suitable for analysing the majority of samples, no matter the matrix in which the analyte is included. Liquids, solutions, pastes, powders, films, fibers, gases and surfaces can all be examined with a proper choice of sampling technique. As a consequence of the improved instrumentation, a variety of new sensitive techniques have now been developed in order to examine formerly intractable samples.

2.2. Basics in IR spectroscopy : vibro-rotational transitions

In order to better understand the application of IR spectroscopy to the research issues presented in this thesis, some basic recalls of IR spectroscopy may be useful to readers with low expertise in physical chemistry.

In any molecule, it is known that atoms or groups of atoms are connected by bonds. These bonds are in a continuous motion in a molecule, as a result they maintain some vibrations with some frequency, characteristic to every portion of molecule. This is called natural frequency of vibration. A suitable model for describing molecular vibrations is the harmonic oscillator.

From quantum mechanics we know that vibrational transitions in a molecule occur between distinct vibrational energy levels. Both intramolecular and intermolecular vibrations contribute to the spectrum. The simplest possible situation is a vibration between two atoms of a diatomic molecule. The vibration between two molecules can be expressed by Hook's spring law. The atoms are located at an average inter-nuclear distance *r*, the bond length. An attempt to bring the atoms more closely together will lead to a rapid increase of the repulsive force between the two atoms. An attempt to pull them apart is resisted by the attractive force. Both displacements require an input of energy which can be described as a function of the distance between the two atoms.



The molecular bond between two atoms is equivalent to a spring between two spheres of the masses m1 and m2. The force of compression or expansion is then given by Hook's law:

$F=-k\,(r-r_0)$

k is a constant and $r-r_0$ is the distance difference to the equilibrium distance generated by the force F. The force is directed against the displacement of the atoms and thus has a negative sign. The potential energy of the oscillating system

$$E=\frac{1}{2}k(r-r_0)^2$$

increases symmetrically when the distance between both atoms is decreased or increased compared to the equilibrium distance.

The vibration of such a diatomic molecule is characterized by an oscillation frequency that is given by classical mechanics:

$$\nu_{vib}=\frac{1}{2\pi}\sqrt{\frac{k}{\mu}}$$

Eq. 2.2

where μ is the reduced mass of the system:

$$\mu = \frac{m_1 \cdot m_2}{m_1 + m_2}$$

From Eq 2.2. we see that the oscillation frequency is dependent on the force constant (k) and on the masses of the oscillating atoms. The larger the mass of an atom, the smaller the frequency of the vibration.

We also know that in contrast to classical mechanics, vibrational energies of molecules are quantized like all other molecular energies.

$$E_n = \left(n + \frac{1}{2}\right) h\nu_{vib}$$

where n is called the vibrational quantum number. From the above equation arises that the lowest vibrational energy is $E = ½\ h\ \nu$ when n=0. Therefore, a molecule can never have null vibrational energy, *i.e.* the atoms can never be without a vibrational motion.

The quantity $E = ½\ h\ \nu$ is called the zero-point energy; it depends on the classical oscillation frequency and hence on the strength of the chemical bond and on the masses of atoms that participate in this bond.

Further application of the quantum mechanic laws leads to a simple selection rule for the harmonic oscillator undergoing vibrational changes:

$\Delta\nu = \pm 1$

and the energy of a transition between two vibrational states is then given by

$\Delta E = h\ \Delta\nu$

For arising absorption, the frequency of the radiation must be identical to the frequency of the vibration.

Indispensable requirement for a molecule to show infrared absorptions is that it must possess an electric dipole moment. This is the selection rule for infrared spectroscopy. Fig. 2.2 illustrates an example of an ‘infrared-active’ molecule, a heteronuclear diatomic molecule. The dipole moment of such a molecule changes as the bond expands and contracts. By comparison, an example of an ‘infrared-inactive’ molecule is a homonuclear diatomicmolecule because its dipole moment remains zero no matter how long the bond.

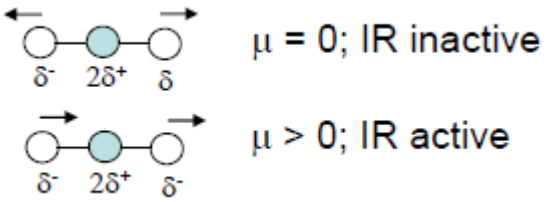


Fig. 2.2

We can sum up selection rules in IR spectroscopy

1) Change in dipole moment. Since no net change in dipole moment occurs during the vibration or rotation of homo nuclear species (i.e. O_2 ,N_2 ,Cl_2) such compounds cannot absorb IR radiation

2) Applied IR frequency should be equal to the natural frequency of radiation. Otherwise compound do not give IR absorption peaks.

When in a molecule N atoms are present, it can be described by three spatial coordinates (x, y, z) of each atom in space. The total number of such coordinates is therefore 3N and the molecule has 3N degrees of freedom. In such a description, the position of the molecule and the bond-angles are fixed. The translation of the molecule in space is described by three degrees of translational freedom. In addition to translation, for the rotational motion of a nonlinear molecule we need additional three degrees of freedom, while for a linear molecule just two (the rotation around the bond axis of a linear molecule does not result in a change of the coordinates of the atoms). Consequently, a non-linear molecule must have 3N-3-3=3N-6 (non-linear molecule) 3N-3-2 = 3N-5 (linear molecule) degrees of freedom of internal vibration. A linear molecule like CO_2 therefore has 4 vibrational degrees of freedom, while H_2O (non-linear) has only 3 vibrational degrees of freedom. The number of vibrations derived in this way are called the number of fundamental vibrations or the normal modes of vibration. There are symmetrical and asymmetrical vibrations depending on the maintenance of symmetry in the molecule. The internal molecular vibration can be subdivided into stretching and bending vibrations. Change in the inter-atomic distance along the bond axis is called symmetric and asymmetric stretching vibrations and can be seen in Fig. 2.3

Fig. 2.3

If the motion does not involve changes along the bond axis but involves changes between bond angles the vibration is called bending. Here there are two types of bending vibrations: in-plane (ip) bending vibrations and out-of-plane (oop) bending vibrations. The (ip)-bending vibrations are thus named because they happen in the plane of the molecule. Anti-symmetric and symmetric (ip)-bending vibrations are known as rocking and scissoring vibrations. The (oop)-bending vibrations are a change in bond angles across the plane of the molecule and these vibrations are known as wagging and twisting bending vibrations (Fig. 2.4)

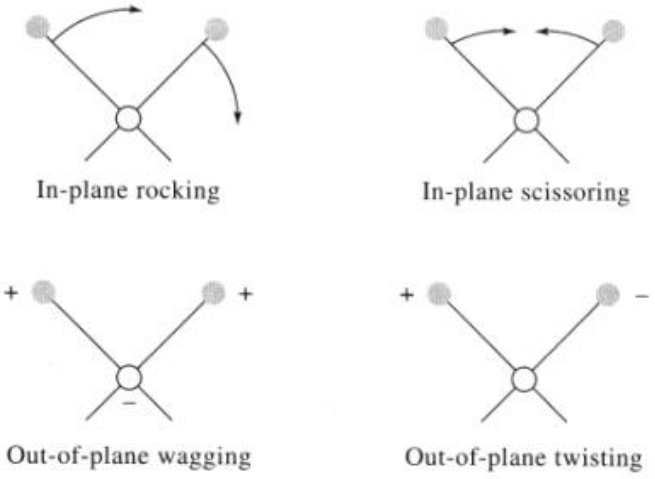


Fig. 2.4

2.3 Overtones and IR spectra complexity: the anharmonic oscillator

Harmonic oscillator model is an approximation that cannot totally describe IR spectra. In fact, molecules do not simply behave like to masses that are connected by a spring and described by the harmonic motion. Although chemical bonds are elastic they do not obey Hook's law totally. They can break if they are stretch beyond a limiting distance and the molecule will dissociate. When amplitudes of vibration exceed a certain value, the system must be described differently. An empiric description of such behavior was given by P.M. Morse and is called the Morse function. However, also the Morse function is an approximation and more accurate descriptions require additional cubic and terms of higher order and other harmonicity constants. The magnitude of these additional an harmonicity constants is rapidly getting smaller with increased order of the an harmonicity term. The selection rules for the anharmonic oscillator are $\Delta\nu = \pm 1$, ± 2, ± 3, ... so that they are the same as for the harmonic oscillator, with further possibility of larger jumps. However, the intensity of the resonance line decreases strongly with the difference in the energy levels and transitions with $\Delta\nu = \pm 3$ or higher are only seldom observed.

Chiefly transitions of $2h\nu$ or $3h\nu$ are sometimes observed, these transitions are called overtones or combination band.

2.4 IR spectrum as fingerprint of molecules

From quantum mechanics and the oscillator models, we know now that the internal vibrations observed in molecules are fundamental vibrations, which are quantized in specific energy levels. When energy in the form of IR radiation is irradiating a sample and when applied IR frequency = natural frequency of vibration, absorption of IR radiation takes place and a peak is observed. Every bond or portion of a molecule requires different frequency for absorption. Hence characteristic peaks are observed for every functional group or part of molecule. In other words, IR spectra is nothing but a fingerprint of a molecule. Fig. 2.5 shows absorption ranges of the most common chemical bonds observed in IR spectroscopy

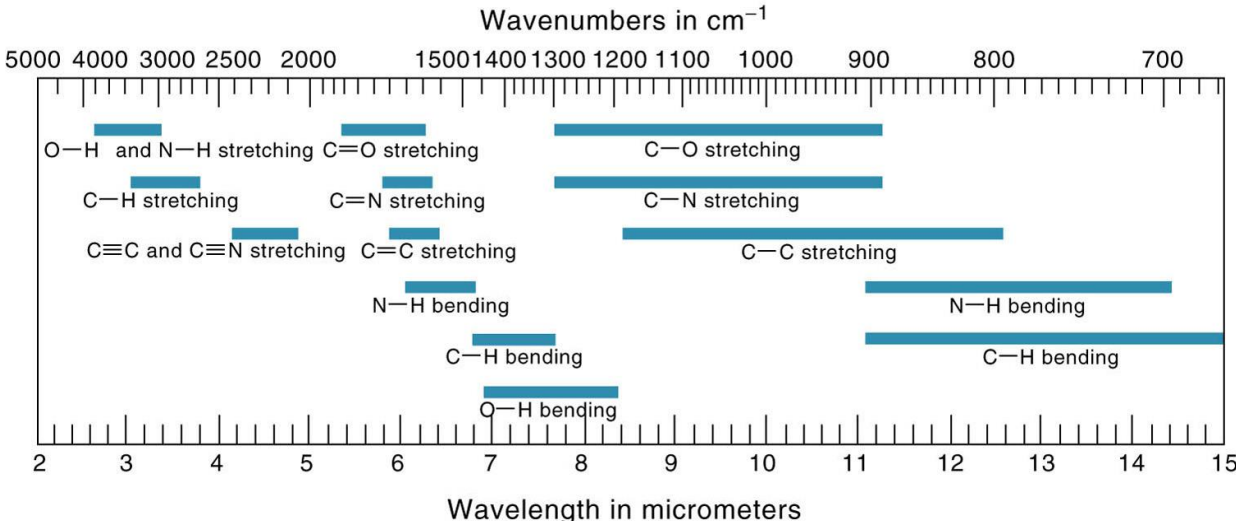


Fig. 2.5

Some features are common to all spectra

- Stretching frequencies are higher than corresponding bending frequencies. It is easier to bend a bond than to stretch or compress it
- Bonds to hydrogen have higher stretching frequencies than those to heavier atoms. In fact, the energy difference in chemical bond excitation, is caused by the atoms involved in the bond. With increased atomic mass the bond length between the atoms will be increasingly longer, and the vibrational excitation energy becomes lower
- Triple bonds have higher stretching frequencies than corresponding double bonds, which in turn have higher frequencies than single bonds

2.5 Transmission and Reflectance

IR measurement can be performed using either transmission or reflectance setup. The theory behind them is quite different considering that they are both based on the absorption. In transmission spectroscopy an infrared spectrum is commonly obtained by passing infrared radiation through a sample and determining what fraction of the incident radiation is absorbed at a particular wavelength, which corresponds to the vibration frequency of a part of the molecule. The sample is normally embedded in a substrate, KBr for the MIR region and in CsI for the FIR region, as these are inert in the respective regions. Infrared transmission spectroscopy relies on Lambert-Beer's law (empiric). Lambert's law states that the fraction of the incident light absorbed is

independent of the intensity of the source. Beer's law states that the absorption is proportional to the number of absorbing molecules. In combining the two law's we express the Lambert-Beer Law:

$$A = \log_{10} \frac{I_0}{I} = \varepsilon l c$$

where A is absorbance, c is the concentration of the sample, I is the path length travelled by the light through the sample and ε is the absorptivity (also know as extinction coefficient). The absorptivity is a peculiar physical property of the specific molecule under investigation and is function of wavenumber.

For historical reasons in some cases transmittance T is used, because when grating or prismatic instruments were employed, they gave a directly readable output-signal as a difference between sample and no-sample in the pathway of the IR beam

$$T = \frac{I}{I_0}$$

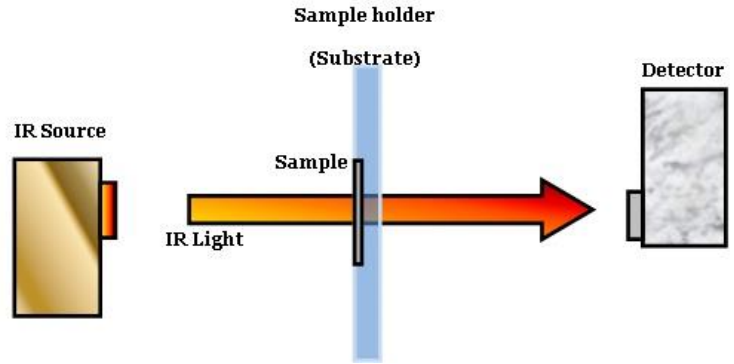
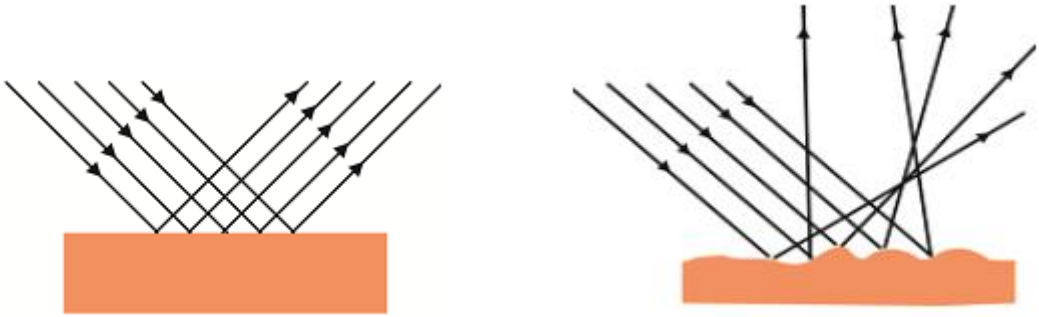


Fig. 2.6

Aside from the conventional IR spectroscopy of measuring light transmitted from the sample (Fig. 2.6), the reflection IR spectroscopy was developed using combination of IR spectroscopy with reflection theories. In the reflection spectroscopy techniques, the absorption properties of a sample can be extracted from the reflected light. IR reflectance techniques can be divided into two categories: external reflection and internal reflection. External reflection covers two different types of reflection: specular (regular) reflection and diffuse reflection (Fig. 2.7). The former is usually associated

with reflection from smooth, polished surfaces like mirror, while the latter is associated with the reflection from rough surfaces. In internal reflection method, interaction of the electromagnetic radiation on the interface between the sample and a medium with a higher refraction index is studied.



specular reflection diffuse reflection

Fig. 2.7

Internal Reflectance Spectroscopy (IRS) theories were independently built by Jacques Fahrenfortand (1959) and N.J.Harrrick, (1967) and soon became popular techniques due to the possibility to be employed in a wide range of applications. IRS is also known as Attenuated Total Reflectance (ATR) and was emplyed in this work. ATR technique is a surface analytical technique. An Internal Reflecting Element (IRE) is used to focus and direct the light beam to the investigated surface.

The concept of internal reflection spectroscopy originates from total internal reflection phenomenon. An internal reflection occurs when a beam of radiation enters from a more dense medium (with a higher refractive index, n_1) into a less-dense medium (with a lower refractive index, n_2), The fraction of the incident beam reflected increases as the angle of incidence rises. When the angle of incidence is greater than the critical angle θ_c, all incident radiations are completely reflected at the interface, results in total internal reflection (Fig. 2.8).

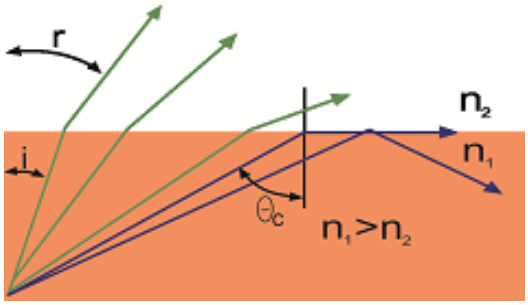


Fig. 2.8

In ATR spectroscopy a crystal with a high refractive index and excellent IR transmitting properties is used as IRE and is placed in close contact with the sample. When the angle of incidence at the interface exceeds the critical angle θ_c, the beam of radiation propagating in IRE undergoes total internal reflection at the interface IRE/sample. Total internal reflection of the light at the interface between two media of different refractive index creates an "evanescent wave" penetrating into the medium of lower refractive index that decays exponentially with distance from the boundary

$$I_{ev} = I_0 \exp\left[-\frac{z}{d_p}\right]$$

where z is the distance normal to the optical interface, d_p is the penetration depth (distance through which the evanescent wave travels path length), and I_0 is the intensity at z = 0.

d_p is given by

$$d_p = \frac{\lambda}{2\pi\sqrt{\sin^2\theta - \left(n_2/n_1\right)}}$$

where n_1 and n_2 are the reflective index of the rarer medium and the denser medium An ATR spectrum arises by reflection of the evanescent wave with the sample. When an absorbing material undergoes ATR measurement, the evanescent wave will be absorbed by the sample and its intensity is reduced in regions of the IR spectrum where the sample absorbs, thus, the intensity of reflected beam is attenuated. As can be seen in the formula of d_p, the resultant attenuated radiation is a function of wavelength so that, at

longer wavelength, the evanescent wave penetrates deeper into the sample. Consequently, the absorption bands at longer wavelengths will be relatively more intense than those shorter wavelengths. Therefore, an ATR spectrum compared with a transmission spectrum is similar except or the band intensities at longer wavelengths.

Additionally, compared to the transmission spectrum, small differences may be seen in an ATR spectrum which arise from dispersion effects (variation of refractive index of a material with change of wavelength). An anomalous dispersion produces changes in refractive index and in penetration depth through an absorption band. The penetration depth changes strongly at wavelength in which the dispersion is the highest. Other differences may occur due to the surface effects between the sample and internal reflection element (IRE crystal). For instance, the degree of physical contact between IRE and the sample influences the sensitivity of an ATR spectrum. Since the evanescent wave only propagates 2-15 μm beyond the surface of the crystal, thus, an intimate contact of the IRE with the sample is essential.

Material	Refractive Index ($1000cm^{-1}$)	Spectral Range (cm^{-1})
Zinc selenide	2.4	20,000-650
AMTIR (As/Ge/Se glass)	2.5	11,000-750
Germanium	4.0	5,500-870
KRS-5 ($TlI_2/TlBr_2$)	2.37	20,000-350
Zinc sulfide (ZnS)	2.2	17,000-950
Cadmium telluride (CdTe)	2.65	10,000-450
Sapphire (Al2O3)	1.74	25,000-1800
Cubic Zirconia (ZrO2)	2.15	25,000-1800
Diamond	2.4	45,000-2500; 1650-<200

IRE materials used in ATR applications

2.6 Fourier Transform

A great significant advance in infrared spectroscopy, occurred as a result of the introduction of Fourier-transform spectrometers. This type of instrument employs an interferometer and exploits the well known mathematical process of Fourier-transformation. Fourier-transform infrared spectroscopy (FTIR) has remarkably improved the quality of IR spectroscopy and reduced the time required in performing

measurement. In addition, with the introduction of computers, infrared spectroscopy has gained further improvements

The goal of any absorption spectroscopy (IR, UV-visible spectroscopy etc.) is to measure how much a sample absorbs light at each wavelength. The most straightforward way to do this, the "dispersive spectroscopy" technique, is to shine a monochromatic light beam at a sample, measure how much of the light is absorbed, and repeat for each different wavelength: this is how some UV-Vis spectrometers work. The monochromatic light beam is obtained through a dispersive element (a prism or a grating) that separates the frequencies. One of the major disadvantages of the dispersive spectroscopy to produce a spectrum is its slowness. As matter of fact, each wavelength of the spectrum has to be recorded separately. The frequency is spread smoothly across the whole range of the spectrum, then the detector signal is monitored and recorded. Lately, a very different method of obtaining an infrared spectrum replaced the dispersive instruments. Fourier transform infrared spectrometers are now predominantly used and have simplified the acquisition of infrared spectra dramatically. Instead of shining the sample with monochromatic beam of light, this technique shines a beam containing many frequencies of light at once, and measures how much of that beam is absorbed by the sample. Thanks to Fourier transform spectroscopy, simultaneous and almost instantaneous a recording of the whole spectrum in the magnetic resonance, microwave and infrared regions is possible.

As simple example we can consider a radiation composed by different waves with two different frequencies. A detector receiving such radiation will show an oscillating signal due to the frequencies of the two superimposed waves, but also to a periodic change in the amplitude which slowly increases and decreases (Fig. 2.9)

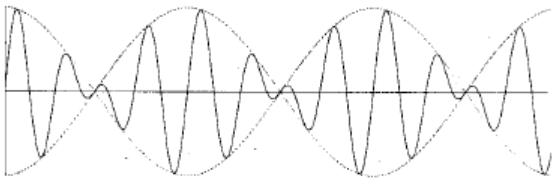


Fig. 2.9

The periodic change in the amplitude arises from the interaction of the two components; when the amplitudes reinforce each other (the waves are “in step”) or cancel each other the (the waves are “out of step”). To resolve a combined wave such as those shown above into its components require to evaluate four unknowns. As matter of fact each component wave has its own frequency and maximum amplitude. Adding more than two waves makes the situation even more complicated. The Fourier transform process provides a simple and general way to resolve a complex wave into its frequency components.

An other advantage of the Fourier transformation technique is that the whole radiation emitted by the source is monitored. Conversely, in dispersive technique monochromator discards most of the radiation, therefore Fourier transform spectrometers have a higher sensitivity than traditional spectrometers.

Modern spectrometers, in particular those used in infrared spectroscopy, today almost always make use of Fourier transform techniques to record the spectrum. The heart of the Fourier transform infrared spectrometer is the interferometer, a device for analyzing the frequencies present in a composite signal.

The purpose of an interferometer is to split a beam of light into two beams, ensuring that one of the light beams travel a different distance than the other. Fourier transform infrared (FTIR) spectroscopy is based on the concept of the interference of radiation between the two beams that produces an interferogram. Interferogram will show a signal that is function of the difference of path length between the two beams and the mathematical method of Fourier-transformation can convert the two domains of distance and frequency. When the two beams are recombined, the difference in the path leads to a phase difference between them, and they interfere either constructively or destructively, depending on the difference in path lengths. The detected signal oscillates as the two components alternately come into and out of phase as the path difference is changed. The radiation emitted by the source passes through an interferometer and then to the sample before reaching a detector. The detected signal is amplified, then the data are converted to digital by an analog-to-digital converter and transferred to the computer. The most common interferometer used in FTIR spectrometry is a Michelson interferometer that consists of a fixed mirror and a movable mirror located at a right angle to each other and oriented perpendicularly, with a semi-reflecting film, the

beamsplitter, placed at the vertex of the right angle and oriented at a 45°angle relative to the two mirrors (Fig. 2.10)

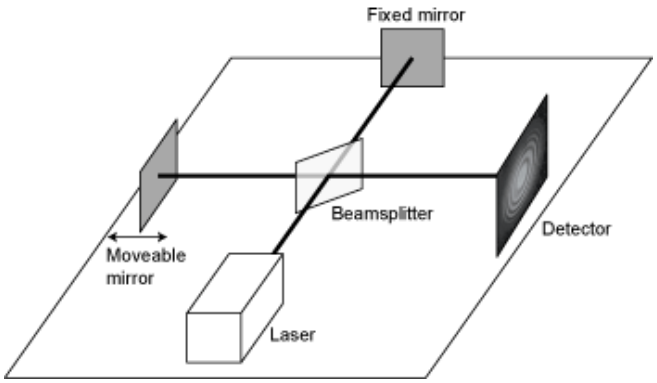
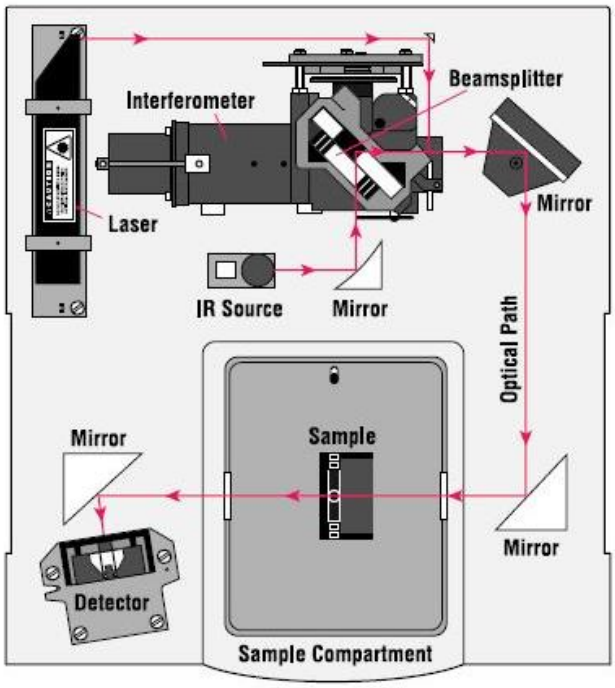


Fig. 2.10

The beamsplitter material has to be chosen according to the region to be examined. Materials such as germanium or iron oxide are coated onto an 'infrared-transparent' substrate such as potassium bromide or cesium iodide to produce beam splitters for the mid or near infrared regions. Thin organic films, such as polyethylene terephthalate, are used in the far-infrared region. The beam splitter splits the incident beam in two equal beams. When a collimated beam is passed into an ideal beamsplitter, half of the incident radiation will be reflected to one of the mirrors (*i.e.* the translating one) while half will be transmitted to the fixed mirror. The beams are both reflected by these mirrors, and return to the beamsplitter where they recombine and interfere. Fifty percent of the beam reflected from the fixed mirror is transmitted through the beamsplitter while the other fifty percent is reflected back in the direction of the source. The beam which emerges from the interferometer at 90° to the input beam is called the transmitted beam, and this is the beam detected in FTIR spectrometry. The moving mirror produces an optical path difference between the two arms of the interferometer. If the translating mirror and the fixed mirror are the same distance from the beam splitter, the distance travelled by the two light beams are the same. This is called Zero Path Difference. If the path difference is $(n + ½)$, the two beams interfere destructively in the case of the transmitted beam and constructively in the case of the reflected beam.



Spectrometer Layout

2.7 Preprocessing

Building a practically useful calibration model using spectroscopic data involves preprocessing and data manipulation. No doubt that the first step consists in *baseline correction*. It is usual in quantitative infrared spectroscopy to use a baseline joining the points of lowest absorbance on a peak, preferably in reproducibly flat parts of the absorption line. The absorbance difference between the baseline and the top of the band is then used. However, there are many other harmful effect that must be approached with more elaborated techniques.

After proper data collection, pre-processing of spectral data is the most important step before chemometric modeling (*e.g.*, Principal Component Analysis or Partial Least Squares). Regarding solid samples, when undesired systematic variations occur, they are mostly due to light scattering and differences in the effective path-length. Such undesired variations often constitute the major part of the total variation in the sample set, and can be observed as shifts in baseline (addictive and multiplicative effects) and other phenomena called non-linearities. Scattering is a general physical process in which light in the form of propagating energy is spread owing to non-uniformities in the

medium in which radiation propagates. Light scattering can be thought as the deflection of a ray from a straight path, for example by interactions with the propagation medium, particles, or in the interface between two media. For instance, deviations from the law of reflection due to irregularities on a surface are also usually considered to be a form of scattering. When these irregularities are considered to be random and dense enough that their individual effects average out, this kind of scattered reflection is commonly referred to as diffuse reflection. Hence electromagnetic waves are scattered by a system owing to its heterogeneity, whether on the molecular scale or on the scale of aggregations of many molecules. In any case, physical reasons of scattering are the same for all systems. Matter is composed of electrically charged particles: electrons and protons. When one molecule is shined by an electromagnetic wave (Fig 2.11) the electron orbits are perturbed periodically with the same frequency as the electric field of the incident wave. Such perturbation of the electron cloud produces a periodic separation of charge within the molecule, which is called an induced dipole moment. The oscillating induced dipole moment in turn becomes source of electromagnetic radiation, thereby generating scattering phenomenon. The greater part of light scattered by the particle is emitted at the identical frequency of the incident light, a process referred to as elastic scattering. In addition to reradiating electromagnetic energy, the excited elementary particles may transform part of the incident energy into other forms (thermal energy, for example), according to a process called absorption underlying the spectroscopy. In summary, light scattering can be thought as a complex interaction between the incident electromagnetic wave and the molecular/atomic structure of the scattering object.

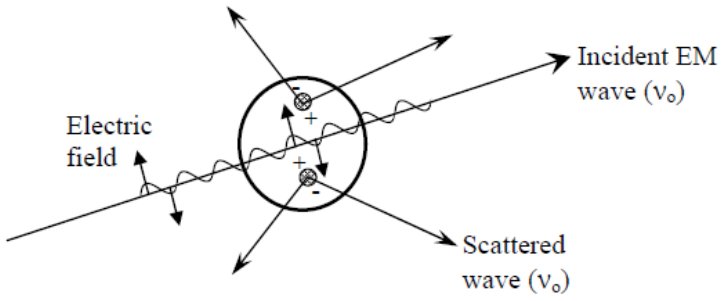


Fig. 2.11

The most important forms of light scattering (in which energy transfer with the sample is not included) are Rayleigh and Lorentz-Mie (Fig. 2.12). Both are processes where the electromagnetic radiation is scattered (*e.g.*, by small particles, bubbles, surface roughness, crystalline defects and so on). Rayleigh scattering is, strictly speaking as originally formulated, applicable to small, dielectric (non-absorbing), spherical particles: the theory is strongly wavelength dependent and occurs when the particles are much smaller in diameter than the wavelength of the electromagnetic radiation. On the other hand, Mie scattering theory includes the general spherical scattering solution (absorbing or non-absorbing) without a particular reference to particle size. Therefore, Mie scattering theory has no size limitations and converges to the limit of geometric optics for large particles. Consequently, Mie theory is more general and can be used for describing most spherical particle scattering systems, including Rayleigh scattering. However, Rayleigh scattering theory is generally preferred if applicable, due to the complexity of the Mie scattering formulation.

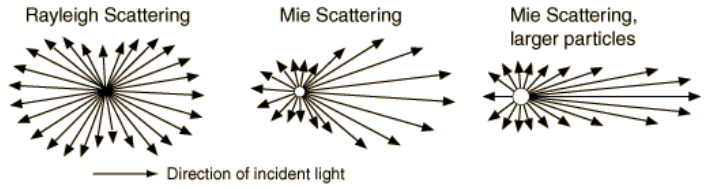


Fig. 2.12

In the case of environmental samples, the scattering effects make spectra very complicated: so spectral preprocessing techniques are demanded to remove the scatter and obtain a pure, suitable absorbance spectrum.

All pre-processing techniques have the goal of reducing variability in the data in order to highlight the supposed relationship underlying the phenomenon of interest. By using a proper pre-processing technique this can be achieved, but there is always the risk of applying the wrong type or applying a too severe pre-processing that will remove the valuable information. The proper choice of pre-processing method is difficult to assess but, in general, performing several pre-processing steps is not advisable, and, as a minimum requirement, pre-processing should maintain or decrease the effective model complexity. The most generally used pre-processing procedures (in both reflectance

and transmittance mode) can be divided into two categories: scatter correction methods and spectral derivatives. Scatter corrective pre-processing methods includes Multiplicative Scatter Correction (MSC), Inverse MSC, Extended MSC, Extended Inverse MSC, Standard Normal Variate (SNV) and normalization. The main spectral derivation techniques are: Norris-Williams (NW) derivatives and Savitzky-Golay polynomial derivative filters which use a smoothing of the spectra prior to calculating the derivative in order to lower the harmful effects on the signal-to-noise ratio that conventional derivatives methods involve. The aims of the pre-processing treatment are primarily to improve a subsequent exploratory analysis and to allow a building of a calibration model. In this work, spectroscopic data arise from reflectance measurement so, in order to build a model, they are supposed to obey to Lambert-Beer law, an empirical equation that suggest a linear relationship between the absorbance and the concentrations of the constituent

$$A_\lambda \lambda = -\log_{10}(T) = \varepsilon_\lambda lc$$

where A_k is the wavelength-dependent absorbance, T is the light transmittance, $\square_k$ is the wavelength-dependent molar absorptivity, l is the effective path length of the light through the sample matrix and c is the concentration of the constituent of interest. Lambert-Beer's was originally derived for transmittance systems. In reflectance measurements, it can be redefined in analogy to transmittance measurements as:

$$A_\lambda \lambda = -\log_{10}(R) \cong \varepsilon_\lambda lc$$

where R is the detected reflectance.

Since reflectance spectra, the subject of the present work, are highly affected by scattering, the following method were here applied: MSC, SNV and normalization. As matter of fact, these techniques were in the beginning designed to reduce the (physical) variability between samples due to scatter. All three also adjust for baseline shifts between samples. Multiplicative Scatter Correction (MSC) is probably the most widely used anti-scattering technique. MSC in its basic form was first introduced by Martens et al. in 1983 and further elaborated on by Geladi et al. in 1985. MSC method relies on the assumption that a corrected spectrum (x_{corr}) can be expressed by original spectrum

(x_{org}) according the equation $x_{org}=b_0+b_1x_{corr}$ where x_{org} is the recorded spectrum and x_{corr} is the "corrected" one.

$$x_{corr} = \frac{x_{org} - b_0}{b_1}$$

In turn the x_{org} is related to a reference spectrum by

$$x_{org} = b_0 + b_{ref} x_{ref} + e$$

in which e is the un-modeled part of x_{org}. Hence the scalar correction parameters b_0 and b_1 are found as the intercept and the slope of the line obtained by a least square regression plotting x_{org} against a reference spectrum. Correction coefficients b_0 and b_{ref} account for additive and multiplicative contributions in original spectrum. and differ for each sample. Fig. 2.13 illustrates the interpretation of the scalar parameters

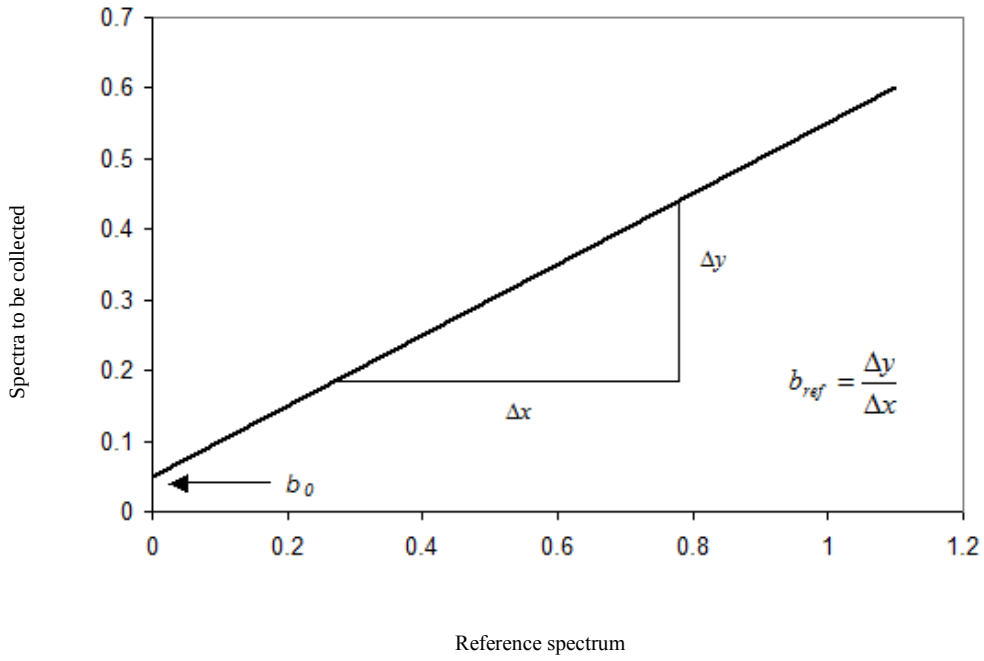


Fig. 2.13

Therefore, in order to obtain a corrected spectrum, a reference spectrum is needed. In most applications the average spectrum of the calibration set is used as the reference spectrum. However, reference spectrum can also be screened according with different criteria. In the original paper by Martens et al. (1983), it was suggested to use only those parts of the spectral axis that do not include relevant information. Although this makes

good spectroscopic sense, it is difficult to determine such regions in practice, especially in ATR measurements where the signals from different chemical components are strongly overlapping and correlated, and little or no true baseline is found. This is the reason why, in most cases, the entire spectrum is used to find the scalar correction parameters in MSC. The above described is the basic form of MSC, that can be expanded into more elaborate forms where more sophisticated corrections can be performed using higher order polynomial fitting (EMSC) or other transformations of the wavelength dependency.

Further pre-processing method applied in this study for scatter correction are Normalization and Standard Normal Variate (SNV). Their underlying basic concept is the same as that for the traditional MSC: $x_{corr} = (x_{org} - a_0)/ a_1$. For SNV, a_0 is the average value of the sample spectrum to be corrected, while for Normalization a_0 is set equal to zero. Another difference between SNV and Normalization lies in a_1: in the former, this parameter is the standard deviation of the sample spectrum while in Normalization different vector-norms can be used, for instance total sum of the absolute values of the elements in the vector (so-called Taxicab norm) or the square root of sum of the squared elements (Euclidian norm). Other options that are sometimes used are normalizing to the maximum absorbance variable and normalizing towards a single selected wavelength. Both these last options should be used with caution, since they can have undesirable effects on the subsequent analysis in cases of noisy data. It should be noted that in Normalization and SNV, unlike MSC, a reference signal is usually not required: each observation is processed on its own, isolated from the rest of the set. The lack of need for a common reference might be a practical advantage.

Spectral derivatives techniques have the capability to remove both additive and multiplicative effects in the spectra and are widely used in analytical spectroscopy. Alternatively methods to scatteing-correction, based on spectral derivatives (like Norris-Williams derivation and Savitzky-Golay derivation) were attempted but no significant improvement was observed, so they will not discussed in detail.

Section 3

3.1 Marine sediments: tipology and composition

The sediments deposited in the ocean are an archive of historical information about the Earth. As a matter of fact, their distribution in the ocean is determined by biological and chemical processes and they provide information about global biogeochemical cycles. The constituents of a marine sediment are often classified according to their origin.

Detrital: brought into the ocean from outside, are subdivided in:

- Terrigenous: those where the ultimate source is weathering and erosion of rocks on land. The materials composing these sediments are introduced to the ocean by water, wind or ice.

- Volcanic: composed of minerals brought into the ocean mostly by wind, as dust and ash from volcanic eruptions.

- Cosmogenic particles that arrive from outer space and survive the Earth's atmosphere to enter the sedimentary record.

Authigenic: these components are oceanic inorganic minerals that precipitate directly from the seawater. These minerals makeup only a small fraction of deep-sea sediments, but in special environments and certain geological times, they comprise the bulk of the sedimentary sequence

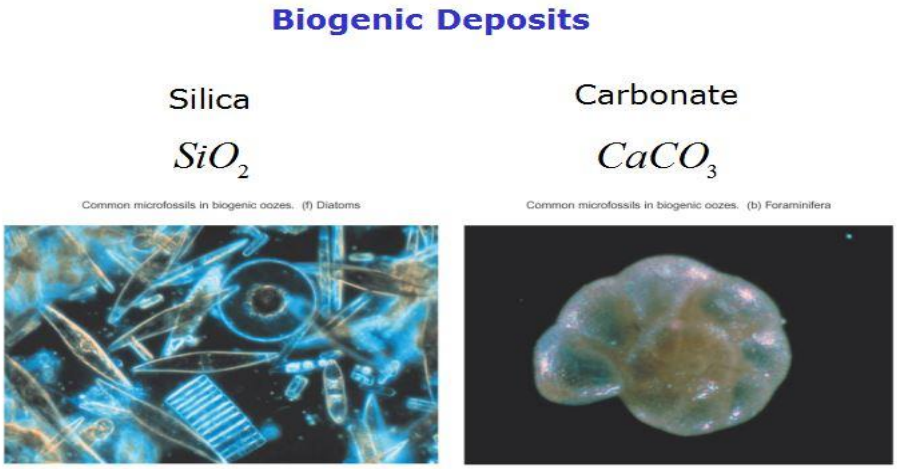
Biogenic: these are among the most important constituents of marine sediments. These sediments are widespread on the sea floor, covering one half of the shelves and more than one half of the deep ocean bottom (total ~55%). They constitute ~30% of total volume of sediment being deposited. As the name implies, these form directly or indirectly through biological activity and can be subdivided in inorganic and organic matter.

- Inorganic are made of a variety of delicate and intricate structures mostly of structural part remains of marine phytoplankton and zooplankton. The life span of most of these organisms is on the order of weeks, so there is a slow continuous "rain" of their remains down through the water column to build successive layers of sediment. The distribution of these sediments depend on the abundance of organisms

- Organic: although not a mineral, organic matter is an important component of biogenic sediments. There are a lot of organisms that do not form hard parts and what is preserved from them is the organic matter. On average in the ocean about only 1% of the organic matter that sinks to the bottom of the ocean is preserved. The amount of organic matter preserved in the sediment depends on how much is produced and the preservation efficiency.

3.2 Composition of deep ocean sediment

In this work we are interested in deep ocean sediment that take great contribution from biogenic sediments. Biogenic sediments, which are defined as containing skeletal remains of marine organisms, cover approximately 55% of the deep ocean floor. Clay minerals make up most of the non-biogenic constituents of remaining matter. The most important biogenic minerals are of calcareous and siliceous origin. Calcareous sediments are composed principally of calcite or aragonite arising from the remains of organisms like plankton with calcium-based skeletons (also called tests), such as *coccolithophores* (plants) and *foraminifera* (animals), while siliceous sediment are formed from the remains (hard part like frustles and spicules) of organisms with silica-based skeletons like *diatoms* (plants) or *radiolarians* (animals).



When the organisms floating in the surface waters die, they settle to the bottom of the sea. When these tests make up greater than 30% of the sediment it is called an ooze. Distributions and accumulation of biogenic oozes in oceanic sediments depend on rates of production of biogenic particles in the surface waters and dissolution rates of those particles either in the water column or after they reach the bottom. Biogenic oozes accumulate very slowly in the deep ocean. This is because the surface waters of the central oceans are very poor in the nutrients (mostly land-derived), such as nitrogen and phosphorus, that are required by the surface sea creatures. Therefore, these waters are inhabited by only small populations which contribute very slowly to the development of the deep ocean sediment accumulation. Moreover, in some regions of the oceans the tests of these organisms re-dissolve before they reach the bottom. In these regions the sediments are dominated by abyssal clays.

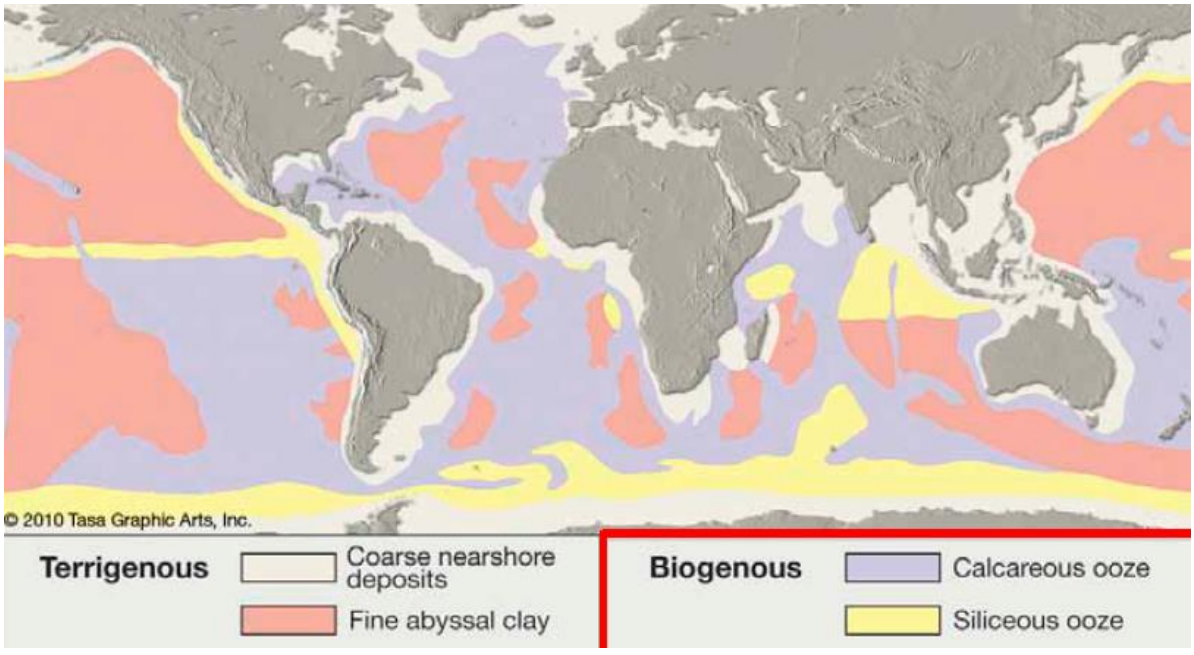


Fig. 3.1 Sediment deposit distribution in sea floor

3.3 Behavior of calcareous and siliceous compounds in seawater

Conditions favoring deposition of silica or calcium carbonate are quite different. Calcareous oozes are never found deeper than about 4,000 to 5,000 meters because the calcium dissolves at deeper depths. Silica, even though is under-saturated in the oceans, is less under-saturated in deep water. The patterns of carbonate and silica deposits reflect different processes of formation and preservation, so that carbonate oozes that are poor in biogenic silica and *vice versa* (Fig. 3.1)

3.4 Carbon

The marine carbon cycle involves the production and recycling of two types of carbon-rich materials: organic matter and carbonates (inorganic). Over 95% of oceanic carbon is in the form of inorganic dissolved carbon (Fig. 3.2). The remainder is comprised of various forms of organic carbon, namely living organic matter as well as particulate and dissolved organic carbon. The primary processes responsible for variations in the deep sea CO_2-carbonic acid system are oxidative degradation of organic matter, dissolution of calcium carbonate, and oceanic circulation patterns. Temperature and salinity variations in deep seawaters are small and of secondary importance compared to the major variations in pressure with depth.

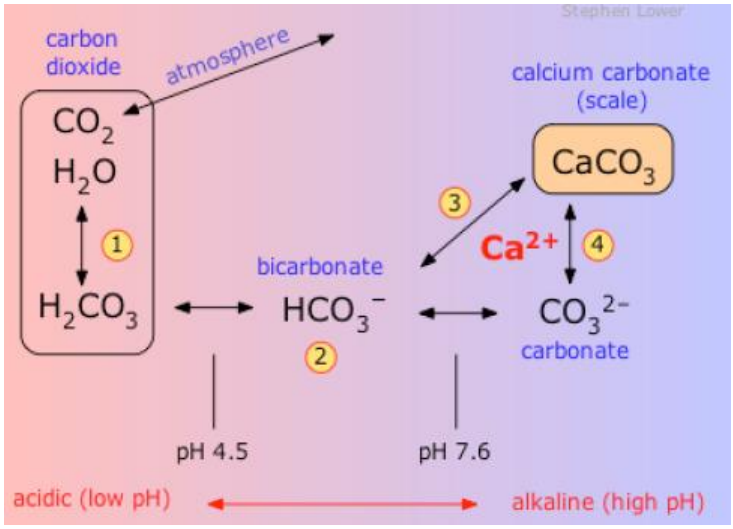


Fig. 3.2 Inorganic carbon equilibria

In warm tropical surface waters $CaCO_3$ does not readily dissolve. However, in colder deeper waters the presence of increased amounts of CO_2 in the water enhances the dissolution of $CaCO_3$ causing the breakdown of calcareous tests, so that bottom waters become more under-saturated in calcium carbonate. Carbonate solubility increase with depth can be explained with physical and chemical effects.

Physical effect: the effect of pressure is the most important environmental variable affecting the solubility product of $CaCO_3$. K_{sp} increases with increasing pressure. Thermodynamically, the pressure effect is related to the partial molar volumes of Ca^{2+}, CO_3^{2-} and $CaCO_3$. Like free energies, the partial molar change for the reaction is calculated from the sum of the products minus the sum of the reactants:

$$\Delta V = V\,Ca^{2+} + V\,CO_3^{2-} - V\,CaCO_3$$

The ΔV for calcite is negative, meaning that the volume occupied by $CaCO_3$ is greater than the combined volume of Ca^{2+} and CO_3^{2-} in solution; so, $CaCO_3$ becomes more soluble with depth. $CaCO_3$ is an unusual mineral in that it is more soluble at lower temperatures (K_{sp} increases with decreasing temperature); the effect is only about 4% over a temperature range of 20°C. Since the temperature range in the deep sea is only a few degrees, temperature effect is less important than pressure effect in dissolving carbonates

Chemical effect: remineralization of organic matter in the water column produces CO_2

$$CH_2O + O_2 \rightarrow CO_2 + H_2O$$

Carbon dioxide and water combine to form carbonic acid which dissolves the $CaCO_3$

$$CaCO_3\,(s) + H_2O\,(l) + CO_2\,(g) = Ca^{2+}\,(aq) + 2\,HCO_3^-\,(aq)$$

An increase of CO_2 at depth lower CO_3^{-2} at depth. In the deep ocean, the decrease in $[CO_3^{-2}]$ from this reaction has a marked effect on $CaCO_3$ solubility. The point where dissolution increases markedly is called lysocline (Fig. 3.3). The depth below which calcareous skeletons dissolve as fast as they accumulate is called Carbonate Compensation Depth (CCD**).** In warm latitudes the CCD occurs at 4-5 kilometers.

Therefore, calcareous oozes will be found only at depths less than 4-5 kilometers. Where the bottom of the ocean is deeper than 4-5 kilometers, calcareous tests will not accumulate. Calcareous oozes, therefore, are found mostly on the oceanic ridges and plateaus.

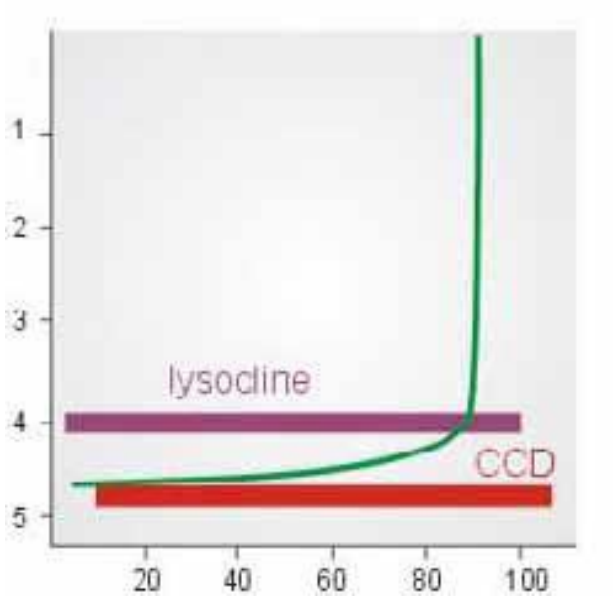


Fig. 3.3

3.5 Silicon

The oceanic silicon cycle is driven by the production of biogenic silica (BSi) by marine organisms, primarily diatoms (Brzezinski et al., 2003; Treguer et al., 1995) that extract silicic acid from seawater to make their structures. During such process the siliceous organisms expend metabolic energy to extract dissolved silicates from seawater and precipitate amorphous silica. Nowhere in the oceans does silica precipitate spontaneously without the intervention of an organism. In fact, silica is thermodynamically unstable under conditions encountered in the water column. Under-saturation of seawater with respect to amorphous silica causes dissolution of the silica hard parts following death of the organisms, thereby recycling most of the dissolved silicates (Hurd, 1972; Nelson et al., 1995; Treguer et al., 1995). Accordingly, the tendency for silica is to dissolve everywhere it occurs in the oceans. Besides silicates arising from decomposition of organic remains, deep ocean waters are rich in nutrients, such as nitrate and phosphate, themselves the result of decomposition of sinking organic matter (mainly detrital of dead plankton) from surface waters. When brought to the

surface, these nutrients are utilized by phytoplankton, along with dissolved CO_2 and light energy from the sun, to produce organic compounds, through the process of photosynthesis. Such phenomenon is called upwelling and involves wind-driven motion of dense, cooler, and usually nutrient-rich water towards the ocean surface, replacing the warmer, usually nutrient-depleted surface water. The nutrient-rich upwelled water stimulates the growth and reproduction of primary producers such as phytoplankton. The only regions in which siliceous oozes are abundant are in regions where the nutrient supply is so large that diatom and radiolarian tests accumulate faster than the seawater can re-dissolve them after death. These regions are along the equator in the central Pacific and in high latitudes near Antarctica. In these area, small amounts of BSi escape dissolution and are buried in marine sediments. Over geological time, the burial flux of BSi exerts a major control on the bio-siliceous productivity of the oceans (DeMaster, 1981; Van Cappellen et al., 2002).

The fundamental difference between carbonate and silicate compounds in sediments is their behavior with respect to pressure and temperature. Contrarily to carbonates, silica solubility increases with decreasing pressure and increasing temperature. The solubility of silica decreases with decreasing temperature by about 30% from 25-5 ºC, so that less silica dissolves when the waters are colder. The solubility increases slightly with pressure, providing some offset to the temperature effect. Although a big share of silica dissolves and only 1-10% of the flux survives dissolution, siliceous sediments are therefore found in zones of high productivity and high sedimentation rates but only below the CCD, where less carbonate dilution occurs and specifically at high latitudes where the water is colder; in such areas, diatom productivity is typically high.

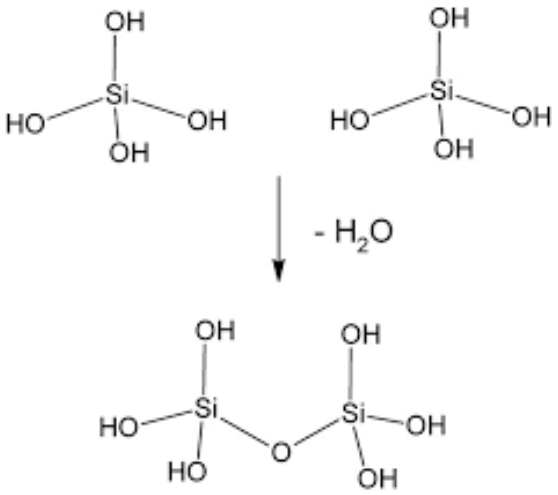
3.6 Origin of BSi: Diatoms

The main contributors to biosilica in today's oceans are diatoms, although radiolarians *silicoflagellates, discoasters* and *sponge spicules* can also contribute. Diatoms are single cellular organisms (algal photo synthesizers) that live at varying depths in the water column. They are abundant in nearly every habitat where water is found: oceans, lakes, streams, mosses, soils and even the bark of trees. These algae form part of the base of aquatic food in marine and freshwater habitats. Diatoms grow as single cells or form filaments and simple colonies. Assemblages of diatom species are often specific to particular habitats and can be used to characterize those habitats. As algae, diatoms are

protists. placed in the division *Bacillariophyta*, which is distinguished by the presence of an inorganic cell wall (called frustle, an hard and porous external layer) composed of hydrated silica. There are an estimated 20,000 to 2 million species of diatom on Earth. This range is so large because scientists are still working to understand basic aspects about "what is a diatom species" and because new and diverse forms are still being discovered and described in scientific publications. Nearly all diatoms are microscopic cells range in size from about 2 µm to about 500 µm. Scientists use light microscopes or scanning electron microscopes (SEM) to visualize diatom structures. When diatoms are observed with a light microscope, the frustules appear clear (we are seeing through glass). When diatoms are obsrved with a scanning electron microscope, the frustules appear opaque.

3.7 Diatoms and Silicon

Diatom distinctive, transparent cell walls are made of silicon dioxide hydrated with a small amount of water ($Si0_2 + H_20$). Silica is the main component of glass and hydrated silica is very like the mineral opal, making these algae, often called "algae in glass houses" more like "algae in opal houses". The biosilica wall of a diatom cell is constructed in a petri-dish like fashion being composed of a top half (epitheca) that overlaps the slightly smaller bottom half (hypotheca). Since silica is impermeable, diatoms have evolved elaborate patterns of perforations in their valves to allow nutrient and waste exchange with the environment. These valve patterns can be quite beautiful and are also helpful for classifying diatoms. Formation of these biosilica structures takes place in specialized intracellular compartments called Silica Deposition Vesicles (SDV). Studies on other silicifying protists have shown that SDV are not a speciality of diatoms but rather represent general organelles for silica biogenesis. The immediate precursor for biosilica formation inside the SDV is unknown, even though monosilicic acid $Si(OH)_4$, which occurs in natural habitats in concentrations between 1 and 100 mM, clearly represents the original source for silica formation. There is general agreement that polymerization, that is, the reactions that result in an increase in molecular weight of the silica, involves the condensation of silanol groups:



Silicic acid polymerization involves three distinct stages. First, monomeric silicic acid polymerizes by condensation of silanol groups to form dimers, trimers, and cyclic oligomers. Oligo silicic acid species have a strong tendency to further polymerize in such a way that siloxane bond (Si–O–Si) formation is maximized. These early processes create highly branched polysilicic acids as nuclei for silica formation. Second, the nuclei grow to form spherical particles either by continuous polymerization with monomeric and oligomeric silicic acids or by fusion of particles. Finally, the silica nanospheres can form a three-dimensional network forming branched particle chains that are cross-linked by siloxane bonds.

3.8 Diatoms as proxy

Microfossils can be used as proxies for climate. When identified by species, they can tell us about the range of oceanic temperatures because every species has specific parameters for survival. Besides temperature they are an effective proxy for climate change due to their sensitivity to a variety of ecological conditions. Therefore, past changes in climate can be inferred from changes in species abundance within a sediment core, as the ecological requirements are well known for a number of ‘indicator’ species. These species are indicative of several variables linked to nutrient availability. These variables are dependent upon a combination of primary factors (for example precipitation, solar output, and wind strength) and secondary factors (*i.e.* upwelling and erosion). For instance, the presence of diatoms themselves tells us one extremely important fact: ice in the form of permanent sheets must not have been present. Diatoms cannot photosynthesize under ice sheets, although some of them live in and on ice

sheets. If they are present in the sediment record, ice sheets must not have extended to this area. By examining the fluctuations in diatom levels we can begin to establish a pattern of warm interglacial periods with high diatom abundance and cooler glacial periods with little or no diatom presence. Assuming a constant sedimentation rate at the coring site, core depth is analogous to time before present. As diatoms bloom and die, their skeletons settle to the bottom of the sea and are incorporated into the sediments. Diatom valves, or skeletons, are made up of silica, which preserves well for a very long time. This allows for the reconstruction of past climate by analyzing the changes in the preserved diatom valves through time.

3.9 Colleting sediment samples: coring

The basic design of the coring devices consists of one or more steel tubes or boxes attached below a lead weight unit. This set-up is dropped to the sea floor and pushed into the sediment to recover a core. There are four main types of coring devices.

The simplest design is the gravity corer, consisting of an up to 20m-long steel tube attached to a lead weight of 1–2 tons. Longer cores can, however, be recovered with the piston corer. Originally invented in 1947, the piston corer has been further developed during the last few decades and is one of the most used coring devices within the marine coring community. Attached to the piston corer there is a trigger arm, which carries a wire with a small weight or a small gravity corer device (trigger corer) extending below the base of the piston corer tube. When the trigger corer penetrates the sea floor to collect the uppermost sediment sequence, the trigger arm is lifted and the piston corer is released falling freely with its own gravity into the sediment. When contact is made with the sediment surface, a piston, located inside the coring tube, is lifted up at the speed of penetration. Such a design reduces the friction inside the tube and allows for the collection of long cores.

Another simpler device is the kasten corer. This coring device also penetrates marine sediments by gravity and consists of long, rectangular boxes. Because of the large volume of sediment sampled, this coring technique is beneficial for multiproxy paleoceanographic studies. Sediment coring is generally accompanied by surface sediment sampling for undisturbed recovery of the sediment/water interface. This is most often achieved using a multicorer, which samples up to 12 individual core. Surface sediment samples may also be obtained using different designs of grabs and box corers,

but generally these do not result in the same quality of sampling as the multicorer. The surface sediment sampling is of importance for understanding modern sediment deposition and the development of reference data sets for paleoceanographic transfer functions. It also provides material for the reconstruction of the most recent ocean history.

3.10 Assessment methods

Techniques to estimate the BSi content of sediments can be classified into 5 broad categories: 1) X-ray diffraction (Goldberg, 1958), 2) point counting of diatoms (Pudsey, 1992), 3) infrared analysis (Fröhlich, 1989), 4) a normative calculation technique whereby BSi is estimated by difference from mineral silicates (Leinen, 1977), and 5) the wet-alkaline digestion techniques (Hurd, 1972; DeMaster, , 1981, 1991; Eggiman et al., 1980; Mortlock & Froelich, 1989; Müller & Schneider, 1993).Wet-alkaline techniques are most often used because of their ease of use and reliability.

The wet-chemical method (DeMaster 1991), in principle, is based on the idea that the dissolution of silica is a surface process requiring the presence of a catalyst. In the case of alkaline leaching, the catalyst is a hydroxyl ion that can be chemisorbed, thereby increasing the coordination number of a silicon atom at the amorphous silica surface to more than four, and as result weakening the oxygen bonds with the underlying oxygen atoms. After the adsorption of OH^-, silicon is released into solution as a silicate ion, $Si(OH)_5^-$, which hydrolyzes to soluble silica, $Si(OH)_4$, once the pH goes below 11. In alkaline leaching procedures, either Na_2CO_3 or NaOH is used to supply the catalytic hydroxyl ion needed as a catalyst**.** Soviet scientists (Bezrukov, 1955) were the first to apply a wet-chemical method for the determination of BSi in marine sediments and suspended material, but the description of this technique is oversimplified and does not provide information on effectiveness of extracting systems or the extent to which silica is dissolved from coexisting aluminosilicates.

Indeed, the big issue in the determination of biogenic silica in marine sediments is how to distinguish BSi dissolved with respect to silica coming from non-BSi compounds. As a matter of fact, marine sediments consist of some reactive silica fractions among which, within the time span of natural leaching experiments, quartz, and alumino-silicate minerals are virtually insoluble, but nevertheless the dissolution of clay minerals severely complicates the determination of the BSi content. To overcome this, many wet

chemical methods have been proposed and tested, but none have been completely acceptable.

For instance, Hurd (1973) prepared a calibration curve constructed from mixtures of typical clay minerals and BSi. The difficulty of this method is to prepare an artificial matrix similar to natural sediments.

DeMaster (1981; 1991) used a sequential alkaline extraction taking samples after 1, 2, 3 and 5h. His results showed that the diatoms are quantitatively dissolved within 2h or less and that Si from clay minerals dissolves at a constant, much slower rate. DeMaster therefore concluded that the S dissolution from clay minerals occurred independently from the dissolution of BSi. Extrapolation of the linear 'clay dissolution line' to time zero would thus correct for the contribution of non-biogenic silica and gives the BSi content of the sample (Fig. 3.4).

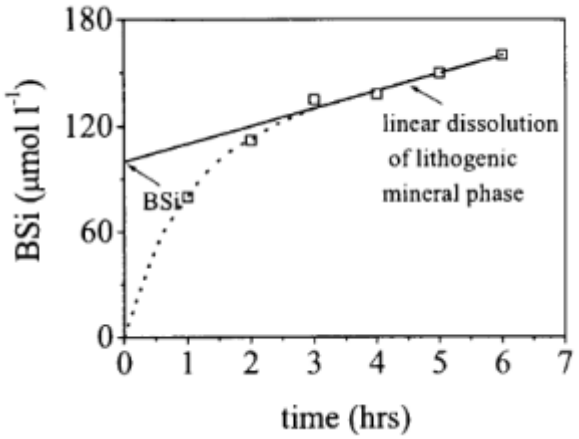


Fig. 3.4 'Ideal' dissolution curve for biogenic silica in sediments as proposed by DeMaster (1981). Extrapolation of the clay line to time zero would subtract the contribution of silica from the lithogenic fraction and would thus give the BSi content

Several corrections and variants to DeMaster procedure were proposed. Mortlock and Froelich (1989) showed that for diatom-rich and clay-poor sediments a single 5h extraction with 2M Na_2CO_3 could be as accurate a measure of biogenic silica as the sequential leaching procedure, although BSi concentrations below 2% were to be regarded with suspicion. In sandy opal-poor sediments from the North Sea, however, the single leach in 2M Na_2CO_3 overestimated the BSi content substantially (Gehlen and van Raaphorst, 1993). In all sequential leaching procedures, correction for the lithogenic mineral phase was made on the basis of a small number of samples, commonly 5 or 6, introducing considerable uncertainties in the intercept determined

from extrapolating of the ‘clay-line’ (Conley,1998). More importantly, this error is determined solely by the composition and quantity of the clay fraction in the sample and can thus be relatively large at low BSi contents. In 1993, Müller and Schneider refined DeMaster’s method to an automated continuous digestion method. They employed an auto-analyzer system that allowed for continuous monitoring of the increase of the dissolved silica concentration in a 1M NaOH leaching solution. The continuous recording of the ‘clay-dissolution line’ improved the accuracy of BSi estimated from the extrapolated intercept.

Ragueneau and Treguer (1994) recognized that for the determination of BSi in suspended matter it is essential to make a correction for the silica leached from lithogenic silicates, and proposed a statistical method to correct for the interference from lithogenic silicates. Recently, Kamatani and Oku (2000) described an alternative approach to correct for the non-biogenic content of the sample. They estimated BSi through linear regression of the extracted SiO_2 plotted against extracted aluminum (Fig. 3.5). The sample was subjected to a sequential digestion for 120 min at 1008C with 0.2 M NaOH solution. The extracted SiO (y-axis) was plotted as a function of the extracted AlO (x-axis), which yields a good linear relationship between the two elements. The straight line is extrapolated to the y-axis, where the intercepting point is used to estimate the BSi content of the sample. However, this is a complicated method and is not recommended for routine work.

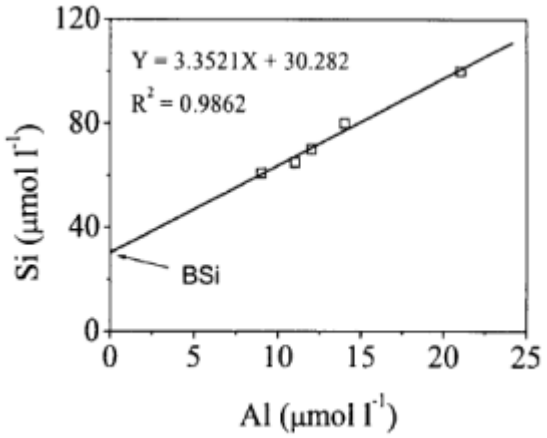


Fig. 3.5 Plot of dissolved silica vs. dissolved aluminum. Kamatami and Oku proposed that all aluminum was associated with lithogenic fraction of the sample. Silica and aluminum dissolve at constant ratio and extrapolating their liner relationship to time zero would subtract the contribution of silica from the lithogenic fraction and would thus give the Bsi content of the sample

3.11 Wet Method: interferences

The wet chemical digestion techniques rely on the ability of a weak base solution to quantitatively dissolve all amorphous Si components of the sediments, while dissolving only a small fraction of the mineral silicates. Either a "mineral correction" is made for separation between the two components (e.g., DeMaster, 1981) or in sediment samples with high BSi concentrations the relative difference between Si extracted from amorphous compounds and from mineral silicates is large so that the "mineral correction" is ignored (Mortlock & Froelich, 1989; Conley, 1998). However, the difficulty of the BSi measurement lies not only in the correction for the dissolution of co-existing aluminosilicates, but also in the extraction efficiency of $BSiO_2$, which must be 100%. Moreover, if the goal of BSi analysis is to estimate diatoms content in sediment, further components can affect the result. For example, in sediments where sponge spicules as well as diatoms and mineral silicates are present, the diatoms are rapidly dissolved (<2 h), sponge spicules are dissolved during the first 8–12 h of the digestion, and increases in Si extracted after that time period are due to digestion of mineral fraction. The procedure for separating diatom BSi from sponge BSi probably overestimates diatoms BSi because smaller and/or lightly silicified sponges can be completely dissolved early during the digestion process. Sponge BSi can comprise a significant portion of the total amorphous Si extracted from sediments and may act as a significant interfering species (Conley & Schelske, 1993; Bavestrello et al., 1996).

Each of mentioned procedures has inherent systematic problems or is analytically cumbersome. Despite these circumstances, the wet-chemical methods have been used widely by many marine scientists, principally because of being simple and economical in handling. Much controversy still remains surrounding the methodology for BSi determination and there is a need for further study**.**

3.12 Wet Method variability

Besides the problem of interference of clay minerals, the wet method is affected by an inherent variability. Conley (1998) proved that there is a wide range of variability in the measurement of BSi across the community of aquatic scientists. In the study an inter-laboratory a comparison was made with the purpose to show the amplitude of variability

in the measurement of BSi in sediments among the community of aquatic scientists and to determine if patterns in the measurement were related to specific methodologies used or to treatments. 30 selected laboratories used a variety of different wet chemical extraction techniques and X-ray method. Six samples were used in the inter-laboratory comparison collected from modern freshwater and coastal marine depositional environments. The samples were chosen to cover a wide range of BSi concentrations. Independently measured BSi concentrations of the same sediment ranged widely. BSi concentrations determined by X-ray diffraction were significantly higher than concentrations determined by wet chemical methods. It can be noticed (see Tab. 3.1) that the percent standard deviation of the mean in samples analyzed by wet chemical digestion techniques was highest in the samples with the lowest BSi concentration (sample 6) and lowest at the highest BSi concentration (sample 2)

Sample	BSi	SD	%SD of mean
1	2.82	±1.17	±41.6%
2	44.3	±9.38	±21.2%
3	6.49	±2.09	±32.1%
4	38.2	±9.48	±24.8%
5	7.37	±2.56	±34.8%
6	1.31	±0.88	±67.5%

Tab. 3.1 Overall neam biogenica silica (BSi) concentration (wt% as SiO_2) of samples ±1 standard deviation (SD) about the mean and the percent standard deviation of the mean (SD of mean) from all laboratories using wet chemical digestion techniques (Conley 1998)

3.13 Wet Method final thoughts

All wet-chemical methods, in principle, are based on the idea that BSi and aluminosilicates have different dissolution rates even in a weakly alkaline solution such that BSi dissolves faster than aluminosilicates. The dissolution rate of BSi depends on its physical and chemical characteristics such as its origin, age, specific surface area and the concentration of silanol radicals, Si–OH (Kamatani, 1971; Hurd, 1983). Therefore, the recovery of BSi is probably a function of extraction conditions: pH, temperature, nature and concentration of the alkaline solution, and digestion time. Different

conditions are shown in Tab. 3.2. Moreover, the method as described by DeMaster is quite complex and time-consuming: several step are required before the spectroscopic measurement. The major problem, however, is how to make the correction of silica leached from the non-BSi compounds which coexist with biogenic phases in samples.

Solutions used for digestion	Digestion conditions	Sample type	Authors
5% Na_2CO_3	85°C, 5 h	Sediments	Hurd, 1973
5% Na_2CO_3	100°C, 100 min	Sediments	Kamatani, 1980
2 M Na_2CO_3	90–100°C, 4 h	Sediments	Eggimann et al., 1980
1% Na_2CO_3	85°C, 5 h	Sediments	DeMaster, 1981
2 M Na_2CO_3	85°C, 2 h	Sediments	Shemesh et al., 1988
2 M Na_2CO_3	85°C, 5 h	Sediments	Mortlock and Froelich, 1989
1.0 M NaOH	85°C, 30 min	Sediments	Müller and Schneider, 1993
2 M Na_2CO_3	85°C, 5 h	Sediments	Gehlen and van Raaphorst, 1993
0.2 M NaOH	100°C, 40 min	Particulates	Ragueneau and Tréguer, 1994

Tab. 3.2 A list of conditions used wet in alkaline method

Section 4

In the last few decades, many research has been dedicated to polar ecosystems, which are generally regarded as the last uncontaminated environments on Earth. Antarctica is the coldest, windiest and most isolated continent, therefore it is practically unaffected by anthropogenic activity. For this reason, Antarctica is scientifically attractive and is a unique natural laboratory. The study of marine sediments represents an important tool to obtain information about past conditions in the ocean. Sediments are also studied for their role in controlling the biogeochemical cycles of seawater. They constitute repository of biologic trace elements which act as micronutrients and characterize the ecosystem. High-latitude environments experience frequent algal blooms during the spring–early summer retreat of the seasonal sea ice cover. These can generate high pulses of biogenic particulate export from surface waters, especially when algal assemblages are composed of diatoms. The Ross Sea is a deep bay of the Southern Ocean in Antarctica, between Victoria Land and Marie Byrd Land. Although the Ross Sea is covered with ice for most of the year, thanks to the circumpolar deep water current, the water mass that flows onto the continental shelf is relatively warm, salty and nutrient-rich. Therefore, the Ross Sea is one of the last sea areas on Earth that is still relatively unaffected by human influence. Because of this, it is still almost totally free from pollution and the introduction of harmful agents. Accordingly, this area has become a subject of numerous environmentalist groups for its feature of a world marine reserve. The Ross Sea is regarded by marine biologists as having a very high biological diversity: for this reason, it is the target of many scientific research as well focus of some environmentalist groups.

Satellite observations showed that each year the Ross Sea exhibits the most spatially extensive biomass in the Southern Ocean (Comiso et al. 1993, Sullivan et al. 1993, Arrigo et al. 1998). Recently, Smith & Gordon (1997) and Smith et al. (2000) were able to confirm the hyperproductive nature of the Ross Sea through a series of spring process studies in its southern portion. Due to the high preservation potential, the Ross Sea continental shelf is an area of high accumulation of biogenic silica in the sediments (Ledford-Hoffman et al. 1986; DeMaster et al. 1996, Langone et al. 1998). Sediments record environmental conditions at the time of their formation, and the study of sediments can give information on water column processes. Coupling data obtained

from water column fluxes measured by traps and data of sediment deposition and accumulation onto the sea bottom can allow a better understanding of the relative importance of particle sinking processes and the factors that influence particle biogeochemistry and transport.

The Ross Sea has a distinctive geomorphology, characterized by a deep and irregular continental shelf, with an average depth of 500 m. The central portion alternates banks and basins, characterized by an elongate shape and oriented to north-east. The shelf slopes towards the continent and is sharper and deeper on its western side. Near Victoria Land, glacial erosion has created tight transverse channels, that can exceed 1000 m depth.

The samples object of the present study come from a site (named site D, Fig. 4.1), located in the western sector of the Ross Sea continental shelf. Site D is within the polynya (a polynia is an area of open water surrounded by sea ice) of Terra Nova Bay, at 75°06'S and 164°13'E. This is an area of high productivity of biological organism (Saggiomo et al. 2002). Although sediment texture and composition in this area have been already described (Dunbar et al. 1985), biogeochemical processes at the seafloor are poorly known. Surface sediments in the Ross Sea are composed of unsorted ice-rafted debris, siliceous and calcareous biogenic debris and terrigenous silts and clays (Dunbar et al. 1985). In site D predominate coarse terrigenous deposits. Sediment gravity cores and box-cores (gravity core 148c and box core 148bc) were collected at the Site D during the 1994/95 Italian Antarctic expedition.

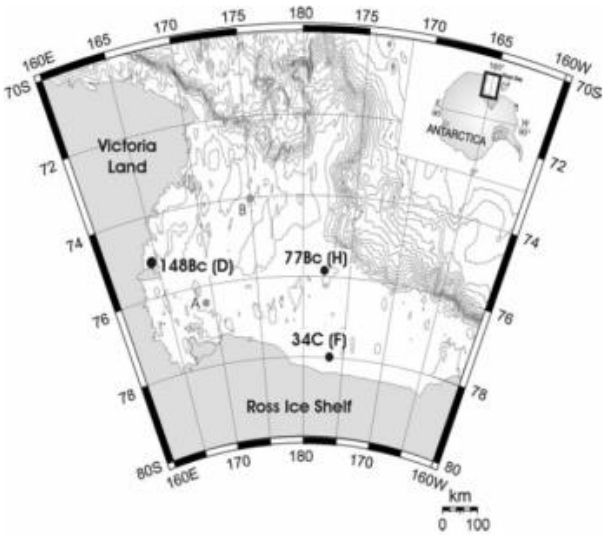


Fig. 4.1 Study area and samples site D

In this section the application of NAS method (based on ATR measurement) is described to determine biogenic silica in marine sediment originated from site D. In order to have some reference values, the samples were formerly analyzed with wet-method in cooperation with Institute of Marine Science (ISMAR-CNR)

4.1 Wet Method performed in ISMAR

Biogenic silica (BSi) content was determined through a progressive dissolution method, followed by colorimetric analysis. Sediments were previously dried at 60°C for 24h at least, and then pulverized in an agate mortar. 0.5g of sediment was transferred in a teflon falcon tube and 35ml of 0.5M NaOH solution as an extractant were added. Tubes were immersed in a water bath heated to 85°C and periodically (every 15 min) shaken throughout the digestion period to ensure full exposure of the sediment to the solution. For each core, 3 sample replicates were analysed to check method reproducibility, and a set of blanks was run every ten samples. After 1, 2, 3 and 4 hours of digestion, tubes were centrifuged at 2000 rpm for 3 min and then 0.2 ml of sample aliquots were collected. Such aliquots were diluted in MilliQ water to a volume of 5 ml (Solution A).

Separately a solution containing ammonium paramolybdate, NaOH 0.5M and water in ratios1:0.2:2.3 is prepared (Solution B). For each sample, 1 ml of solution A and 2.8 ml of solution B are mixed in order to obtain a volume of 3.8 ml (Solution C). The reducing agent is obtained by mixing metol, 4-(methylamino)phenol hemisulfate, saturated solution of oxalic acid, sulphuric acid 50% and water in ratios 10:6:6:8.

1.2 ml of reducing solution is added to sol. C, at this point total volume for each sample will be 5ml. After 3 h, spectrophotometric measurement are carried out at □=810 nm.

Concentrations are determined by interpolation using a calibration line built with a reference solution of BSi.

4.2 ATR measurement

The analytical reproducibility and accuracy of the FTIR-ATR method are strictly related to the optimization of the experimental conditions such as drying process, sample

deposition, instrumental calibration. In ATR measurement two requirement are particularly important to obtain an adequate spectrum

1. The size of particles mustn't be greater than the wavelength at which absorption is measured
2. The sample must be homogeneous

The results also depend strictly on the effect of the spectral resolution on the intensity of the signal. The improvement in resolution (when the resolution decreases at 1 or 2 cm^{-1}) produces a general improvement of intensity measurements. In any case, the intensity of the signal is also influenced by the change in the refractive index, depending on the concentration of the absorbing species. When the effect of the change in the refractive index is negligible, a linear relationship is observed between the intensity of the signal and the concentration, according to the Lambert–Beer law. A spectral resolution of 4 cm^{-1} is an acceptable trade-off , otherwise the relationship can result nonlinear.

Samples were first mechanically grinded down to a size smaller than the shortest wavelength used (2.5 mm) before mixing with Celite (Sigma) in various amounts. The powder was dried at 110 °C overnight in order to remove absorbed water and then analyzed with an FTIR spectrometer (65 scans) in the 1400–400 cm^{-1} range. FTIR-ATR spectra were obtained from approximately 0.01 g of sample material using an instrument Bruker Model ALPHA (Bruker Optik GmbH, Leipzig, Germany).

The FTIR spectra of BSi (Fig. 4.2) exhibit four vibration bands. The two main bands at 1100 and 471 cm^{-1} are attributed respectively to triply degenerated, stretching and bending, vibration modes of the [SiO_4] tetrahedron. The band at 800 cm^{-1} corresponds to an inter-tetrahedral Si–O–Si bending vibration mode and the band near 945 cm^{-1} to a Si–OH mode .

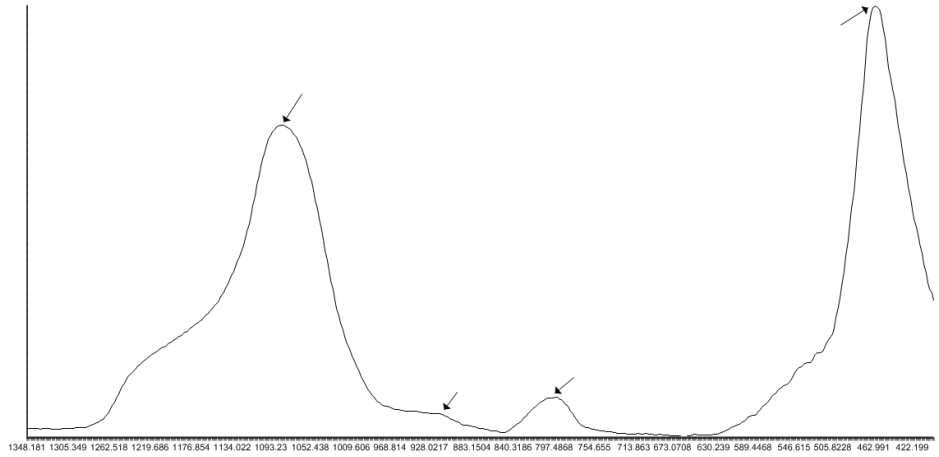


Fig. 4.2 Biosilca ATR spectrum

After having collected spectra, all data were subjected to the following procedure

1) Spectral pre-treatment: MSC and normalization
2) Esplorative analysis: PCA
3) Mathematical treatment: PLS /NAS

5.3 Data pre-processing.

Pre-processing of the spectral data was performed by The Unscrambler (Version 10.3 Camo, Norway) software packages. For optimal modelling, Multiple Scatter Correction (MSC) and baseline correction were used to linearize spectra and remove variation in spectra caused by noise. Baseline correction performs a linear correction of the spectra so that two points equal zero. MSC removes spectral variation arising from different effective path lengths and particle sizes.

4.3 Determination of BSi by NAS

The quantitative assesment of BSi is performed by Multivariate Standard Addition Method (MSAM). Partial Least Squares (PLS) Regression (par. 1.10), was used for building the calibration model. By PLS the explained variance and the correlation are optomized at the same time (Geladi 1986). According to standard addition method, known amounts of dependent variable (celite in this case) are added to the sample. Then the unkwnown value is determined in extrapolation mode, which allows to bypass the eventual matrix effect. The suitability of multivariate models to be applied in

extrapolation mode has been demonstrated by some Authors (Lorber et al. 1997). In that work, MSAM was applied to a bulk method, while in the present work the sample was in solid state: this is another very useful novelty, since analysing solid samples as such is a very interesting target for analysts.

One futher novelty here proposed with respect to the previuos work on MSAM is how to extrapolate the signal. In fact, in that previous work a blank solution was available, and the blank signal was the unkwnown to be predicted by the MSAM model. On the contrary, in the present case the blank is not available. We don't have a marine sediment surely devoid of biogenic silica. The solution to this issue comes from chemometrics again: the Net Analyte Signal (NAS) method may solve the problem. This method was introduced by Lorber (1997), and Hemmateenejad (2009) applied it to MSAM in similar mathematical mode as it is applied in the present work. The present procedure was followed

- A first interference reduction is obtained by computing a PLS model from the original data and choosing the proper number of PCs (the ones which minimizes RMSE). By a matrix product between the chosen scores and loadings we rebuild the original data to a R_{reb} matrix
- A mathematical combination of the pure analyte signal (which is represented by the vector **r**) with the matrix R_{reb} extract the blank contribution to signal (matrix R_{-k} so called because contains all compounds signals except the kth analyte signal)

$$R_{-k} = R_{reb} - \alpha \hat{c}_k r^T$$

Where α and $\hat{c}_k$ are calculated as described in par. 1.16

- For each object (row) in the dataset, a "net" signal is calculated

$r^*_k=[I- R_{-k} R^+_{-k}] r_k$

r^*_k is the vector of the kth analyte. This mathematical operation performed to extract the blank contribution consist in projecting the pure-anayte signal on a plane and the sample-signal on another plane that is perpendicular to the first: normal planes correspond to independent mathematical quantities. In other words, NAS is based on the idea of extracting that part of the signal which is mainly related to the concentration of the analyte of interest and independent from all other components

The result is the NAS vector r^*: by taking the Euclidean norm we obtain a scalar signal directly related to the analyte concentration. Also the added concentration vector is projected onto the PLS-NAS space to give a new concentration vector (here called $\hat{c}_k$). In this way, a multivariate model is well represented by an univariate linear regression between the pseudo-univariate NAS signal and $\hat{c}_k$. This method is well developed for the standard addition method. The extrapolated value doesn't need further calculation and it corresponds to the extrapolated concentration value of the starting data. If necessary, the described PLS and NAS procedures can be combined with chemometric tools for variables selection. In fact, the dataset here analysed is characterized by a low number of rows (some units of samples which, by replicating measurements, become some tens of objetcs, the spectra) and a very high number of variables (some hundreds of frequencies). The LASSO and SPLS variable-selection (decribed in shrinkage techniques) were applied, and among them the best performing was chosen.

4.4 Results

All the samples were both analyzed as such and emplyed to prepare several samples for the standard addition method.

Sample D9

In the case of sample D9, four additions of celite were performed: 2.2, 5.2, 7.9 and 12.5 $\%_{w/w}$.

In figg. 4.3-4.4 the original acquired spectra (4.3) and the spectra after MSC treatment (4.4) are shown. As for treated spectra (likewise for the following samples), also the pure-analyte spectrum is plotted (marked 100%). In the score plot of PCA (fig 4.9) it can be seen that samples with the same content of celite form clusters which lie on horizontal axis: cluster with increasing amount of analyte are aligned from right to left.

For this sample the LASSO resulted the most suitable shrinkage technique (fig.4.12) and after PLS three components were chosen. In fig. 4.14 NAS regression line is shown, together with the fundamental figures of merit (quality parameters). The extrapolated value is (3.79±0.09) $\%_{w/w}$, whreas the wet-method result is (4±1) $\%_{w/w}$.

Sample D18

Sample D18 resulted more difficult to analyze compared to the others. In fact, even though five addition on the analyte were performed (1.8,3.2,7.2,10.3 and 14 $\%_{w/w}$), samples with 1.8% $_{w/w}$ and 10.3% $_{w/w}$ content of celite were rejected owing to a big overlapping of spectra. In PCA (fig. 4.10) score plot, samples with increasing amount of analyte are ordered from left to right: we can see that sample 1.8% $_{w/w}$ is not distinguishable from sample as such (0% $_{w/w}$ of added celite) and the same is for sample 10.3% $_{w/w}$ compared to samples 7.2% $_{w/w}$ and 14% $_{w/w}$. Two further outliers are marked with a circle. In fig. 4.6. just spectra of samples used in calibration (0, 3.2, 7.2 and 14 $\%_{w/w}$ of added celite) are shown.

No shrinkage methods were applied because no improvement was observed. Despite these problems, the extrapolated value (5.2±0.3) $\%_{w/w}$ is somewhat close to the one obtained with wet method (4.3±0.6) $\%_{w/w}$. NAS line and relative parameters are reported in fig. 4.15.

Sample D21

For this sample, just three addictions of celite (5,10 and 14 $\%_{w/w}$) were enough to build a good calibration model. In score plot (fig 4.11.) it can be seen that clusters with increasing amount of celite are well divided and ordered along horizontal axis from left to right.

After MSC pre-treatment, there is a little overlapping in some region of the spectrum (fig.4.8). However, the extrapolated value (fig. 4.16) is in excellent agreement with "classical" method (3.17±0.04) $\%_{w/w}$ for ATR and (3.5±0.5) $\%_{w/w}$ for the wet method).

In the case of sample D21, SPLS method (fig. 4.13) instead of LASSO was emplyed, and three components were used.

Table 5.1 summes up the results with reference values obtained with wet method and NAS regression lines.

Sample	Shrinkage	Number of components	NAS	Wet Method
D9	LASSO	3	3.79±0.09	4±1
D18	-	3	5.2±0.3	4.3±0.6
D21	SPLS	6	3.17±0.04	3.5±0.5

Tab. 5.1

Conclusions

In this work, a new chemometric method applied to the analysis of environmental samples is presented. The aim of the project was to develop new multivariate procedures to be applied to data obtained by direct analytical techniques. The great advantage of direct chemical analysis is that we can to investigate samples without altering them, keeping the sample available for further analysis. Specifically, infrared spectroscopic measurements in Attenuated Total Reflectance (ATR-IR) were performed. In addition to being a non destructive technique, ATR is rapid and useful to characterize materials with minimal sample preparation. Compared with transmission spectroscopy, ATR sample preparation is less labor-intensive, spectra variation due to sample preparation is minimal and the impact on results of sample preparation due to KBr grinding and particle size differences is greatly reduced.

Although ATR is a widespread technique in qualitative investigations, its use in quantitative analysis is not yet well-established. Moreover in the case of environmental samples, due to their very complex matrices, matrix effect hinder the use of calibration methods in interpolation mode. For this reason univariate approach may be not exaustive and Multivariate Standard Addition Method (MSAM) was chosen as suitable alternative.

In multivariate analysis, and in particular in creating and validating calibration models, real data are often contaminated with atypical samples or instrumental anomalous responses (outliers), which may have harmful consequences on the estimations of the parameters. A big issue in this work was that spectroscopic data are often affected by undesired systematic variations primarily caused by phenomena like difference path-length and light-scattering. In the case of ATR, where the penetration depth of incoming beam is proportional to wavelength and to incident angle, such phenomena produce changes in intensities of some bands. Moreover, light scattering due to physical conditions of the samples (*e.g.* surface roughness, droplets, crystalline defects, density fluctuations) can cause shifts in baseline and other phenomena called non-linearities. For these reasons, pre-processing techniques were required in order to remove physical phenomena and reduce the un-modeled variability in the data, thus enhancing the feature sought in the spectra. Such preprocessing methods are also aimed to adjust baseline shifts among samples, thus improving subsequent multivariate analysis. The goal was to obtain a simple (linear) relationship with the constituent of interest. Among

scatter-correction methods, we considered three preprocessing methods: MSC, normalization and derivative. Therefore, as first step after we collected data, we experienced these techniques to reduce the physical variability between samples due to scattering.

Besides the above mentioned experimental shortcomings, the heart of this work consists in mathematical treatment of data. As matter of fact, in certain circumstances calibration methods cannot be performed in classical mode, *i.e.* with Ordinary Least Square regression. When high correlated variables are present and/or the number of independent variables >> number of experiment, Principal Component Regression or Partial Least Square Regression are required. Such methods are based on so-called inverse regression and provide models in which the signal (IR spectrum in this case) is the *independent variable*. Hence, in order to extrapolate the unknown, blank signal is needed. In environmental analysis the blank is typically not available: we don't have the matrix without analyte. The theory of Net Analyte Signal (NAS) originally proposed by Lorber (1986) and subsequently applied and improved by many authors allows to separate, by a mathematical procedure, the signal of analyte of interest from other interfering components. Implementing NAS principle, for the first time MSAM was applied to solid state samples: this is a very useful novelty, since analysing solid samples as such is a very interesting target for analysts.

Spectroscopic measurement and following multivariate treatment were applied to determine biogenic silica (BSi) content in marine sediment. In order to confirm the obtained results, samples were also analysed in cooperation with ISMAR-CNR (Institute of Marine Science), where the alkaline methods were applied according the protocol proposed by DeMaster(1981). A quite good agreement was observed.

Perspectives

Since the results obtained in this work are very encouraging, in perspective the same work may be applied to all solid-state direct-analysis analytical techniques.

In fact, preliminary work applied to X-Ray Diffractometry (not here reported) has already been done with very good results.

Moreover, this work, together with previous similar work in Raman Spectroscopy, has already been published (Melucci et al., 2016).

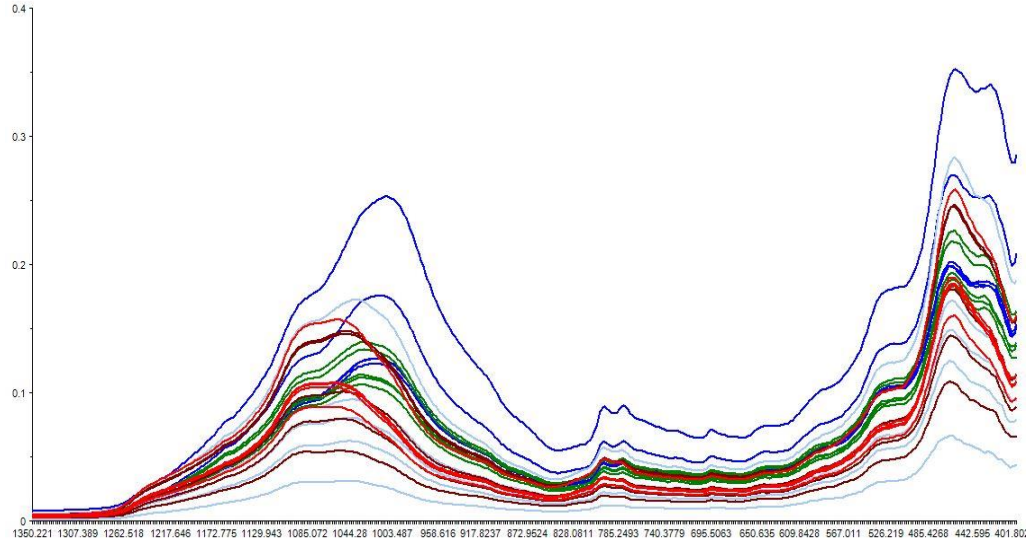


Fig 4.3 D9 original spectra

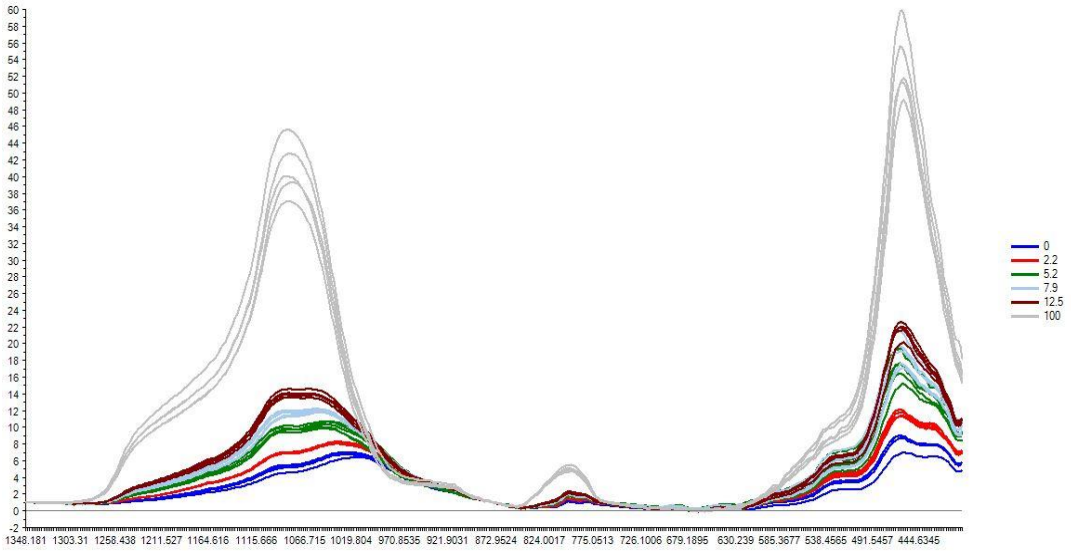


Fig. 4.4 D9 treated spectra

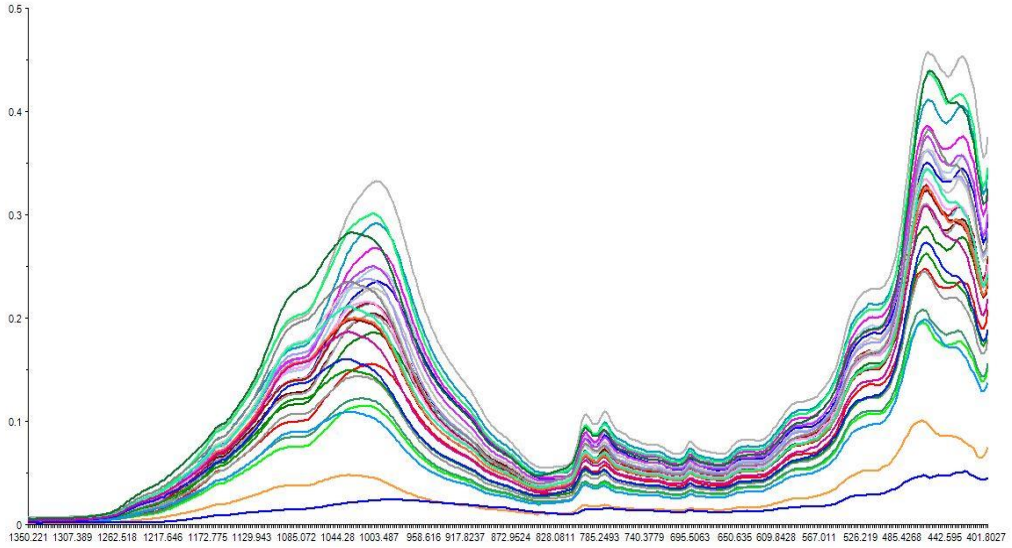


Fig. 4.5 D18 original spectra

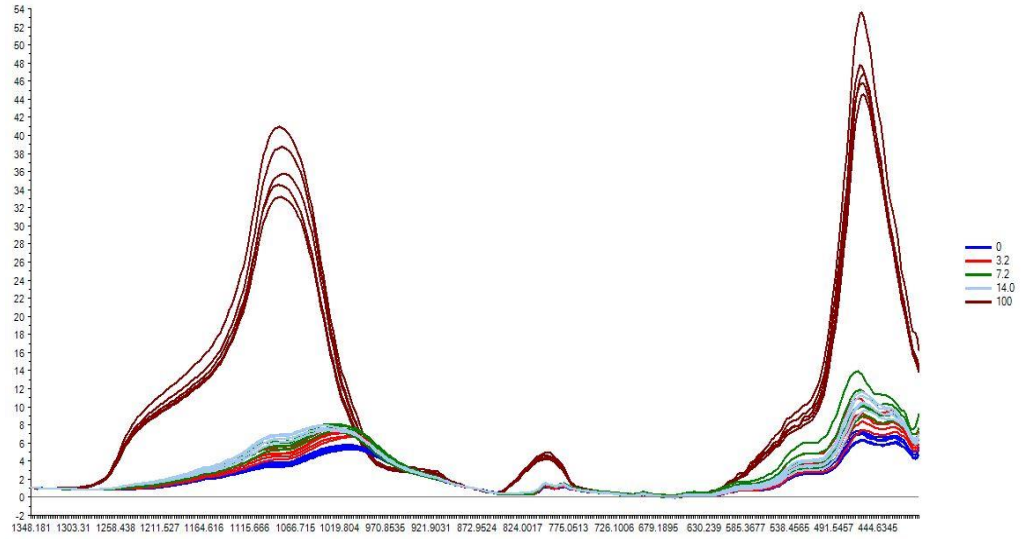


Fig. 4.6 D18 treated spectra

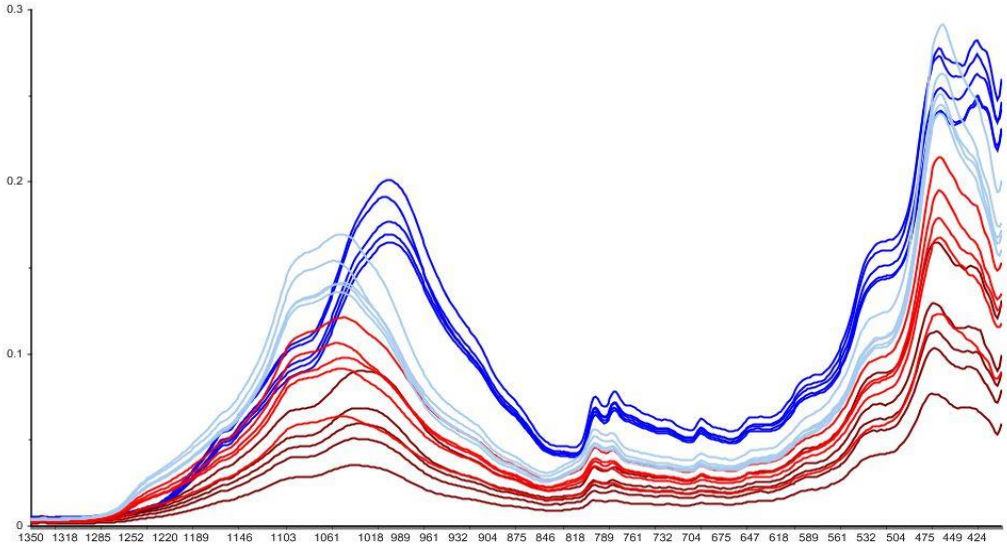


Fig. 4.7 D21 original spectra

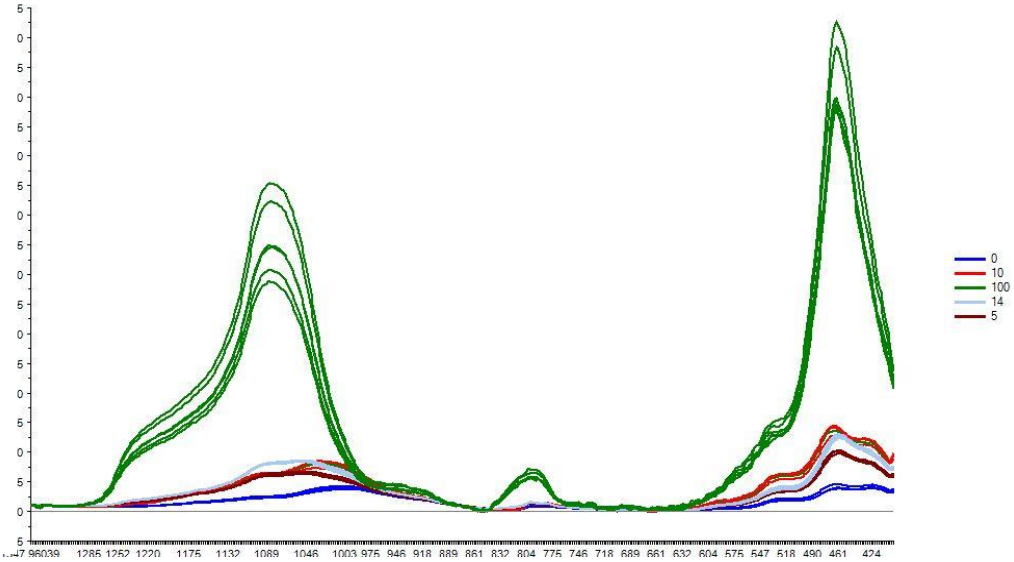


Fig. 4.8 D21 treated spectra

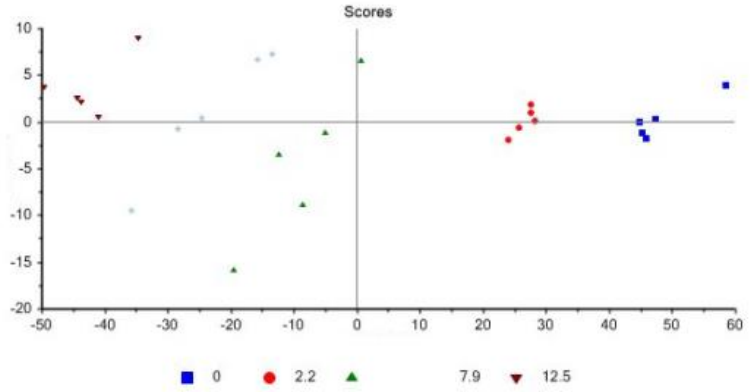


Fig. 4.9 D9 PCA

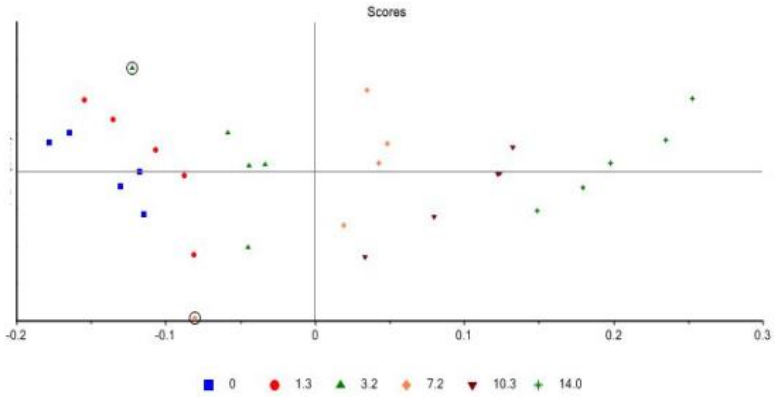


Fig. 4.10 D18 PCA

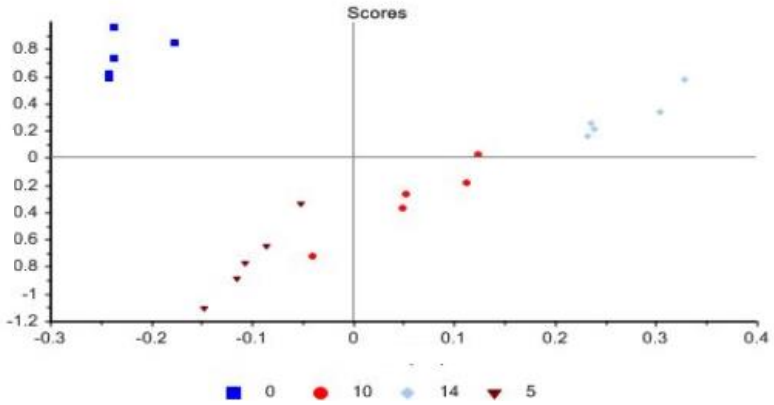


Fig. 4.11 D21 PCA

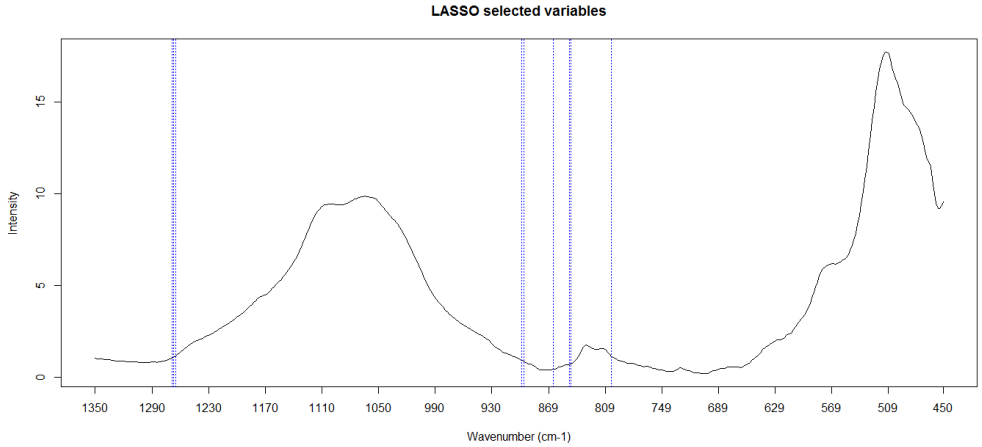


Fig 4.12 D9 LASSO

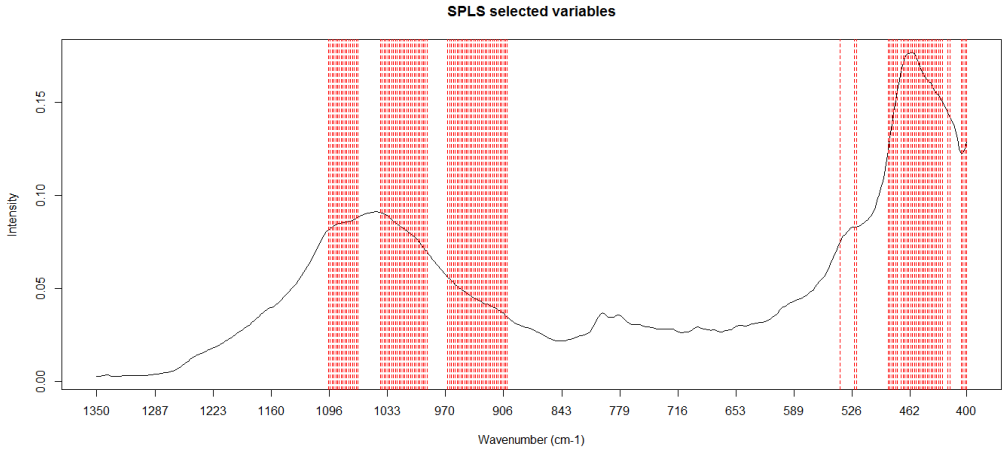
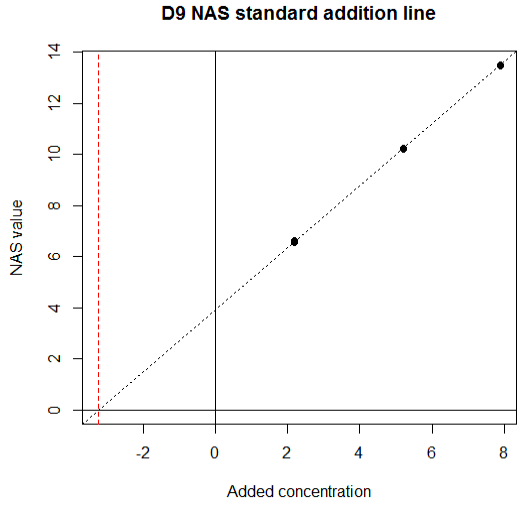
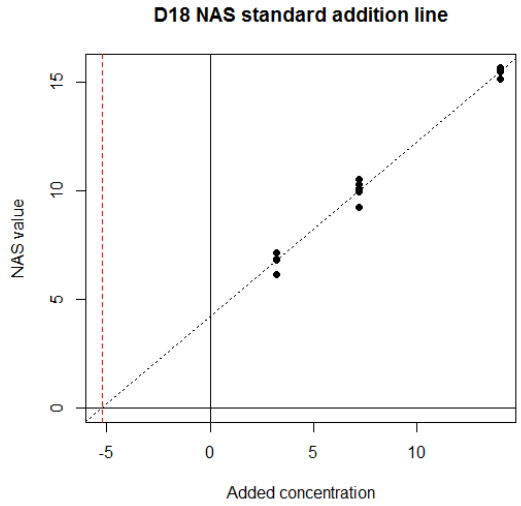


Fig. 4.13 D21 SPLS



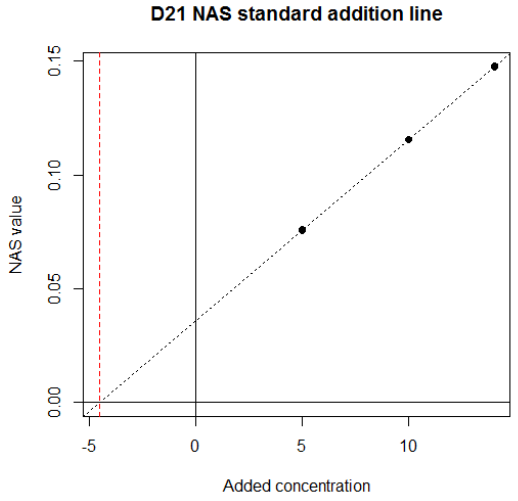
Intercept = 3.917±0.008
Slope =1.212±0.0.002
RMSE=0.01401
Radj=0.999

Fig 4.14 D9 NAS



Intercept = 4.2±0.2
Slope =0.80±0.02
RMSE=0.3591
Radj=0.991

Fig 4.15 D18 NAS



Intercept = 3.63e-02±2e-04
Slope =7.95e-03±1e-05
RMSE=0.002
Radj=0.999

Fig. 4.16 D21 NAS

Bibliography

Arrigo, K.R., Weiss, A.M. & Smith JR, W.O.,1998, Physical forcing of phytoplankton dynamics in the southwestern Ross Sea. Journal of Geophysical Research, **103**, 1007–1021.

Bavestrello, G., R. Cattaneo-Vietti, C. Cerrano, S. Cerutti & M. Sara, 1996. Contribution of sponge spicules to the composition of biogenic silica in the Ligurian Sea. P.S.Z.N. I. Mar. Ecol. 17: 41–50.

Bezrukov, P.L., 1955, Distribution and rate of deposition of silicate sedimentations in the Sea of Okhotsk. Dokl. Akad. Nauk SSSR 103, 473–476.

Booksh, K. S.; Kowalski, B. R., 1994**.** Anal. Chem., *66*, A782-A791.

Brzezinski MA, Jones JL, Bidle K.D., Azam F. ,2003, The balance between silica production and silica dissolution in the sea: insights from Monterey Bay, California, applied to the global data set. Limnol. Oceanogr.48:1846–54

Comiso, J.C., McClain C.R., Sullivan C.W, Ryan., J.P. & Leonard,C.L., 1993.Coastal Zone color scanner pigment concentrations in the Southern Ocean and relationships to geophysical surface features. Journal of Geophysical Research, 98, 2419–2451.

Conley, D. J. and Schelske C.L., 1993. Potential role of sponge spicules in influencing the silicon biogeochemistry of Florida lakes. Can. J. Fish. Aquat. Sci. 50: 296–302.

Conley, D. J., 1998, An interlaboratory comparison for the measurement of biogenic silica in sediments. Mar. Chem. 63: 39– 48

De Master, D.J. 1981. The supply and accumulation of silica in the marine environment Geochim. Cosmochim. Acta 45:1715–1732

DeMaster, D.J., 1991. Measuring biogenic silica in marine sediments and suspended matter. In: Hurd, D.C., Spenser, R.W. _Eds.., Marine Particulates: Analysis and characterization. American Geophysical Union, pp. 363–368

DeMaster, D.J., Ragueneau, O. & Nittrouer, C.A. ,1996. Preservation efficiencies and accumulation rates for biogenic silica and organic C, N and P in high-latitude sediments: the Ross Sea. Journal of Geophysical

Dunbar, R.B., Anderson, J.B., Domack, E.W. & Jacobs, S.S.,1985. Oceanographic influences on sedimentation along the Antarctic continental shelf. Antarctic Research Series, 43, 291–312.

Eggiman, D.W., Manheim, F.T., Betzer, P.R., 1980. Dissolution and analysis of amorphous silica in marine sediments. J. Sed. Petrol. 50, 215–225
Faber, N. M. ,1998**,** *Anal. Chem, 70,* 5108-5110.

Fahrentfort J,1959. Attenuated total reflection- A new principles for the production of useful infrared reflection spectra of organic compounds. Molecular Spectroscopy(Proceeding IV Int. Meeting, Bologna), 2, Mangini A, editor, Londan:Pergamon, 1962, 437.

Frohlich, F. ,1989. Deep-sea biogenic silica: new structural and analytical data from infrared analysis—geological implications, Terra Res. 1:267–273.

Gehlen, M., van Raaphorst, W., 1993. Early diagenesis of silica in sandy North Sea sediments: quantification of the solid phase. Mar. Chem. 42, 71–83.

Geladi, P., Kowalski, B.R.,1986. Partial Least-Squares Regression: a Tutorial, Analytica Chimica Acta 185:1-17.

Geladi P.. MacDougal D., Martens H., ,1985,Appl. Spectrosc. 39 491-500.

GoicoecheaHC,OlivieriAC.,1999.Enhanced synchronous spectrofluorometric determination of tetracycline in blood serum by chemometric analysis. Comparison of partial least-squares and hybrid linear analysis calibrations. Anal. Chem.; 71: 4361–4368.

Goicoechea HC, Olivieri AC. ,2001. A comparison of orthogonal signal correction and net analyte preprocessing methods. Theoretical and experimental study. Chemometrics Intell. Lab. Syst.; 56: 73–81.

Goldberg, E.D. ,1958, Determination of opal in marine sediments, J. Mar. Res. 17:178–182.

Harrick NJ,1967, Internal Reflection Spectroscopy. New York, John Wiley & Sons, Int.,

Hastie T., Tibshirani R. and Friedman J., 2009.. The Elements of Statistical Learning,:Data Mining, Inference, and Prediction, Springer New York, 2edition.

Hemmateenejad, B., Yousefinejad, S., 2009. Multivariate standard addition method solved by net analyte signal calculation and rank annihilation factor analysis, Anal Bioanal Chem 394:1965–1975.

Hoerl A. E. and Kennard R. W. ,1970. Ridge Regression: Biased Estimation for Nonorthogonal Problems. Technometrics, 12:55–67

Hotelling H., 1933, Analysis of a complex of statistical variables into Principal Components, Journal of Educational Psychology, 24 : 417-441 and 498-520.

Hurd, D.C. 1972. "Factors affecting solution rate of biogenic opal in seawater" Earth Planet. Sci. Lett. 15:411–417.

Hurd, D.C., 1973. Interaction of biogenic opal, sediment and seawater in the Central Equatorial Pacific. Geochim. Cosmochim.Acta 37, 2257–2282

Hurd, DC, Birdwhistell S. (1983). On producing a more general model for biogenic silica dissolution. *Am. J.Sci.* 283:1–28

Kamatani, A. and O. Oku, 2000, Measuring biogenic silica in marine sediments. Mar. Chem. 68: 219–229.

Kamatani, A., 1971, Physical and chemical characteristics of biogenous silica. Mar. Biol. 8, 89–95.

Langone, L., FrignanI, M., Labbrozzi, L. & Ravaioli, M., 1998, Present day biosiliceous sedimentation in the NW Ross Sea (Antartica*).* Journal of Marine Systems, 17, 459–470.

Leconte J., Le Rayonnement Infrarouge, 1949, Gauthier-Villard, Paris,. 5.2

Ledford-Hoffman P.A., DeMaster D.J. & Nittrouer C.A. ,1986. Biogenic silica accumulation in the Ross Sea and the importance of Antarctic continental-shelf deposits in the marine silica budget. Geochimica Cosmochimica Acta, 50, 2099–2110.

Leinen, M. ,1977. A normative calculation technique for determining opal in deep sea sediments, Geochim. Cosmochim. Acta 40:671–676.

Lorber A. ,1986..Error propagation and figures of merit for quantification by solving matrix equations. Anal. Chem.; 58: 1167–1172.

Lorber A, Faber K., Kowalski BR.,1997. Net analyte signal calculation in multivariate calibration. Anal. Chem.1997; 69: 1620–1626

Martens H, Jensen S.A. , Geladi P., ,1983, Multivariate linearity transformations for near infrared reflectance spectroscopy, in: O.H.J. 2.5Christie (Editor), Proc. Nordic Symp. Applied Statistics, Stokkland Forlag, Stavanger, Norway, , pp. 205–234

Melucci D., Cocchi M., Corvucci F., Boi M., Tositti L., de Laurentiis F., Zappi A., Locatelli C., Locatelli M., 2017, Chemometrics for the direct analysis of solid samples by spectroscopic and chromatographic techniques In: "Chemometrics:

methods, applications and new research", Chapter 9, pp. 173-204, Nova Science Publishers, Inc.,

Meyer-Jacob, C., Vogel, H., Boxberg, F., Rosen, P., Weber, M.E., Bindler, R., 2014. Independent measurement of biogenic silica in sediments by FTIR spectroscopy and PLS regression, J. Paleolimnol 52:245–255.

Moenke H.H.W, in: E.V.C. Farmer (Ed.), 1974, The Infrared Spectra of Minerals, Mineralogical Society Monograph, London, , p. 365. 5.2

Mortlock, R.A., Froelich, P.N., 1989. A simple method for the rapid determination of biogenic opal in pelagic marine sediments. Deep-Sea Res. 36, 1415–1426.

Muller, P.J., Schneider, R., 1993, An automated leaching method for the determination of opal in sediments and particulate matter. Deep-Sea Res. 40, 425–444.

Nelson DM, Treguer P, Brzezinski MA, Leynaert A, Queguiner B. ,1995. Production and dissolution of biogenic silica in the ocean: revised global estimates, comparison with regional data and relationship with biogenic sedimentation. Glob. Biogeochem. Cycles 9:359–72

P. Geladi, D. MacDougal, H. Martens, ,1985,Appl. Spectrosc. 39 491 2.5.

Pearson K.,1901. On lines and planes of closest fit to systems of points in space, Philosophical Magazine, 2(6) 559-572.

Pudsey, C.J. ,1992, Calibration of a point-counting technique for estimation of biogenic silica in marine sediments, Journal of Sedimentary Petrology 63:760-762.

Ragueneau, O., Treguer, P., 1994. Determination of biogenic silica in coastal waters: applicability and limits of the alkaline digestion method. Mar. Chem. 45, 43–51

S. Parke, in: E.V.C. Farmer (Ed.), 1974, The Infrared Spectra of Minerals, Mineralogical Society Monograph, London, , p. 483. 5.2

Saggiomo, V., Catalano, G., Mangoni, O., Budillon, G. & Carrada, G.C.,2002. Primary production processes in ice-free waters of the Ross Sea (Antarctica) during the austral summer 1996. Deep-Sea Research II, 49, 1787–1801.

Sanchez, E.; Kowalski, B. R. ,1988. J. Chemom., *2*, 247-264.

Smith JR, W.O. & Gordon, L.I. 1997. Hyperproductivity of the Ross Sea(Antarctica) polynya during austral springs. Geophysical ResearchLetters, 24, 233-236

Smith JR, W.O., Marra, J., Hiscock, M.R. & Barber, R.T. 2000. The

seasonal cycle of phytoplankton biomass and primary productivity in the Ross Sea, Antarctica. Deep-Sea Research II, 47, 3119–3140.

Sullivan, C.W., Arrigo K.R., McClain, C.R., Comiso, J.C. & Firestone, J., 1993.Distributions of phytoplankton blooms in the Southern Ocean. Science, **262**, 1832–1837.

Tibshirani R. 1996. Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, 58(1):267–288.

Treguer P, Nelson DM, van Bennekom AJ, DeMaster DJ, Leynaert A, Queguiner B,1995, The balance of silica in the world ocean: a re-estimate. Science 268:375–379

Van Cappellen P., Dixit S., and Van Beusekom J. E. E. ,2002. Biogenic silica dissolution in the oceans: reconciling experimental and field-based dissolution rate. Global Biogeochem. Cycles 16 (25), 1 – 10.

Xu L., Schechter I.,1997,A calibration method free of optimum factor number selection for automated multivariate analysis. Experimental and theoretical study. Anal.Chem.; 69: 3722–373

Appendix Library and script R

```
library(MASS)
library(pls)
library(glmnet)
library(spls)

# R is the matrix containing samples spectra
# c.agg is the vector of added concentrations
# r.pure is the matrix containing the pure analyte spectra

n.fold<-10

## LASSO ##
grid<-10^seq(10,-10,length=300)
set.seed(1)
cv.model.lasso<-cv.glmnet(R,c.agg,alpha=1,grouped=FALSE,nfolds=n.fold)
plot(cv.model.lasso)

lambda<-cv.model.lasso$lambda.min
out.lasso<-glmnet(R,c.agg,alpha=1,lambda=lambda)
coeff.lasso<-predict(out.lasso,type="coefficients",s=lambda2)

plot(cv.model.lasso$glmnet.fit,xvar='lambda')
abline(v=log(c(lambda)),lty=2)

index.lasso<-which(coeff.lasso!=0)
coefficients.lasso<-coeff.lasso[index.lasso]

## SPLS ##
set.seed(1)
cv<-cv.spls(R,c.agg,eta=seq(0.1,0.9,0.01),K=c(1:10),fold=n.fold)

model.spls<-spls(R,c.agg,K=cv$K.opt,eta=cv$eta.opt,scale.x=FALSE)
coef.spls<-coef(model.spls)

index.spls<-which(coef.spls!=0)
coefficients.spls<-coef.spls[index.spls]


## Dataset definition ##
# If a subset of variables has been chosen by LASSO or SPLS
index<-index.lasso # For LASSO-subset
index<-index.spls # For SPLS-subset
R.new<-subset(R,select=index)
r.pure.mean<-apply(subset(r.pure,select=index),2,mean)
```

```
# ELSE if NO subset has to be used
R.new<-R
r.pure.mean<-apply(r.pure,2,mean)

## PLS Regression ##
n.sam<-nrow(R.new)
p.var<-ncol(R.new)
set.seed(1)
model.plsr<-plsr(c.agg~R.new,validation="CV",ncomp=n.sam-3)
rmse<-RMSEP(model.plsr)
plot(rmse,legendpos="topright")
n.comp<-# Optimal number of components (which minimizes RMSE)

## Recalculation of R ##
T<-as.matrix(model.plsr$scores[,1:n.comp])
P<-as.matrix(model.plsr$loadings[,1:n.comp])
m<-apply(R,2,mean)
R.reb<-t(t(T%*%t(P))+m)

## R.reb decomposition ##
R.reb.pinv<-ginv(R.reb)
c2<-R.reb%*%R.reb.pinv%*%c.agg
w<-r.pure.mean%*%R.reb.pinv
alpha<-as.numeric(1/(w%*%c2))

## Rank Annihilation ##
Rk<-R.reb-alfa*c2%*%r.pure.mean
Rk.pinv<-ginv(Rk)

## Orthogonal projection matrices ##
Id<-diag(1,p.var,p.var)
H<-(Id-t(Rk)%*%t(Rk.pinv))

## Net spectra and NAS ##
r.net<-matrix(NA,n.sam,p.var)
nas<-matrix(NA,n.sam,1)
for(i in 1:n.sam) {
r.net[i,]<-H%*%R.new[i,]
nas[i,]<-norm(as.matrix(r.net.i[i,]),"f")
}

## Univariate regression ##
lin.reg<-lm(nas~c.agg)
summary(lin.reg)

## Extrapolation ##
```

```
c0<-lin.reg$coefficients[1]/lin.reg$coefficients[2]
c0
```