Developments in Clay Science

Volume 8

# Infrared and Raman Spectroscopies of Clay Minerals

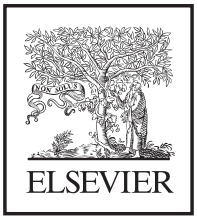
Edited by

**W.P. Gates**
Institute for Frontier Materials, Deakin University, Victoria, Australia

**J.T. Kloprogge**
Department of Chemistry, College of Arts and Sciences University of the Philippines, Visayas, Miagao, Iloilo, Philippines

**J. Madejová**
Slovak Academy of Sciences, Institute of Inorganic Chemistry, Bratislava, Slovakia

**F. Bergaya**
CNRS, Interfaces, Confinement, Matériaux et Nanostructures (ICMN) Orléans, France



Elsevier
Radarweg 29, PO Box 211, 1000 AE Amsterdam, Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, United Kingdom
50 Hampshire Street, 5th Floor, Cambridge, MA 02139, United States



**Notices**

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

ISBN: 978-0-08-100355-8
ISSN: 1572-4352



*Publisher:* Candice Janco
*Acquisition Editor:* Amy Shapiro
*Editorial Project Manager:* Tasha Frank
*Production Project Manager:* Maria Bernard
*Designer:* Victoria Pearson

Typeset by SPi Global, India

# Dedication

**In loving memory of my son Andrew Olav Kloprogge**

**Feb. 2000–May 2016**
**J.T. Kloprogge**

# Contents

# Contributors

**D. Bain**, Middleton Steading, North Deeside Road, Cults, Aberdeen, AB15 9PL, United Kingdom; derekcbain@gmail.com

**E. Balan**, Sorbonne Universités, Paris, France; etienne.balan@impmc.jussieu.fr

**F. Bergaya**, CNRS, Interfaces, Confinement, Matériaux et Nanostructures (ICMN) Orléans, France; f.bergaya@cnrs-orleans.fr

**J.L. Bishop**, SETI Institute, Mountain View, CA, United States; jbishop@seti.org

**J. Carter**, Institut d'Astrophysique Spatiale, CNRS, University de Paris - Sud, Orsay, France; john.carter@ias.u-psud.fr

**G.D. Chryssikos**, National Hellenic Research Foundation, Athens, Greece; gdchryss@eie.gr

**W.P. Gates**, Institute for Frontier Materials, Deakin University, Melbourn, Victoria, Australia; will.gates@deakin.edu.au

**H.P. He**, Guangzhou Institute of Geochemistry, Chinese Academy of Sciences, Guangzhou, China; hehp@gig.ac.cn

**C.T. Johnston**, Purdue University, West Lafayette, IN, United States; clays@purdue.edu

**J.T. Kloprogge**, Department of Chemistry, College of Arts and Sciences University of the Philippines, Visayas, Miagao, Iloilo, Philippines; t.kloprogge@qut.edu.au

**J. Madejová**, Institute of Inorganic Chemistry, Slovak Academy of Sciences, Bratislava, Slovakia; jana.madejova@savba.sk

**J.R. Michalski**, Department of Earth Sciences and Laboratory for Space Research, University of Hong Kong, Hong Kong, China; jmichal@hku.hk

**S. Petit**, Institut de Chimie des Milieux et Matériaux de Poitiers, Université de Poitiers, Poitiers, France; sabine.petit@univ-poitiers.fr

**H. Pálková**, Institute of Inorganic Chemistry, Slovak Academy of Sciences, Bratislava, Slovakia; uachpalk@savba.sk

**J. Zhu**, Guangzhou Institute of Geochemistry, Chinese Academy of Sciences, Guangzhou, China; zhujx@gig.ac.cn

# Acknowledgements

**(1)** *Faïza Bergaya* acknowledges her co-editors for accepting this non-trivial task and for the heavy work they provided during these past 2 years. The author is particularly thankful to Jana Madejová for her enthusiastic participation as co-author of several chapters, and as co-editor of the book. The author also extends her gratitude to Helena Pálková, for help with the conception and development of the cover. She also expresses her thanks to Will Gates who acted as co-author of several chapters and also, later during its final stages, as first main editor of this book. His help in sharing this last responsibility of managing this book was particularly invaluable. A special thought should be addressed to Theo Kloprogge who, despite a personal setback (a sad event that happened while editing this book), continued to be active writing important chapters for this book. Both Faïza and Will are very grateful for his effort and valuable contributions.

**(2)** *Will P. Gates* thanks the co-editors for asking him to join the team and his co-authors for three of the chapters. As co-author of Chapter 7, he acknowledges that some of the data were collected at the Australian Synchrotron during commissioning of the far-IR beamline. The authors of this chapter acknowledge *D. Appadoo* and *J.D. Cashion* for their able assistance in collecting the data displayed in Fig. 7.9.

**(3)** *Jana Madejová* acknowledges her co-authors for three chapters of this book, and *P. Komadel*, for many years of close collaboration in the IR study of clays and clay minerals.

**(4)** *Theo Klopproge* extends his thanks, first of all, to *Faïza Bergaya* for making a long-held dream come true after his forced retirement from the Queensland University of Technology. He specially acknowledges the co-editors for their tremendous help and support in what was for him personally an extremely difficult year. Finally, Theo wants to thank, in particular, *Will Gates* as he took over so much of the work of the main editor in getting this book to the finish line.

**(5)** All four co-editors express their thanks to the external authors and co-authors of the different chapters of this volume for their useful and timely contributions which made this book possible. Last but not least, several anonymous reviewers are thanked for their precious help in improving the scientific quality of this book. The hard work of the authors and reviewers has resulted in a hopefully excellent addendum to Farmer's book.

**(6)** Co-author of Chapter 4, *Georgios Chryssikos*, thanks *V. Gionis* (NHRF) for many years of close collaboration in the study of clays, and *G. Kacandes* (Geohellas S.A.) and *M. Stefanakis* (S&B Industrial Minerals, now Imerys) for setting the challenge and providing the means.

**(7)** The authors of Chapter 14, *Janice Bishop*, *Joseph Michalski* and *John Carter* are grateful to *R.E. Arvidson* and *J.J. Wray* for their helpful comments that greatly improved this chapter.

This page intentionally left blank

Chapter 4

# Spectral Manipulation and Introduction to Multivariate Analysis

G.D. Chryssikos and W.P. Gates

## 4.1 INTRODUCTION

The vibrational spectrum of any clay mineral is complicated for a number of reasons. First, it represents a complex and polyatomic unit cell with both covalent (layer) and ionic (interlayer) features. In addition, many clay minerals of interest are formally defined over a range of compositions with variable type and degree of chemical substitutions: The species montmorillonite, for example, cannot be associated with a unique infrared (IR) spectrum. Further, a smectite-enriched sample, such as bentonite, may contain many clay mineral phases (e.g. montmorillonite, beidellite, illite and kaolinite) and variable amounts of accessory minerals (e.g. quartz, carbonates and sulphates). All these factors of chemical and mineralogical complexity are convoluted in the spectrum of a single sample and need to be distinguished prior to or during analysis (e.g. Balan et al., 2001; Madejová et al., 2002).

Another form of complexity related to modern IR and Raman spectral analysis comes from the spatial distribution of minerals or other species within samples which may have been collected to represent the larger object of study. Modern IR and Raman instrumentation offers the possibility of collecting hundreds or thousands of spatially resolved spectra efficiently and often noninvasively (see Chapter 3). Spatial complexity, for example, in thin sections studied in reflectance by mid-IR (MIR) microscopy, or in a set of drill cores from the exploration of a deposit studied by near-IR (NIR) spectroscopy, is recorded in large databases. These are called hyperspectral because they contain a (two- or, often, multiple–dimensional) grid of sampling points expressed in spatial coordinates and an extra dimension of a full spectrum


Developments in Clay Science, Vol. 8. https://doi.org/10.1016/B978-0-08-100355-8.00004-7

for each point. As expected, hyperspectral data contain large amounts of information that can be difficult to handle, much less interpret. But they can be analysed to yield the key spectral components that describe the variance of the system of interest, as well as the spatial distribution of these components within the macrosample. This leads to applications like hierarchical classification schemes and identification algorithms for the improved understanding of complex mineralogy. The same spectral databases can also be correlated with independent chemical, mineralogical and physical parameters of the sample, and trained to provide quantitative predictions on unknown samples. The development and application of such qualitative and quantitative prediction using multivariate analysis and chemometric tools can reduce drastically the time, cost and environmental footprint associated with the analysis of a large number of individual samples by conventional methods.

The terms *multivariate analysis* and *chemometrics* are often used interchangeably in the literature, but they are not equivalent. Multivariate analysis is a method for revealing underlying latent descriptors in a dataset and can be called chemometric only when calibrated to predict properties quantitatively. Multivariate analysis can be done independently of chemometric methods, but the opposite is not true. For example, principal component analysis (PCA) (see Section 4.4.1) is often used purely as a data exploration tool. Multivariate analysis and chemometric methods are not limited to vibrational techniques and clay minerals, and can be applied to any large datasets obtained by any technique that can provide an information-rich output regarding the materials of interest and their properties. Among vibrational techniques, multivariate analysis and chemometric methods have been historically based first on NIR spectra, primarily because these were the first that could be collected noninvasively and with sufficiently high signal-to-noise ratio, but also because NIR features were too complicated to assign and analyse using a simple Beer–Lambert approach. Chemometric analysis of MIR data came only after the advent of diffuse and attenuated total reflectance (ATR) techniques. Vibrational chemometric tools are widely applied in diverse fields including food science (e.g. González-Martín et al., 2002; Jie et al., 2013), biology and medicine (e.g. Grootveld, 2015), pharmaceutics (e.g. Reich, 2005), soil science (e.g. Tauler et al., 1996; Viscarra Rossel et al., 2006; Samuels et al., 2006; Reeves and Smith, 2009; Waruru et al., 2014) and mineralogy and ore processing (e.g. Andrés and Bona, 2005). The application of multivariate or chemometric methods to the processing of vibrational (mostly NIR and MIR) datasets has proven to be highly effective in extracting useful information, and this chapter is an attempt to introduce the reader to the salient aspects of these methods. As a prerequisite to multivariate analysis, but also as an independent part of spectral analysis, the following section covers the most common methods used in the mathematical treatment of

vibrational spectra to enable their semiquantitative (identification) and quantitative analysis.

## 4.2 OVERVIEW OF POSTCOLLECTION SPECTRAL PROCESSING

The use of any spectroscopic technique based on resonance absorption is underpinned by an understanding of the Beer–Lambert Law, namely:

$$A = -\log(T) = -\log(I/I_o) = \epsilon \cdot b \cdot c \tag{4.1}$$

where $A$ is the absorbance expressed as the negative logarithm of the transmittance ($T = I/I_o$) of light through a sample ($I$) to the initial light intensity ($I_o$), $\epsilon$ is the wavelength-dependent molar absorptivity (also known as the absorption or extinction coefficient), $b$ is the path length of light through the sample and $c$ is the molar concentration. The Beer–Lambert law describes the experimentally observed exponential attenuation of electromagnetic radiation through a homogeneous absorbing (and nonscattering) medium as a function of both distance and concentration (Griffiths, 2002). As such, the Beer–Lambert law is applicable to IR absorption but not to Raman scattering. In IR spectroscopy, the absorbing medium can be a molecule (e.g. interlayer $H_2O$), a polyatomic ion (e.g. $CO_3^{2-}$ or $NH_4^+$) or a functional group (e.g. structural OH bonded to different octahedral cations in montmorillonite) or, collectively, a crystalline or amorphous network of covalent bonds. The exact energy and intensity profile of IR absorption, expressed by the wavenumber-dependent coefficient $\epsilon$, is a useful fingerprint property specific to the vibrating species involved and depends on the dipole moment changes associated with its symmetry-dependent vibrational transitions (selection rules), as well as its local and crystal symmetry (see Chapter 2).

The Beer–Lambert law applies to multicomponent systems provided that there are no chemical interactions or matrix effects. In the ideal case where the sample is of known thickness and the observed vibrational transitions are well separated from each other by a zero baseline, the law allows for both the identification and the quantification of the absorbing species via their unique wavenumber dependence of $\epsilon$. If $\epsilon$ is not known, it can be obtained from a series of standards of variable concentration $c$, or estimated by theoretical analysis (e.g. Balan et al., 2009).

The Beer–Lambert relation typically behaves linearly over the range $0.2 < A < 1$, and for many oscillators of interest in clay science (such as the O—H or Si—O bonds) this range may impose unrealistic constraints on concentration and/or sample thickness. Depending on the sample, these constraints may be critical in determining the proper IR measuring technique (see Chapter 3). Additional reasons causing nonlinearity, such as light

scattering, electronic absorption, reflection phenomena and associated optical dispersion effects, may be difficult to be explicitly accounted and call for determining the linearity regime of the law, prior to its application.

The ideal spectrum that can be fully assigned and quantified as described in the previous paragraph is rarely encountered, even less so in clay minerals. In addition to the chemical and mineralogical complexity that yields spectra with many strongly overlapping features, noise and broadly varying non zero baselines are common features of these spectra. In fact, this situation applies to both IR and Raman spectroscopies. As a result, the analysis of the vibrational spectra requires mathematical pretreatments to reduce or eliminate non vibrational features, as well as to identify and assign discrete vibrational bands, prior to extracting quantitative (or semiquantitative) information. Unless otherwise indicated, these pretreatments apply to the analysis of both, single spectra and large spectral datasets.

### 4.2.1 Smoothing

Smoothing aims at reducing the noise level in a spectrum without reducing the number of spectroscopically significant variables. Usually smoothing is applied in cases where noise reduction by increasing acquisition time is impractical (as is the case in high-throughput measurements) or not possible. Application of smoothing is based on the assumption that noise peaks are much sharper than the vibrational bands of interest, and smoothing thus serves as a low-pass filter. Smoothing routines use different equation forms where the values of the points on either side of a datum may be 'averaged' or applied algorithmically to estimate the central value in some way, and this average then replaces the original central datum. The most common smoothing routines used in spectroscopy are based on the Savitzky–Golay filter (Savitzky and Golay, 1964), which is also used in conjunction with differentiation (see Section 4.3.2). The filter employs polynomial functions, fit to either side of each successive individual datum point within a spectrum. The functions are calculated from a regression fit of each successive polynomial to a moving window of user-chosen filter-width (e.g. $n = 1$, 2, 3 or more) points on either side of each successive datum in the spectrum, to project calculated intensities for each datum in turn. The user should ensure that the filter width (i.e. the total number of points used in the smoothing, $2n+1$, multiplied by the spacing of the data, $\Delta\nu$, in $cm^{-1}$) and the order of the polynomial are chosen carefully. The filter should be of the same or smaller width than the nominal width of non noise features present in the spectrum, and the polynomial order should be adjusted to remove only noise. The larger the filter width and the lower the polynomial order chosen, the greater will be the resulting smoothing effect.

Smoothing results in decreasing spectral resolution and can affect the position, width and intensity of vibrational bands, as well as the ability to resolve components of vibrational envelopes. Therefore smoothing should be applied

only if needed, and with the appropriate adaptation of its parameters to the specific spectrum and its acquisition conditions. Any reference to the number of smoothing points employed should be accompanied by the frequency spacing, $\Delta\nu$, of the data, which is not necessarily identical to the spectral resolution setting of the experiment.

### 4.2.2 Baseline Corrections

The baseline of IR and Raman spectra are often excessively non zero in some regions that may make the detection of weak vibrational bands difficult. Baseline correction is used to minimise the presence of non informative background in the measurements, thereby enhancing the spectral features of interest. A strongly sloping background is very common in diffuse reflectance spectra due to variations in light scattering from irregularities in the sample: non uniform particle size, shape, sample packing or irregular surface texture (Martens et al., 2003), but also in the case of NIR, from the high-frequency tails of strong MIR bands. Similar effects are common in Raman spectra due to luminescence (fluorescence). Consistent data acquisition can minimise some of these effects, but it is usually not possible to remove all baseline issues. Such unwanted baseline effects can propagate in the spectra during subsequent treatment (e.g. normalisation, see the following) unless they are properly compensated for.

The simplest baseline correction is an offset correction applied to bring relatively flat (non sloping) baselines to zero. Similarly, a sloping baseline can be corrected by subtracting a suitable straight-line segment from the experimental spectrum over the wavenumber range of interest. More typical is the situation where the features of interest are located on a curved baseline, usually the tail of a broad and strong band of vibrational or electronic origin. This is, for example, the case of the OH deformation modes of smectites that appear on the low-frequency side of the strong Si—O stretching envelope (e.g. Vantelon et al., 2001; Gates, 2005), or the corresponding, NIR-active, OH combination modes (Post and Noble, 1993), which are observed on the high-frequency side of the strong, MIR-active, O—H stretching envelope (e.g. Madejová et al., 1994; Bishop et al., 1999; Gates et al., 2002; Zviagina et al., 2004; Petit et al., 2015). This unwanted contribution from the broad overlapping features is typically removed by subtracting background Gaussian functions from one or both sides of the wavenumber range of interest (e.g. Fig. 4.1). Obviously, this type of correction is critical in estimating the occupancy of octahedral sites in clay minerals from their IR spectra (e.g. Petit et al., 2015). Another common approach to baseline correction has been the simple subtraction of an artificial baseline 'spectrum' created to simulate the wavenumber dependence of the baseline over the spectral range or ranges of interest (e.g. Petit et al., 2015). Whereas these baseline correction

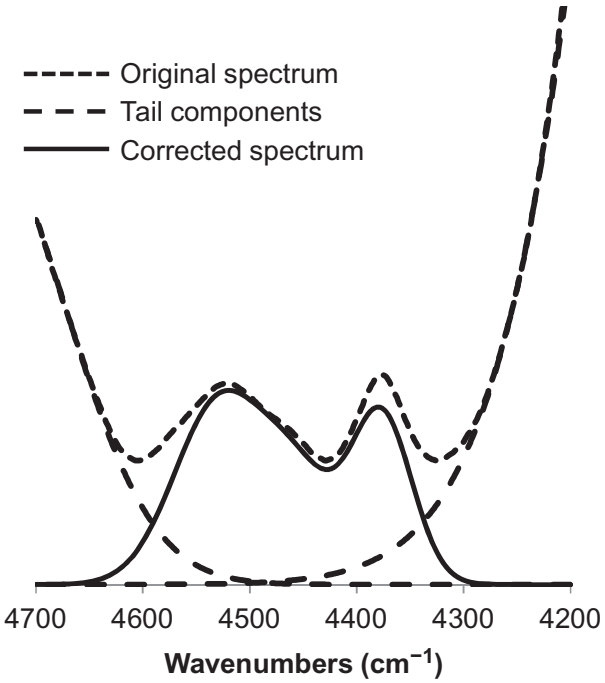



**FIG. 4.1** Baseline correction by removing Gaussian-shaped tails from the raw NIR spectrum of a synthetic smectite. Such correction enables least squares quantitative decomposition analysis of bands of interest. Spectral treatment applied by the authors to the IR spectrum of smectite synthesised and characterised by Andrieux and Petit (2010).

approaches tend to be somewhat empirical and subjective, they have served sufficiently for most applications, as long as they have been employed consistently across a spectral dataset and documented in detail.

### 4.2.3 Atmospheric Compensation

Compensation for the residual spectra of atmospheric gases is a special form of baseline correction in IR spectroscopy. Unless specifically removed, air within the optical path of an experiment often contains enough $H_2O$ vapour to produce a very rich vibrational-rotational spectrum consisting of a series of sharp bands at $\sim$1900–1300 $cm^{-1}$ ($\delta$), 4000–3400 $cm^{-1}$ ($\nu$), 5600–5100 $cm^{-1}$ ($\nu+\delta$) and 7400–7000 $cm^{-1}$ ($2\nu$). Similarly, residual atmospheric $CO_2$ yields a characteristic doublet at 2360, 2340 $cm^{-1}$ and a sharp bending mode at 668 $cm^{-1}$. The latter features are usually of a lesser concern, except perhaps in the case of deuteration studies or in the detailed analysis of the OH bending and lattice deformation modes of smectites. The bands of atmospheric gasses are very prominent in the single-beam IR spectra and should ideally cancel out when intensity ratios of the sample and background spectra are calculated. In many cases, however, this cancellation is incomplete, either because the optical path is open or, more seriously, because process monitoring measurements need to be performed over long times and cannot be interrupted for updating the background single-beam measurement

(Li et al., 2000). Despite overlap with the vibrational features of structural OH and $H_2O$ in clay minerals, the positive or negative residual rotational-vibrational bands of $H_2O$ vapour are easily discerned in high-resolution (e.g. 1 $cm^{-1}$ or higher) spectra because they are much sharper than the vibrational features of the sample and often can be corrected by application of an appropriate smoothing filter. Under more routine resolution settings (e.g. 4 $cm^{-1}$ or lower), however, the bands of $H_2O$ vapour appear broader and may therefore be confused with the sharpest bands of the sample (e.g. the structural $Mg_3OH$ stretch of talc) and as such, should not be remedied with smoothing. Unfortunately, the incorrect assignment of $H_2O$ vapour bands to all sorts of fictitious chemical species in the sample is not at all rare in the literature.

Modern commercial spectroscopic software offers general-purpose atmospheric compensation options that can be satisfactory for the routine presentation of absorbance and reflectance spectra, but are unsuitable for subsequent derivative analysis. The best practise to avoid interference from $H_2O$ vapour is to isolate the optical path as much as possible from the atmosphere (which is more easily achieved in ATR spectroscopy) and to equilibrate by desiccants and/or purging with dry $CO_2$-free air or $N_2$. Any remaining positive or negative changes of $H_2O$ vapour within the instrument itself are best compensated by measuring a second background spectrum after the completion of the experiment and using it to remove by subtraction any residual contribution for the vapour (e.g. Bukas et al., 2013). This procedure essentially interpolates two reference background spectra collected before and after sample measurement to compensate in full the individual spectra of a series of successive measurements. The only requirement for the successful vapour band removal by subtraction is that temperature remains constant during measurement to avoid shifting the $H_2O$ vapour spectrum.

All of this described here refers to IR spectroscopy. Raman spectroscopy with excitation in the visible spectrum is generally immune to such atmospheric effects because the absolute frequency of the measurement is well away from the vibrational spectrum of $H_2O$ (see Chapter 3). FT-Raman spectra, however, collected with 1064-nm excitation ($\sim$9400 $cm^{-1}$) can exhibit atmospheric interference, for example, in the $\Delta\nu = 2400–2000$ $cm^{-1}$ range from the $2\nu$ vibrations of $H_2O$. Such interference is in the form of negative peaks on a nonzero, sample-dependent background (due to luminescence or blackbody radiation, for example), and its compensation is difficult.

### 4.2.4 Normalisation

Normalisation is used to adjust the intensity of a spectrum to the same scale as other spectra in a dataset in order to assist visual comparisons or quantitative analysis. Normalisation helps to remove non systematic effects, for example

when sample concentrations may differ in dilutions, or when the path length of a transmission cell may differ between experiments, and serves to give all sample spectra equal impact for comparison. Another important use of normalising spectra is when the signal is a function of source power rather than sample concentration, as may occur in some instances in synchrotron-based IR or Raman spectroscopy or when a comparison between data obtained by different techniques over the same energy range is required (e.g. IR and Raman). Several procedures exist for the normalisation of IR spectra, the most common of which are area normalisation, vector normalisation and multiplicative scatter correction (e.g. Randolph, 2006). Other normalisation procedures exist, but are less commonly used in the vibrational spectroscopy of clay minerals and will not be discussed here. Normalisation can be performed on a specific band, or envelope of bands, even the entire spectral range, depending on needs. It can be applied to spectra in all forms, including absorbance, transmittance, derivatives, and others.

Most normalisation procedures transform the spectral features within a set of spectra to represent the same number of oscillators distributed over two or more component bands within the frequency range of interest. This is approximated by converting the spectra to the same integrated intensity (area) over the range of interest and explains why this should be done on baseline-corrected spectra. The critical assumption underlying this approximation when normalised data are used for quantitative analysis is that the absorptivity, $\epsilon_i$, of the component bands (scattering cross-sections in Raman experiments) are identical or, at least, similar to each other.

*Area and intensity normalisation*. Area normalisation transforms the observed data so that the area under the baseline-corrected spectrum in the frequency range of interest is the same for all samples. Intensity normalisation on the same band (usually the maximum of the strongest) within the region of interest can be applied instead of area normalisation, provided that the band position and width remain fixed across a series of spectra. This is the typical normalisation applied in cases where the sample is spiked with a suitable non-interacting, IR- or Raman-active admixture that serves as an internal standard. Maximum-weighted normalisation divides each spectrum by the maximum intensity observed, and therefore all spectra have vector norms that are scaled by their maximum values.

Vector normalisation refers to a group of normalisation methods that are based on linear algebra rather than empirical as is often the case for intensity or area normalisations. These normalisations attempt to give each spectrum equal impact on any model subsequently developed, and also assist with visualisation. Vector normalisation is based on the vectorial definition and properties of a spectrum $\boldsymbol{x}$, where $\boldsymbol{x}$ is a $1 \times k$ matrix row vector described by:

$$\boldsymbol{x} = [x_1, x_2, \ldots x_n] \tag{4.2}$$

where $x_i$ is the measured intensity and $n$ is the number of points over the wavenumber range of interest. The method transforms spectrum $\boldsymbol{x}$ so that its vector norm becomes equal to 1. This is done by subtracting first the average spectral intensity from the spectrum, leading to a new spectrum, $\boldsymbol{x}' = \left[x'_1, x'_2, \ldots x'_n\right]$, with both positive and negative intensities so that $\Sigma x' = 0$. The sum of the squares of the intensities $\Sigma(x'^2)$ is then calculated and the spectrum $\boldsymbol{x}'$ is normalised by the square root of this sum. Similarly, standard normal variate (SNV) normalisation scales $\boldsymbol{x}'$ to the standard deviation of the data points in the spectrum (Barnes et al., 1989; Rinnan et al., 2009). Absolute value vector normalisation divides each spectrum by the sum of the absolute value of all intensities, $|x_{ij}|$, in the spectrum. This returns a vector norm with a unit area, $w_i$, under the spectrum equal to 1. An example is given in Fig. 4.2, where spectra (Fig. 4.2A) have been vector normalised (Fig. 4.2B).

*Multiplicative scattering correction*. Within the general frame of normalisation transformations, multiplicative scatter correction (MSC) is a special type of compensation for variable individual baseline contributions within a set of spectral data. As discussed earlier, baseline variability within the dataset can be caused by particle size, sample packing and density variations, inhomogeneous particle distribution as a function of depth of sample or sample surface roughness, and is particularly common in diffuse reflectance spectroscopy (Afseth and Kohler, 2012). All these poorly controlled factors influence the optical path length, $b$ (Eq. 4.1), and create additive or multiplicative effects on the spectral baseline (Rinnan et al., 2009; Huang et al., 2010). Some of these effects can be taken into account by application of the Kubelka–Munk theory (Yang and Kruse, 2004), but further normalisation within the spectral dataset is frequently needed. MSC cannot be applied to individual spectra and is mostly relevant to the analysis of large spectral datasets by, for example, PCA (see Section 4.4.4). MSC operates on the mean of the entire dataset, and it scales a group of spectra to an equalised reflectance for quantitative comparison. In its simplest form MSC relies on the assumption that any measured IR or Raman spectrum can be successfully approximated by the sum of a baseline offset and a variably amplified 'chemical' Beer–Lambert absorbance. Mathematically, this is expressed (Huang et al., 2010) as:

$$x_{ij}(\mathrm{MSC}) = a_{ij} + b_{ij}x_{ij} \tag{4.3a}$$

where $x_{ij}$ is the absorbance value of the spectrum $i$ at wavelength $j$; $a_{ij}$ is the corresponding baseline offset; $b_{ij}$ is the reference absorbance value for $i$ at each wavelength $j$ (Rinnan et al., 2009) and accounts for signal amplification effects due to variations in optical pathlength; and $x_{ij}$(MSC) is then the absorbance associated with the sample chemistry. Extended-MSC (EMSC) allows for further compensation of effects like sample mass differences

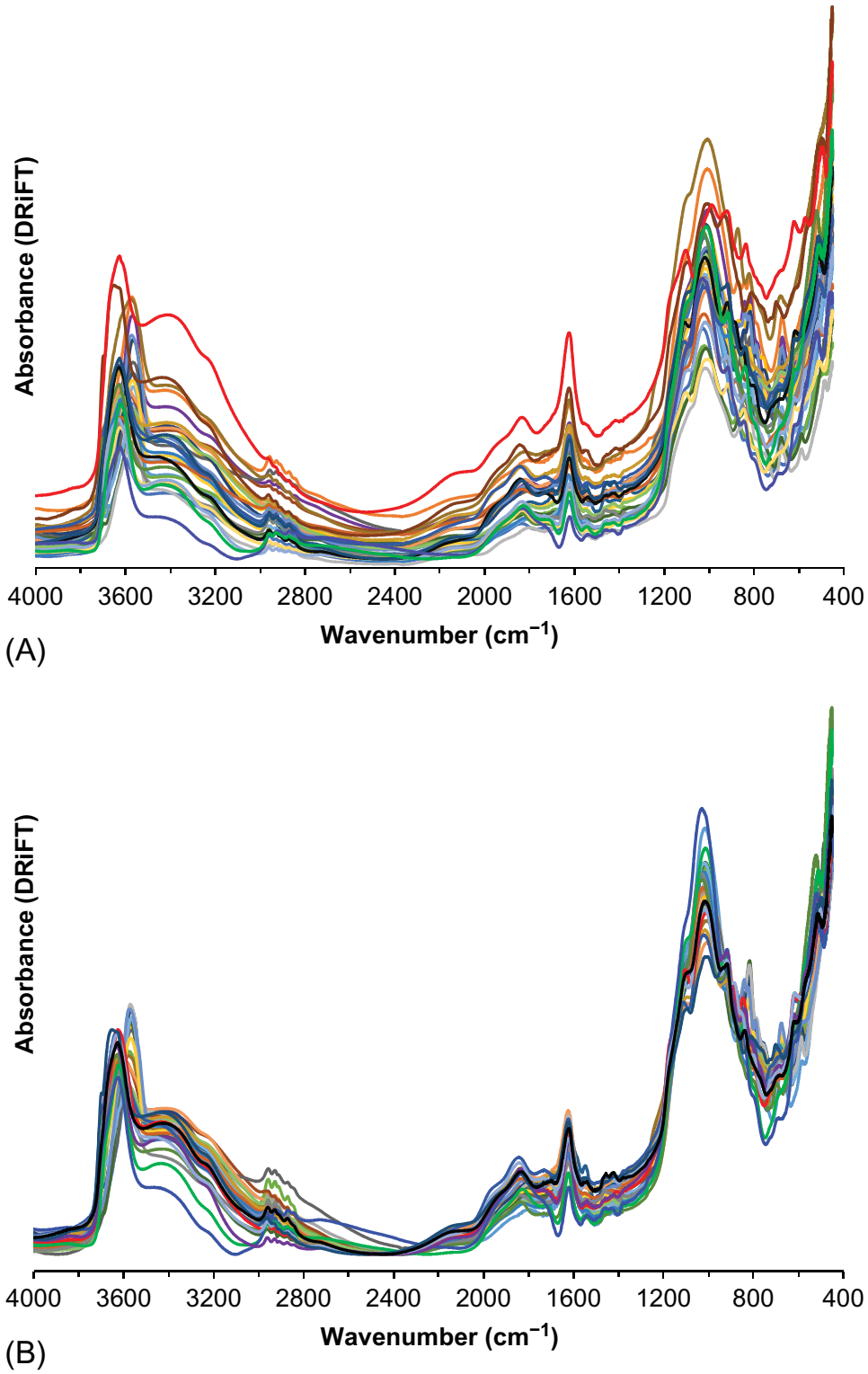



**FIG. 4.2** (A) MIR spectra of 32 dioctahedral smectites (neat air-dried, $Ca^{2+}$ form). (B) The same dataset after unit vector normalisation and offset correction. Absorbance values are the Kubelka–Munk transformations of diffuse reflectance (DRiFT) spectra.

and wavelength-dependent spectral effects (e.g. interference and path length variations) in order to remove their variance from the spectral information of interest (Martens et al., 2003). In EMSC a spectrum is expressed linearly as:

$$x_{ij}(\text{EMSC}) = a_{ij} + b_{ij}x_{ij} + d_{ij}\lambda + e_{ij}\lambda^2 \tag{4.3b}$$

where additional $d_{ij}$ and $e_{ij}$ terms, calculated from the dataset, are introduced to describe wavelength-dependent, $\lambda$, variations.

Both MSC and EMSC methods regress each individual spectrum, $i$, of a set of spectra against a reference, $b_i$, typically the average spectrum of the set, and it is important to realise that the (E)MSC corrected output depends on the set of spectra where this reference spectrum belongs (Maleki et al., 2007). In mathematical terms, correction has the form:

$$x_{ij\,\text{MSCcor}} = \left(x_{ij} - a_{ij}\right) \Big/ b_i \tag{4.4a}$$

$$x_{ij\,\text{EMSCcor}} = \left(x_{ij} - a_{ij} - d_i\lambda - e_i\lambda^2\right) \Big/ b_i \tag{4.4b}$$

In both expressions, $x_{i(\text{E})\text{MSCcor}}$ returns the frequency-dependent absorbance of the spectrum containing the chemical information of interest with minimal interference from nonchemical sources of variability (Li-Chan et al., 2011). The corrections are most commonly performed using a first-order polynomial (Rinnan et al., 2009). Since each spectrum is treated as such, direct intensity (area) comparisons become fully quantitative. Examples of both MCS and EMSC corrected spectra are shown in Fig. 4.3.

## 4.3 IDENTIFICATION AND SEPARATION OF OVERLAPPING VIBRATIONAL TRANSITIONS

In most cases, the mere identification, much less quantification, of the mineral components in a sample by vibrational spectroscopic techniques cannot generally be separated from the consideration of the chemical (compositional) variability of these components and its effect on the spectra. In fact, the complex mineralogy and crystal chemistry of the clay minerals call for the separation of overlapping vibrational bands, as a prerequisite to their assignment (identification) and subsequent quantification. Solution of the problem involves ideally the precise and reproducible determination of the exact number of component bands underlying a complex vibrational envelope, as well as the position, width and intensity of each component band. This determination is usually based on band decomposition and/or derivative procedures. The two approaches, each with its own advantages and deficiencies, will be outlined in the following sections.

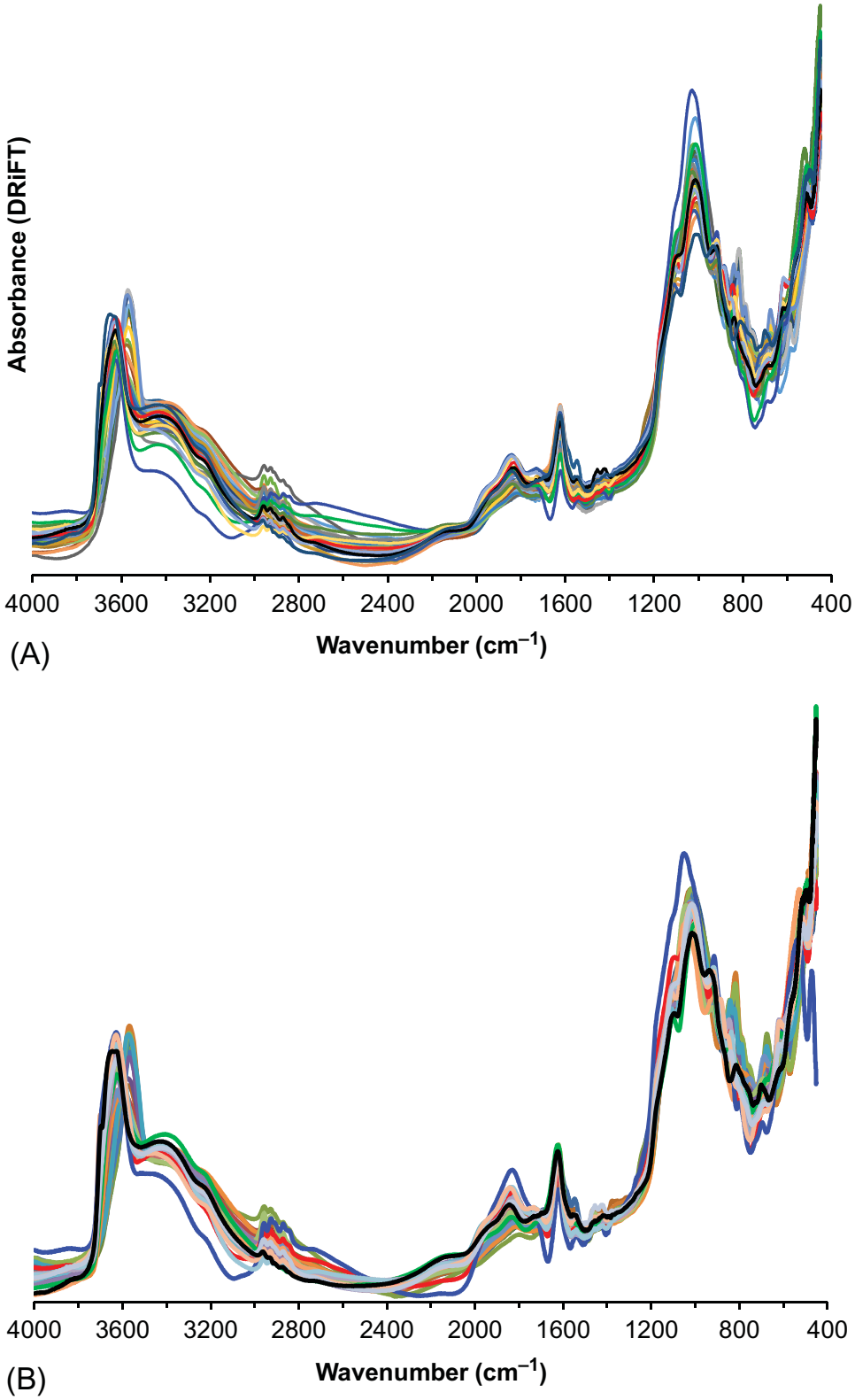



**FIG. 4.3** The (A) MSC-corrected and (B) EMSC-corrected Kubelka–Munk transformed MIR DRiFT spectra shown in Fig. 4.2A. Note the differences from vector normalisation (Fig. 4.2B).

### 4.3.1 Decomposition of Overlapping Bands

The compositional and structural variations inherent in smectites and other clay minerals result in band broadening and overlap, which negates the applicability of a simple intensity measurement for quantification. Instead, the broadened absorption envelopes can be analysed by separating into a number of overlapping components, where each component is made up of a distribution of intensities over a discrete frequency range. This process, called spectral band decomposition fitting, or simply decomposition analysis, provides a means to quantify spectra if the resulting component integrated intensities associated with each particular transition are summed. The goal in decomposition analysis is to simulate the experimental spectrum (typically in absorbance and after correction for baseline effects) over the frequency range of interest with a number of component bands (Fig. 4.4).

Decomposition analysis is an iterative curve-fitting problem, solved by minimising the least squares difference between the simulated (sum of components) and experimental spectra. Most available software allows for choosing the minimisation algorithm (e.g. damped or local least squares) and the least squares threshold for terminating the iteration. Many researchers have performed decompositions using spreadsheet formulations (e.g. Gates et al., 2002; Petit et al., 2015, 2016). Most fitting routines require initial 'guesses' about the number, shape (typically Lorentz, Gauss or Voigt), position, width and intensity of the component bands to produce an initial 'calculated'

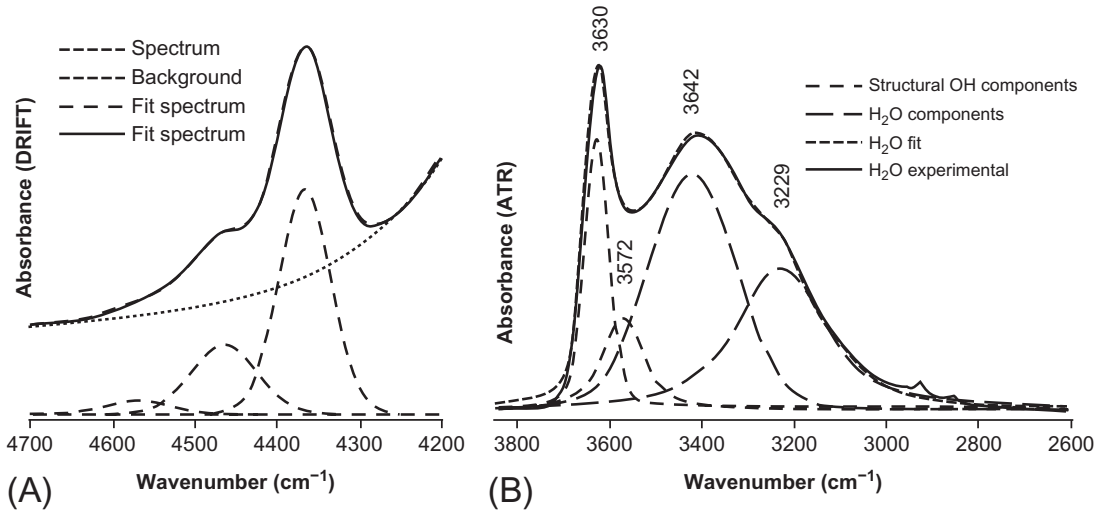



**FIG. 4.4** Representative decompositions of the IR spectra of smectites. (A) Decomposition of the OH combination ($\nu+\delta$) bands of NG-1 nontronite as shown in Gates et al. (2002) following baseline correction as presented in Fig. 4.1. (B) Decomposition of the complex OH stretch ($\nu$) bands of a montmorillonite following the procedure outlined by Madejová et al. (2002a). In (A) baseline correction by removing the tail on the low frequency side of the region of interest had little impact on the sensitivity of the fit and quantification of the two strongest bands. In (B), however, the fitting uncertainty of the 3572 $cm^{-1}$ component (which is not evident in the absorbance spectrum) can have a significant effect on the parameters of the 3422 $cm^{-1}$ component.

spectrum. The calculated spectrum is then least-squares fit to approximate the experimental spectrum by iterative changes of these initial parameters.

Each component is defined by three parameters (position, width and intensity) to be fitted (four in the case a Voigt function having variable Lorentz character), leading to a multidimensional global minimisation problem. In such problems, defining the single global minimum, rather than one among several possible local minima, is not straightforward and may require several optimisation runs starting from different initial conditions. Often it is also necessary for the operator to fix individual parameters in a controlled iterative fashion to ensure that the calculated minimum produces a realistic model of the experimental spectrum. Some programmes enable setting limits, or boundary conditions, to the parameters to minimise the opportunity for the fits to fall into 'false minima.' Obviously, for large spectral datasets, spectral decomposition fitting can become quite cumbersome even with modern software packages.

One pitfall to the method is that the introduction of more component bands into the model spectrum produces better fits without improving necessarily the physical meaning of the outcome (Gates, 2005). To achieve meaningful results, intuition and experience are important for deciding what parameters to constrain, fix, or allow to evolve freely during the fitting routine, as well as when to do so (e.g. Zviagina et al., 2004; Petit et al., 2015). It probably cannot be stressed strongly enough that a consistent approach is key to meaningful application of decomposition fitting.

Despite the difficulties in controlling peak fitting algorithms, several authors have utilised the method effectively to the study of clay minerals. Among the first successful applications of this method to smectites was an adaptation by Muller et al. (1991) of the breakthrough approach of Slonimskayá et al. (1986) on micas. Since then, peak fitting for spectral decomposition has been applied to the study of a variety of smectite samples for various purposes (e.g. Madejová et al., 1994, 2002a; Yan et al., 1996a; Besson and Drits, 1997a,b; Petit et al., 1999b, 2015; Frost et al., 2001b; Vantelon et al., 2001; Gates et al., 2002; Zviagina et al., 2004, 2015; Bishop et al., 2011). The popularity of spectral decomposition analysis stems from the fact that the relative integrated intensities (areas) of resolved components with common origin (e.g. structural OH) can be fed to a Beer–Lorentz-type quantification. Converting relative intensities to relative concentrations is based on the assumption that the absorptivity coefficients, $\epsilon_i$, of the resolved component IR bands (or the scattering cross sections of Raman components) are identical. The validity of this assumption should always be tested against independent analytical data (e.g. Gates, 2005; Petit et al., 2016). An overview on the applications of band decomposition analysis to the structural characterisation of dioctahedral clay minerals is discussed in greater detail in Chapter 7.

Spectral decomposition by least squares peak fitting approaches reaches its limits in cases where the experimental spectral envelope to be fitted contains both sharp and broad overlapping bands, especially in cases where the relative integrated intensity of the former is weak. Examples of such cases are encountered in the O—H stretching and overtone regions of the IR spectra of smectites, due to the complex overlap of the sharper bands of structural OH with the broader (and possibly variable) bands of $H_2O$, or to the coexistence of dioctahedral and trioctahedral clay minerals. In these cases a small acceptable error in the optimised solution of the strong broad components can induce large and unacceptable errors in resolving the weak sharp components. This is because fitting is usually done on spectra with fixed frequency intervals, $\Delta v$. Thus sharp features are represented by fewer experimental points than broad ones and have less influence on the minimisation criterion. For this reason, some software routines offer the option to use different frequency intervals within segments of the range of interest, but this option remains scarcely explored in the field.

### 4.3.2 Derivative Analysis

Some of the drawbacks of band decomposition analysis are conveniently remedied by the use of spectral derivatives, $d^nA/dv^n$. Derivatization is a multifunctional mathematical tool that can be used simultaneously for enhancing spectral resolution and filtering out broad spectral features, including sloping and curved baselines. This explains why derivative analysis has become very popular among NIR spectroscopists, although its use on other vibrational techniques is equally important. A deeper insight about the properties of derivatives and their use in spectroscopy can be found in Mark and Workman (2003).

For Lorentzian band shapes the amplitude of the $n^{th}$ derivative decreases with $n$, but also varies inversely as the $n^{th}$ power of the bandwidth. The dependence of Gaussian band shape amplitudes on the order of the derivative is weaker, but still very significant (Maddams and Mead, 1982a; Maddams and Southon, 1982b). This behaviour is the basis for the high discriminating power of derivative spectra for weak sharp peaks superimposed on intense broad peaks or background (Fig. 4.5), which increases sharply with derivative order. The main drawback of using higher-order derivatives is that the derivative amplitude of any random high-frequency noise decreases with increasing order of the derivative at a slower rate than the corresponding amplitude of the broader vibrational bands of interest. As a result, the existing noise in the spectrum becomes more pronounced, and the signal-to-noise ratio deteriorates significantly with increasing derivative order. The compromise between desired resolution and signal-to-noise ratio sets a practical limit on the maximum order of the derivative, which is usually the second derivative. In comparison to second derivatives, first derivatives have a better

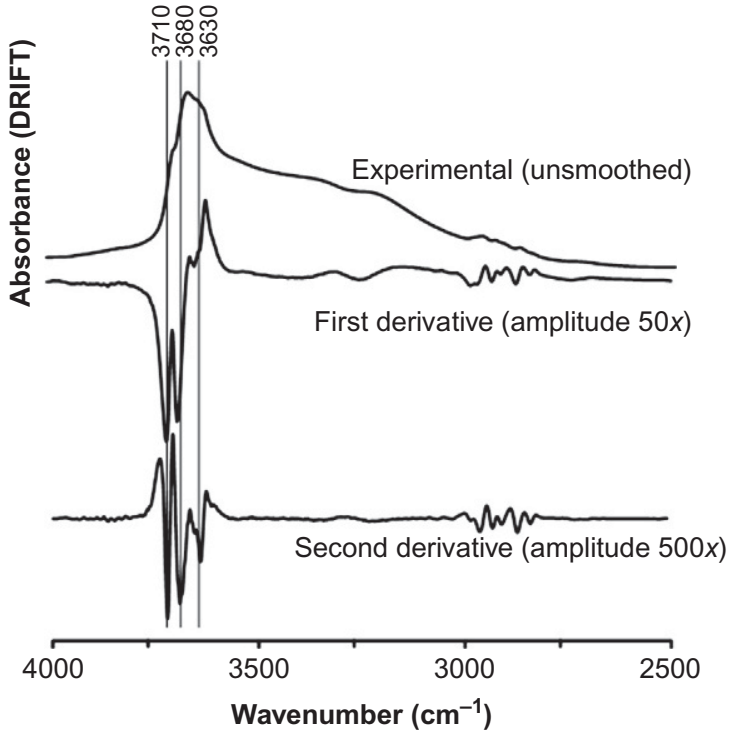



**FIG. 4.5** DRIFT MIR spectrum (2 $cm^{-1}$ resolution with $\Delta v = 0.96\ cm^{-1}$) of the OH stretching region of a smectite sample detailing how absorbance features within complex bands of experimental spectra can be better viewed after derivatization. Application of first derivative with Savitzky–Golay function over $2n+1=11$ points (50×); Application of second derivative $2n+1=15$ points (500×).

signal-to-noise ratio, but a smaller resolving power and a band shape that can be difficult to interpret (zero-crossings at peak maxima).

Derivatives can be computed in various ways, for example, as simple differences of adjacent raw data points, or over fixed-gap distances. Spectral derivatives are usually computed over a finite differentiation interval using an algorithm such as that parametrised by Savitzky and Golay (1964) and corrected by Steinier et al. (1972). This algorithm is applicable on spectra with equidistant data points, $\Delta v$, and computes derivatives at any point of the spectrum (except a few points at either end) on the basis of a range of $n$ points on each side of the central point, that is, over a spectral range of $(2n+1)\cdot\Delta v\ cm^{-1}$, similarly to smoothing routines described previously. Both $n$ and $\Delta v$ need to be selected judiciously because, although the increase of their product results in an increase of the signal-to-noise ratio, it also results in loss of spectral detail.

In practical terms, the aforementioned considerations regarding the properties of the derivative and their computation imply that all the spurious sharp features in a spectrum (noise, spikes, spectrum of atmospheric gases) must be eliminated prior to differentiation. Smoothing should not be used prior to differentiation because, as has been described, it is embedded in the Savitzky–Golay derivative algorithm. Similarly, offset, sloping or broadband

backgrounds will be eliminated or damped significantly by differentiation, so these corrections prior to performing derivatives are unnecessary. For a given experimental frequency spacing $\Delta\nu$ (which is typically controlled by the spectral resolution setting and any zero-filling factor employed), the smallest number of Savitzky–Golay points, $(2n+1)$, that can resolve the spectral bands of interest with satisfactory signal-to-noise ratio should be employed (Fig. 4.6). If the spectrum of interest contains peaks with different bandwidths, more than one value of $(2n+1)$ can be chosen for analysing different parts of the spectrum. At the expense of effort to acquire high-quality spectra, this procedure can yield a very clean picture of the sharper vibrational features of interest, with high discriminating power and free of interference from broader features, either spurious or intrinsic (e.g. bands of variable intensity due to physisorbed $H_2O$) (Fig. 4.7). Properly applied derivatization has been shown to be highly suitable for the qualitative assessment of the mineralogical or chemical composition of clay mineral samples and can be applied quickly on large sets of relevant spectra (Gionis et al., 2006, 2007). In contrast to band decomposition, derivatives can be applied to preprocess large sets of spectra in multivariate analysis where they assist in removing nonvibrational effects to create robust training sets (see Section 4.4.2).

On the other hand, the direct exploitation of derivatives for quantitative determinations on clay minerals is rare, because of the resulting lower signal-to-noise ratio of the derivative spectrum. Nevertheless, such quantitative analysis applications have been demonstrated on the basis of second derivative amplitudes as an extension to Beer–Lambert's law (Gionis et al., 2007; Chryssikos et al., 2009), or on the basis of the position of sharp $H_2O$ stretching bands (Kuligiewicz et al., 2015a,b).

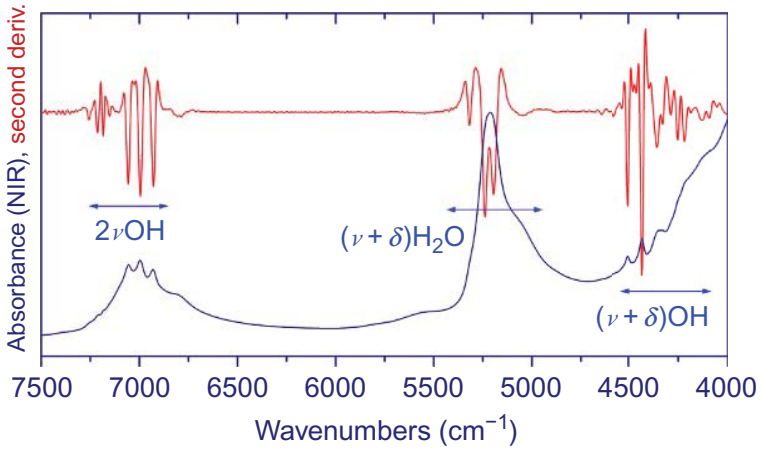



**FIG. 4.6** Vector-normalised NIR absorbance and second derivative spectra of a palygorskite sample demonstrating the excellent resolving power and baseline compensation of a properly tuned second derivative. The spectrum was collected by averaging 100 scans at 4 $cm^{-1}$ resolution ($\Delta\nu = 2\ cm^{-1}$), and the derivative was computed by application of a Savitzky–Golay, $2n+1=13$ function.

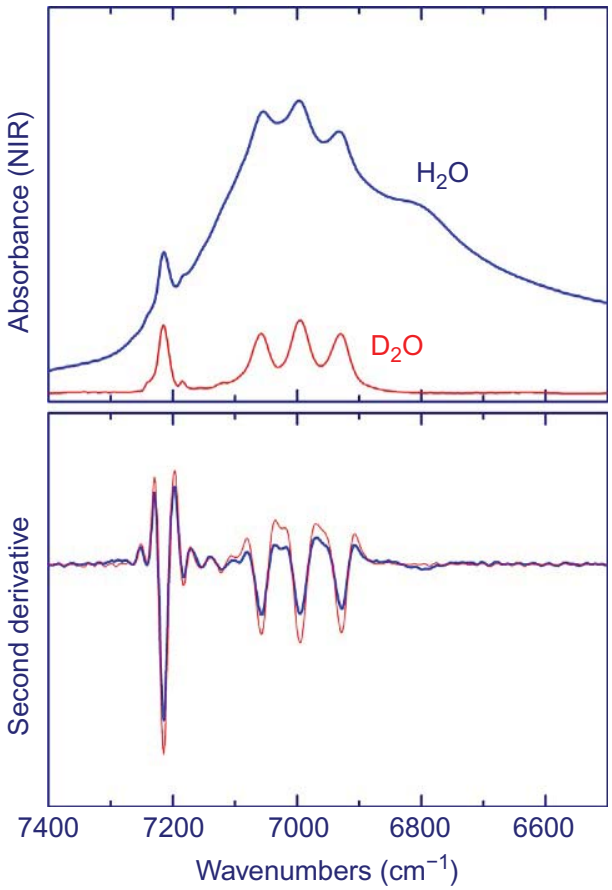



**FIG. 4.7** Experimental NIR absorbance data *(upper spectrum)* of a palygorskite specimen with mixed dioctahedral–trioctahedral character in its natural $H_2O$ form, and after exchange with $D_2O$ to remove the broad and complex spectrum of $H_2O$. The corresponding second derivative spectra *(lower panel)* are nearly identical to each other indicating that the accurate identification of structural OH modes can be obtained directly from the $H_2O$ form, without the need for deuteration. Sample drying was in this case out of the question because it affects the structure of the layers. This is a case where peak fitting the envelope of the $H_2O$ form (as exemplified in Fig. 4.4B) would introduce serious uncertainty, especially in cases of variable relative humidity. For more examples and technical details, see Bukas et al. (2013).

## 4.4 MULTIVARIATE ANALYSIS AND CHEMOMETRIC QUANTIFICATION

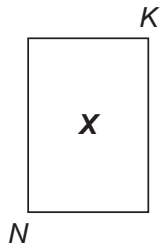
Multivariate analysis and chemometrics are a group of methods that provide a statistically meaningful basis for describing complex multiparametric phenomena (Kramer, 1998). When applied to large sets of vibrational spectra (e.g. Fig. 4.2), multivariate analysis enables the extraction of qualitative information that would otherwise be too complex to obtain by other means. Similarly, chemometric techniques allow for quantitative predictions in cases where the straightforward application of Beer–Lambert law would be impossible. These methodologies for qualitative and quantitative analysis rely on good-quality spectral data (the independent variables in the multivariate analysis) as well as good quality chemical or physical data (the dependent

variables in chemometric analysis). Application of such methods to NIR and MIR spectra of materials has proven to be an inexpensive method for rapidly identifying the mineral and chemical components in soils and rocks (Karstang et al., 1991; Sudduth and Hummel, 1993; Chang et al., 2001; Viscarra Rossel et al., 2006; Reeves, 2010; Reeves et al., 2010) as well as establishing predictive correlations against mineralogical, physical and chemical datasets associated with the same materials (Reich, 2005; Viscarra Rossel et al., 2006; Bona and Andrés, 2007; Gomez et al., 2008; Reeves and Smith, 2009; Mouazen et al., 2010; Kerr et al., 2011; Filgueiras et al., 2014). Recent advances in specialised commercial software packages enable rapid incorporation into the spectroscopy laboratory.

A brief introduction to matrix or vector approaches which form the basis for multivariate analysis is presented. This introduction is intended to provide a starting point for researchers interested in utilising the chemometric approach to the study of clays and clay minerals. For greater detail the reader is referred to many specialised literature sources including Mardia et al. (1979), Naes and Martens (1988), Naes et al. (2002) and Schlens (2014).

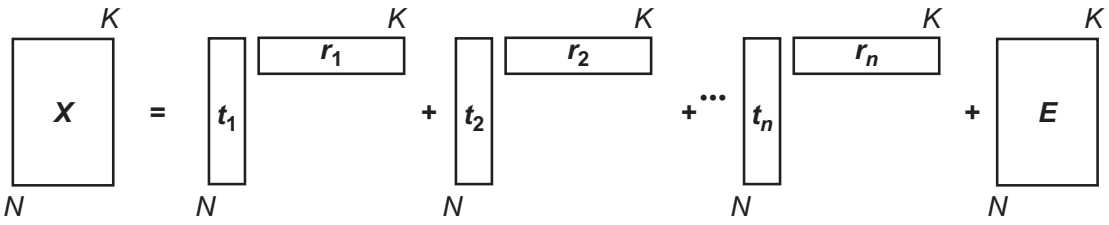
### 4.4.1 Introduction to PCA and PLS

Given the specificity and sensitivity of vibrational spectroscopy to chemical composition and structure, any large sample set can be measured and represented by a hyperspectral database (e.g. absorbance spectra) that is very rich in information content. The database can be represented as an $N \times K$ matrix, $\boldsymbol{X}$, where $K$ is the number of discrete frequency points (independent parameters) in each spectrum, and each of the $N$ rows represents a spectrum as a point in a $K$-dimensional space (see Section 4.2.4 on vector normalisation). This matrix is usually called *training* or *calibration set* in the context of multivariate analysis and chemometrics, respectively.




In the typical case of interest, the number of spectra, $\boldsymbol{N}$, is much larger than the number of chemical or mineral components that can possibly be present in the sample set, $n$. There also is sufficient variability between the spectra (rows of the matrix) at most frequencies (columns). Each individual spectrum is considered to represent a different (for simplicity, linear) combination of this limited number of chemical components (i.e. the spectra are highly correlated). Of course, each spectrum also contains errors from random noise and, possibly, instrumental artefacts.

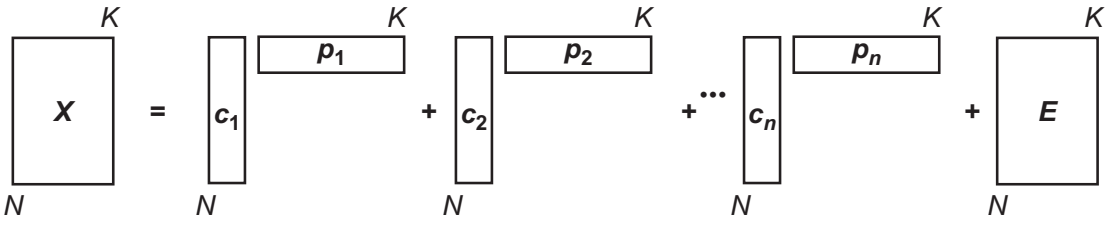
For the simplest analysis of a *multivariate* set of data, one should apply a classical least squares approach in order to express the matrix $\boldsymbol{X}$ as a linear combination of a finite number $n \ll N$ of reference spectra (row vectors $\boldsymbol{r_n}$), one for each of the chemical components of interest, such that the residuals, $\boldsymbol{E}$, are small and contain only error. In this way, every spectrum of the training set can be mapped on a predefined $n$-dimensional space formed by the reference vectors $\boldsymbol{r_n}$. This would be a *supervised* solution (the reference spectra $\boldsymbol{r_n}$ are chosen by the user) expressed in the form:




(4.5)

Eq. (4.5) contains both the qualitative (spectra $\boldsymbol{r_n}$) and quantitative (coefficients $\boldsymbol{t_n}$) information that fully describes database $\boldsymbol{X}$. Nonlinear variations to Eq. (4.5) also occur. It was already argued (see Section 4.3) that this approach can be problematic because of the lack of properly defined reference mineral spectra and the uncertainty surrounding how many spectra would be necessary to describe a database. The problem, therefore, calls for a *multivariate analysis* solution to identify and extract the main features underlying the dataset $\boldsymbol{X}$ in an *unsupervised* manner, without any input of the user. There are several methods available to perform this extraction based on multivariate statistics and linear algebra, and one of the most popular for decoding spectroscopic data is Principal Component Analysis (PCA) (e.g. Cowe and McNicol, 1985; Geladi and Kowalski, 1986; Bro and Smilde, 2014; Schlens, 2014).

PCA expresses the training matrix $\boldsymbol{X}$ (or rather, its mean-centred version) as a linear combination of calculated independent one-dimensional (*rank* 1) variables, called *loadings, components,* or more specifically *principal components*, $\boldsymbol{p_n}$, weighted by appropriate coefficients called *scores*, $\boldsymbol{c_n}$. The principal components represent calculated, *latent,* reference spectra and the scores represent their intensity changes within $\boldsymbol{X}$. The original set of spectra containing, for example, absorbance versus wavenumber data is now expressed into an equivalent set of scores on principal components.




(4.6)

PCA differs from the supervised modelling that was previously based on spectra of reference compounds, $\boldsymbol{r_n}$ (Eq. 4.5), in that the linear terms in the series are not preselected by the user. The linear terms are computed from the data in decreasing order of importance for explaining the variability in $\boldsymbol{X}$. The elements of $\boldsymbol{p_n}$, are linear combinations of the independent variables in $\boldsymbol{X}$, and the vectors $\boldsymbol{p_n}$ are the new, *latent*, independent variables. They are computed to be fully uncorrelated from each other, and they constitute the eigenvectors of the covariance in $\boldsymbol{X}$. As such they form an orthonormal vectorial basis suitable for the projection of any spectrum within $\boldsymbol{X}$. The scores, $\boldsymbol{c_n}$, are the corresponding eigenvalues, and they are scaled so that the sum of their weight coefficients squared is unity.

The PCA calculation starts by mean-centring the spectra in $\boldsymbol{X}$, through subtraction of the means across all dimensions of the data:

$$x'_{ij} = \left(x_{ij} - \bar{x}_{ij}\right) \tag{4.7}$$

where $x'_{ij}$ is the mean-centred result and $\bar{x}_{ij}$ is the mean value. This operation centres the variability along each of the original $K$ dimensions (frequency points) at zero. The first principal component, $\boldsymbol{p_1}$, is determined as the (eigen)vector along which $\boldsymbol{X}$ shows the highest covariance. The second principal component, $\boldsymbol{p_2}$, accounts for most of the remaining variability that is orthogonal (independent) to the first. The iteration continues, always maintaining orthogonality, until the variability that remains unexplained after the calculation of $\boldsymbol{p_n}$ approaches the size of the measurement error. The eigenvalues of the principal component eigenvectors constitute the covariance matrix.

The successful application and usability of PCA relies on the extent to which the dimensionality of the original matrix $\boldsymbol{X}$ can be reduced. In other words it depends on the *rank* of the solution, which is the minimum number of principal components in Eq. (4.6) that explain most of the variance in the original data (without explaining noise). As the principal components are computed in order of decreasing explained variance, the cumulative explained variance approaches 100% asymptotically at $n = N$ (Fig. 4.8). Fortunately, this convergence usually occurs at a number of ranks that is significantly smaller than the original number of spectra in $X$, $n \ll N$. In contrast to Eq. (4.5), where the rank is predefined (and limited) by the number of reference minerals used or available, in unsupervised PCA, a decision must be made about truncating the terms of Eq. (4.6). This decision depends in part on the contents of the training set, the amount and type of information that needs to be explained, and the identification of the principal component(s) along which this information can be retrieved most clearly. General guidelines regarding these aspects and specific examples on the development and application of PCA will be provided in subsequent sections.

Any spectrum within $\boldsymbol{X}$ can be uniquely located in the $n$-dimensional space of the principal components by its scoring on each component. The location

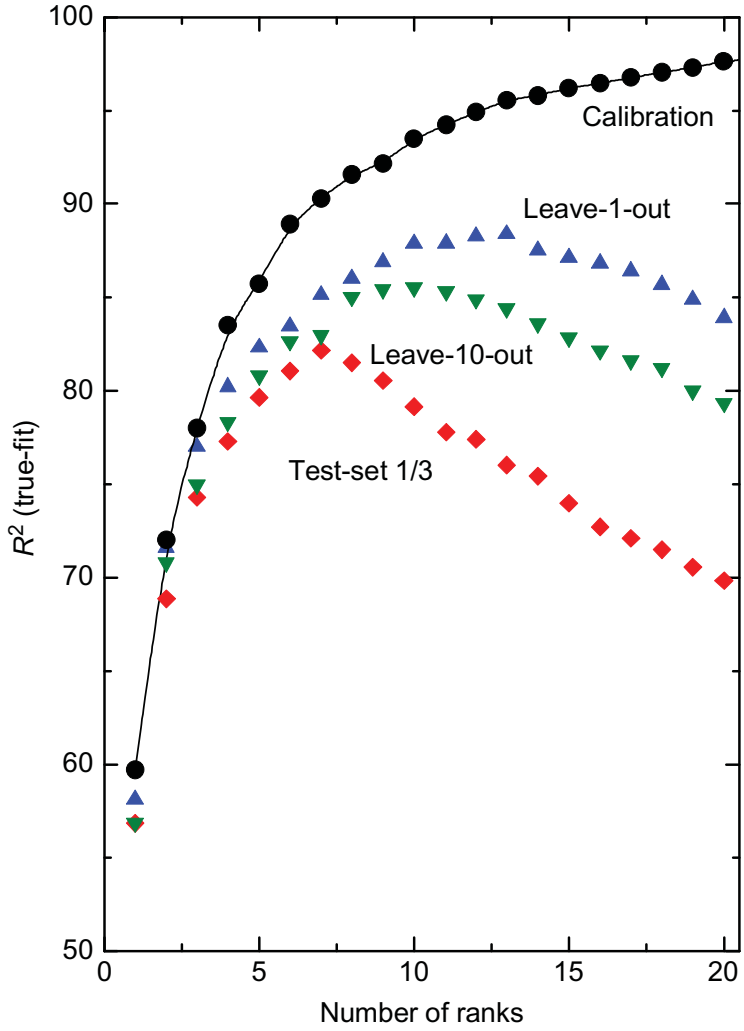



**FIG. 4.8** True-fit correlation coefficient $R^2$ of a case-study PLS calibration and three different internal validations for predicting a property of bentonite based on the NIR spectra of a set of 150 samples. All methods have been optimised in terms of the same spectral pretreatment and frequency ranges. Opposite to the trivial behaviour of calibration, all validations exhibit maxima with position (rank) and performance ($R^2$) that depend on the validation method. Leave-one-out cross-validation is clearly optimistic (overfitting), and the recommended number of ranks (~12) is high in comparison to the relatively small size of the dataset. On the other hand, a test set validation based on 50 spectra (1/3 of the dataset) affords a significantly smaller calibration set and may be underfitting at 7 ranks.

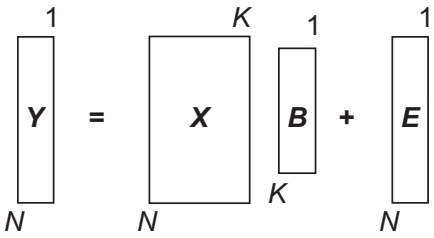
of several spectra can be projected on any plane defined by two principal components and be used for the unsupervised discrimination among groups of spectra, should the latter internal structure exist in the set. By the same means, unknown spectra that are relevant to the spectra in $\boldsymbol{X}$ can be identified as members of these groups. Although the common names of the groups may be coming from other techniques (X-ray diffraction or chemical analysis), the multivariate analysis of vibrational spectroscopic data from clay minerals and related materials is a powerful, independent and self-consistent data exploration tool for discrimination and identification applications.

Besides identification, there is often additional or independent need for quantitative predictions, as multidimensional extensions of the Beer–Lambert law. Such quantitative multivariate tools are commonly called chemometric due to their relevance in analytical chemistry, but they are important in many fields, even outside the physical sciences. The basic operating principles of two of the most popular chemometric methods will be presented here: the Principal Component Regression (PCR) which builds on the results of PCA (e.g. Naes and Martens, 1988), and PLS (standing for Projection to Latent Structures but commonly referred to as Partial Least Squares) which is wholly independent of PCA (e.g. Geladi and Kowalski, 1986; Wold et al., 2001). A detailed presentation of the two methods with common examples and notation can be found in Hasegawa (2006).

Training (more specifically, *calibrating*) a method to provide quantitative predictions for the sample set represented by the spectra within $\boldsymbol{X}$, requires that a set of independent, accurate property measurements $\boldsymbol{Y}$ is available for exactly the same sample set by other techniques (see Section 4.4.2).


PCR is based on calculating the correlations between the scores of each principal component derived by PCA and the property $\boldsymbol{Y}$. The user can then construct a predictive algorithm including only those principal components that correlate with the property and excluding all others as irrelevant. This approach is conceptually identical to testing how the absorbance values in a set of spectra at each individual wavenumber correlate with the property of interest and selecting the region of interest for a predictive model, but it is performed on an orthogonal vectorial basis with reduced dimensionality. Although the predictive principal components of a PCR method are by definition identical to those derived by PCA, they are not necessarily employed in the same order. This is because the principal components of PCA were determined by order of explaining the overall latent structure of the spectra, and not the subset of this structure that may specifically be related to any property $\boldsymbol{Y}$. One would ideally prefer that the $\boldsymbol{Y}$-predictive principal components are also among those that explain a large part of the overall spectral variability (meaning that they have a rather strong spectral signature), but this may not always be the case.

In the PLS method, however, the stage is set from the beginning for predicting $\boldsymbol{Y}$:




(4.8)

where a vector $\boldsymbol{B}$ is sought such that, when applied to $\boldsymbol{X}$, yields a prediction of $\boldsymbol{Y}$ with minimum error $\boldsymbol{E}$ (Geladi and Kowalski, 1986). In this sense, PLS is a method of dimensional reduction which is *a priori* supervised by the property data $\boldsymbol{Y}$. The computation of $\boldsymbol{B}$ involves the usual expression of $\boldsymbol{X}$ as a linear series of scores on loadings (Eqs. 4.5 and 4.6), but the loadings are now different than the principal components of PCA, because they are specifically calibrated to describe the variability in $\boldsymbol{Y}$ and obtained in order of decreasing ability to describe this variability. Further, the PLS loadings are not forced to be orthogonal, but their scores are. Similar to PCA, an important decision must be made about the minimum number of terms (dimensions) that are needed to fit the property without fitting the error (*overfitting*).

### 4.4.2 Training (Calibration) and Property datasets

As elegantly stated by DiFoggio (2000), the fundamental assumption behind any multivariate regression or chemometric method is that the (training or calibration) set of spectra $\boldsymbol{X}$ must contain all the information that is sought by the method. This assumption sets specific requirements about both the contents of the spectral matrix $\boldsymbol{X}$, and the kind of properties, $\boldsymbol{Y}$, that can be reasonably extracted. These requirements will be discussed in this section.

The training set $\boldsymbol{X}$ can be selected from a larger set of spectra (the usual case) or be developed wholly separately from reference materials and then applied to the analysis of another dataset. The quality of the training set controls to a large degree the performance of any multivariate or chemometric application. The number of spectra to be included in $\boldsymbol{X}$ is rarely an issue because matrices of very large size can now be manipulated with commonly available computing power. The number of spectra chosen for the training set, however, must be much larger than the number of anticipated latent variables $N \gg n$ usually by a factor of 10–100. Ideally, every spectrum in $\boldsymbol{X}$ should be a linear combination of many, if not all, anticipated latent variables: The spectra in $\boldsymbol{X}$ must be correlated. In fact, the whole normal distribution of each anticipated variable should be ideally represented in $\boldsymbol{X}$, and this is the main reason behind the need for large sample datasets. A *design of experiments* approach is frequently useful in selecting spectra for $\boldsymbol{X}$ that meet these requirements among even larger groups of available data (Kjeldhal and Bro, 2010).

As a general rule, it is inadvisable to include the spectrum of a pure reference compound in $\boldsymbol{X}$, if the latter is meant to represent a population of real-world mixtures that are too different from this reference. In such a case, the reference compound would be singled out as the first principal component in PCA (because it would be explaining a very large part of the variance in the whole set), but would be too different (distant, in the geometrical sense) from the rest of the samples and hence of little value in describing the finer details of their latent structure. In fact, including the spectra of all anticipated reference compounds in $\boldsymbol{X}$ would reduce the problem to the supervised approach expressed in Eq. (4.5). The same situation is encountered if the variance of nonvibrational features in the spectra (e.g. strong and variable sloping baselines or atmospheric gas signatures) dominates the overall variance in $\boldsymbol{X}$. These features would be promoted to the early principal component ranks, despite the fact that they are completely irrelevant to the chemical structure or mineralogical composition of the samples. This explains why the removal of such non vibrational sources of variance by appropriate mathematical pretreatments (see Section 4.2) is of paramount importance in multiple regression applications.

In practise, the statistics of the scores of each loading do reveal whether the representation of this loading in $\boldsymbol{X}$ is close to the desirable normal distribution or, instead, is mostly determined by a small isolated minority of spectra. If a particular loading is lacking such statistically distributed scores, one should decide whether this loading represents an outlier (that should be rejected prior to redoing the analysis), a spurious effect in the spectra (that should be corrected by an appropriate pretreatment), or a real chemical component which is rarely (but truly) present in the spectra of $\boldsymbol{X}$. In this latter case, one may decide to exclude the minority of spectra that represent the suspect component in $\boldsymbol{X}$, or enrich the spectral set with additional spectra bridging the compositional (or property) gap between the majority and minority populations. The decision depends on the anticipated significance of the chemical component with respect to the problem of interest. Tuning the contents of the spectral set $\boldsymbol{X}$ and deciding on the necessary mathematical pretreatment of the spectra that removes spurious effects and provides a satisfactory compromise between spectral resolution and signal quality typically requires a few trial analyses.

As an additional general rule, the spectra to be included in $\boldsymbol{X}$ should be of the same quality as those that are anticipated for future analysis. Methods based on high-quality laboratory data often lose their robustness when applied to field measurements. This is because the latter may introduce additional sources of spectral variability that are not encountered in the former and have not been considered by the model. In cases where this situation cannot be avoided, a *desensitisation* approach can be adopted, involving the *a posteriori* addition of the missing variability in the calibration spectra (DiFoggio, 2000).

These guidelines are sufficient for exploring the latent structure of spectral sets by multivariate analysis, but additional considerations are needed for chemometric (quantitative) applications. For example, what are the properties, ***Y***, that can be considered for chemometric analysis based on the spectra of a set of mineral samples, ***X***? In principle, ***Y*** can be any property that can be related fundamentally or empirically to the composition and local chemical structure (bonding and symmetry) of the samples; that is, to the latent features normally found in vibrational spectra. The elements of the matrix ***Y*** must come from independent measurements, and these measurements must be as accurate and precise as possible, because their quality characteristics will be transferred at large to the chemometric prediction tool (DiFoggio, 2000).

In practise, chemometrics serve to substitute for the routine measurement of properties that would otherwise be determined by time-consuming, expensive or environmentally problematic standardised measurement protocols (Viscarra Rossel et al., 2006; Waruru et al., 2014). In many cases, chemometrics are called to be based on (and eventually substitute for) existing quality control schemes. These may include mineralogical (e.g. % quartz in bentonite), chemical (e.g. moisture content, cation exchange capacity [CEC], or methylene blue [MB] adsorption, Al occupancy in the tetrahedral sheet, total organic carbon), physical (e.g. refractive index, viscosity of a dispersion) or engineering (e.g. swelling capacity, Atterberg limit) properties. One must be fully aware of the limitations and assumptions inherent to the measurement protocols that are used to produce ***Y***. One should also make sure that these protocols are applied exactly to the same samples that produced the spectra in ***X***, are conducted in the exact fashion for all samples and preferably to the same type of samples, and in the same fashion, anticipated for future analysis. In cases where field, rather than laboratory, applications are sought, the requirement for matching data in ***X*** and ***Y*** can set major difficulties in assembling these matrices. For example, in designing a chemometric application for the field measurement of moisture in soil, one must make sure that the natural moisture of the calibration samples is fully preserved until both spectroscopic and thermogravimetric measurements are completed.

One must also be aware that a chemometric prediction is based on the correlation between the spectra of a particular set of samples and their properties. For this reason, chemometric predictions may perform very well, but only within the domain set by their latent structures. Chemometrics are not direct measurement techniques and therefore cannot be used to predict the properties of any set of samples that lie outside that range of properties of the set they have been calibrated against. Two subtle issues regarding the information content of the spectra in ***X*** may be added at this point:

Can the quality of the chemometric predictions surpass that of the independent measurement? The intuitive answer is ‘no’, but the regression

mathematics used for establishing the chemometric tool average out random error in the reference values and can, in principle, improve the precision (DiFoggio, 2000). In addition, the study of the same sample set by two independent experimental methods feeding $\boldsymbol{X}$ and $\boldsymbol{Y}$ often reveals the presence of *outliers* (wrong samples, wrong measurements, typographical errors, etc.) in either or both sets of data. Outliers are samples that are (1) described badly by $\boldsymbol{X}$ or $\boldsymbol{Y}$ or (2) outside the range of predictive capability of the method. Elimination of these outliers improves the accuracy of the measurement.

Can one employ chemometrics to measure concentrations of a species that are unobserved in spectra? Again, the intuitive answer is 'no', and is correct unless the concentration of the invisible species is (positively or negatively) correlated with that of other constituents that have a pronounced spectral signature. A typical example is the prediction of % quartz in clay mineral samples by NIR spectroscopy, despite the fact that quartz is transparent in the NIR. Such predictions are based on the assumption that quartz is the only NIR-transparent mineral in the samples and that its presence and concentration can be inferred from the absence of the remaining, nontransparent components. Such indirect determinations should be used with great caution.

### 4.4.3 Validation and Optimum Dimensionality

The single purpose of any multivariate analysis or chemometric methodology is the satisfactory reduction of the dimensionality (rank) of the problem from the initial large number of $N$ correlated spectra in $\boldsymbol{X}$, to a much smaller number of noncorrelated latent loadings, $n \ll N$. Recall that the linear decomposition of a matrix $\boldsymbol{X}$ consisting of $N$ spectra can involve asymptotically up to $N$ linear terms (e.g. Eq. 4.6) and that the information content of these terms is expected to progressively drift from significance to triviality. The fundamental question is how one can decide where to truncate the series, thereby defining the *optimum rank* of the method, in order to capture most of the significant latent structure in $\boldsymbol{X}$, or the best possible correlation with the external variable $\boldsymbol{Y}$, without fitting irrelevant variables and errors.

Some insight about the anticipated dimensionality of the solution can be obtained from a general knowledge of the sample set under investigation. For example, the octahedral sheet of palygorskite contains trioctahedral magnesian and dioctahedral aluminous-ferric components which are manifested in the O—H stretching spectra by discrete $Mg_3OH$, $Al_2OH$, AlFeOH and $Fe_2OH$ bands (Chryssikos et al., 2009; Stathopoulou et al., 2011). The palygorskite content and composition is then defined by three independent variables, that is, the trioctahedral fraction, the $Fe^{III}$ occupancy in the dioctahedral domains and the extent to which mixed Al-$Fe^{III}$ pairs can form. Indeed, both a Beer–Lambert approach based on the intensities of the O—H stretching

overtones and a 3-rank chemometric model yield the same $R^2 = 0.94$ performance in predicting the palygorskite content in samples containing other clay and nonclay minerals (Gionis et al., 2007). In the more general case of modelling all species in a mixture, the number of significant principal components would ideally be the same as the number of those mineral components that can be observed by the spectroscopic technique employed. Similarly, in a predictive model for CEC, the rank of the solution should be similar to the number of mineral species with exchangeable cations present, multiplied by the number of spectroscopically active chemical substitutions that are encountered in each of these species and may be determining their layer charge. Such considerations do not take into account the possible presence of other variables or noise that can be strong enough to obscure some of the anticipated loadings.

For these reasons, the optimum rank as well as the figures of merit of any multivariate or chemometric methodology at this rank (e.g. the percent explained variance in PCR, or the root mean square error of prediction in PLS) should never be derived from data that have been used in the calibration of this methodology. As with any scientific methodology, multivariate analysis and chemometrics should be evaluated on the basis of independent, but plausibly similar, data by a process called *validation*. For details about the significance of validation, the various validation strategies and their potential pitfalls, the reader is referred to the specialised literature (e.g. Shao, 1993; DiFoggio, 2000; Wold et al., 2001; Steyerberg et al., 2003; König et al., 2007; Esbensen and Geladi, 2010; Westad and Marini, 2015).

There are many types of validation depending on the way the validation samples are chosen. The strictest and most conservative approach is *external validation* which uses an independent dataset—collected as similarly as possible to the parent dataset, but different from the training set—to test the statistical significance and predictive capability of a chemometric model. The external validation set must be representative and large enough to enable a statistically relevant assessment of the robustness of the method. Depending on the sources of irrelevant variability that must be accounted for, this independent set may arise from different operators, with data obtained over different time periods, or even from a (geological) setting that is analogous, but not identical, to the one used for training or calibration.

Very frequently, the conditions for constructing a true external validation dataset cannot be met, or the externally validated methodology turns out to be too conservative for the specific application of interest. In such cases, validation is based on a subset of data that are removed from the parent dataset and is called *internal validation*. As with its external counterpart, the internal validation dataset must be sufficiently large and well spread over the latent space, but still constitute a small fraction of the parent dataset so that the structure of the remaining calibration set is not disrupted. A common scheme of internal validation, *test set validation*, involves the selection of 10%–35%

samples for validation purposes. The method of selecting validation samples for the test set (random, in blocks, or evenly spaced across $\boldsymbol{X}$) can be critical if there are hidden systematics in the numbering of the spectra in $\boldsymbol{X}$ (time series, blocks provided by different operators or equipment, blocks added to fill-in gaps in original data, etc). There are many possible permutations of test set over parent data, and therefore several test set validations of the same method can be performed to optimise the rank, identify gaps in the calibration set, remove outliers, etc.

Another popular method of internal validation is *cross-validation*, also known as the *leave-$N_v$-out* method. The method removes a set of spectra $N_v$ from the original set of $N$ spectra in $\boldsymbol{X}$, calibrates on the basis of the remaining $N–N_v$ spectra and applies this calibration to the excluded $N_v$ spectra for method evaluation. The process is iterated over all spectra in $\boldsymbol{X}$. In some cases, $N_v = 1$ and the method is called a *leave-one-out validation*. Cross-validations tend to be optimistic, especially in their leave-one-out version, and should be used with caution.

An example about the use of validation in determining the optimum rank of a chemometric method is shown in Fig. 4.8. The example is taken from the PLS modelling of a property based on the NIR diffuse reflectance spectra of ~150 bentonite samples in powder form. The true-fit correlation coefficient $R^2$ is shown as a function of dimensionality (rank) for calibration as well as three different validation runs. Upon increasing number of ranks, the $R^2$ of the calibration run tends asymptotically (and trivially) to 100%. There is no way to define the optimum rank from such data. On the contrary, the $R^2$ of any one of the three validations first increases with increasing dimensionality (albeit remaining always inferior to calibration), reaches a maximum and then decreases, which is a clear sign that the model is now fitting the property on the basis of irrelevant variables (*overfitting*). The optimum dimensionality (and associated performance) of the method can be evaluated from the rank at maximum $R^2$. The case study in Fig. 4.8, however, indicates that the suggested optimum dimensionality can vary between the optimistic leave-one-out validation and the conservative test-set validation. If encountered, the PLS correlation should be repeated after increasing the size of the representative calibration spectra to support a proper test-set validation or reducing possible data clustering in the dataset (which can bias leave-one-out validations).

### 4.4.4 PCA and PCR Chemometrics in the Study of Clay Minerals

This section illustrates the use of PCA and PCR to extract and explore the variability present in a dataset consisting of the MIR diffuse reflectance (DRIFT) IR spectra of 32 dioctahedral smectites. This is a well-studied dataset (Gates, 2005) with accompanying chemical and crystallographic data and has been already shown in Figs. 4.2 and 4.3 to demonstrate the effect of

various mathematical pretreatments. The dataset is very limited in size compared to the anticipated dimensionality of the problem (see Section 4.4.3), but it does cover most of the chemistry range of the dioctahedral smectites. As such it will serve to guide the reader through the PCA process from training to validation and the development of PCR predictions.

*PCA calibration*. The set of 32 vector-normalised and baseline-offset MIR data presented in Fig. 4.2A was subjected first to PCA analysis. Employing the Pearson's correlation coefficient,

$$R = \frac{\sum (x - x_i)}{\left(\sum (x - x_i)^2\right)^{\frac{1}{2}}} \tag{4.9}$$

as a measure of variance, the effect of each calculated component in explaining the total variance of the dataset can be deduced in Fig. 4.9A. As explained in Section 4.4.1, components are calculated in order of decreasing explained variance and their incorporation to the fit results in increasing cumulative explained variance. In the vector-normalised dataset, the first component, PC1, accounts for ~55% of the explained variance in the dataset, PC2 accounts for ~45% of the residual variability, and each successive component describes approximately 33% of the successive residual. The effect of derivatization treatments as well as multiplicative scattering corrections on the resulting explained variance of the training set is also shown in Fig. 4.9. Derivatization (Fig. 4.9A) results in an increased contribution by PC1, but successive components have less influence on the explained variance. Conducting MSC or EMSC corrections on the data prior to derivatization has little influence on the resulting explained variance (Fig. 4.9B). Two remarks can be made on the basis of the data in Fig. 4.9: First, as this is a calibration run,

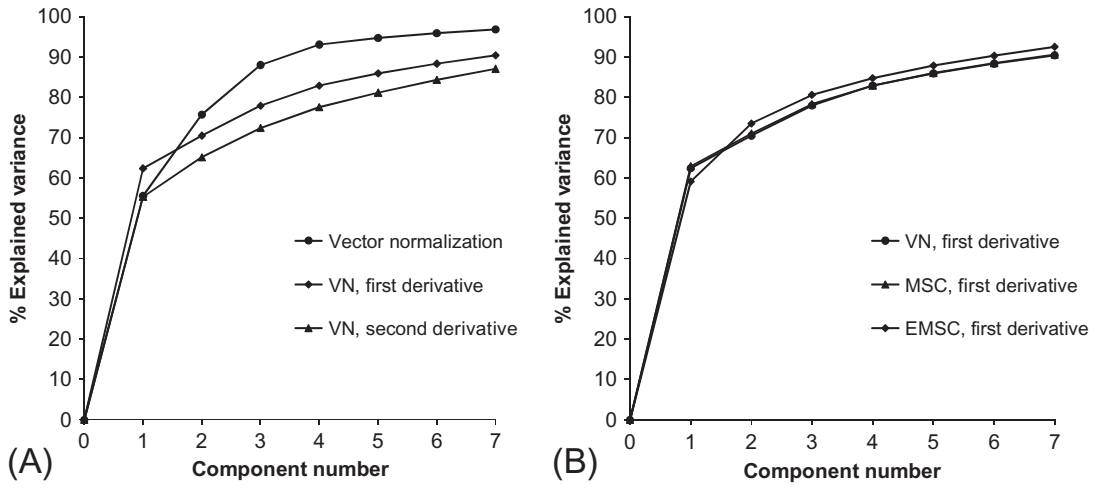



**FIG. 4.9** Cumulative explained variance (based on Pearson's R coefficient) by successive components calculated from PCA on the MIR spectra depicted in Fig. 4.2B. (A) Effect of derivatization on vector-normalised data. (B) Effect of multiplicative scattering corrections (pretreatments) on first derivative data.

the explained variance will be based on increasingly trivial information beyond a certain number of ranks until reaching 100% at 32 principal components (cf. Fig. 4.8). Second, this slowly converging solution is typical for an information-rich system. The description of such systems requires a relatively large number of independent latent variables and should be based on datasets that are significantly larger than the one used in the present example.

Nevertheless, the minimum mathematical preprocessing of the vector-normalised spectral set allows for rationalising the shape of the first principal components (Fig. 4.10) in terms of smectite chemistry. For example, PC1 (Fig. 4.10A) is dominated by positive correlations with the $\nu$(OH) and $\delta$(OH) spectrum of nontronite (e.g. strong positive bands at ~3580, 830 $cm^{-1}$). A sharp positive feature at 1620 $cm^{-1}$ also occurs in PC1, associated with interlayer $H_2O$. Capturing the largest amount of the remaining variability (20% of the total), PC2 (Fig. 4.10B) exhibits an opposite response in the $\nu$(OH) range (positive near 3665, negative near 3570 $cm^{-1}$), as well as a strong negative response associated with ν(SiO) stretch near 1180 and 1060 $cm^{-1}$. Negative features associated with bound water near 3350 $cm^{-1}$ are also present in PC2.

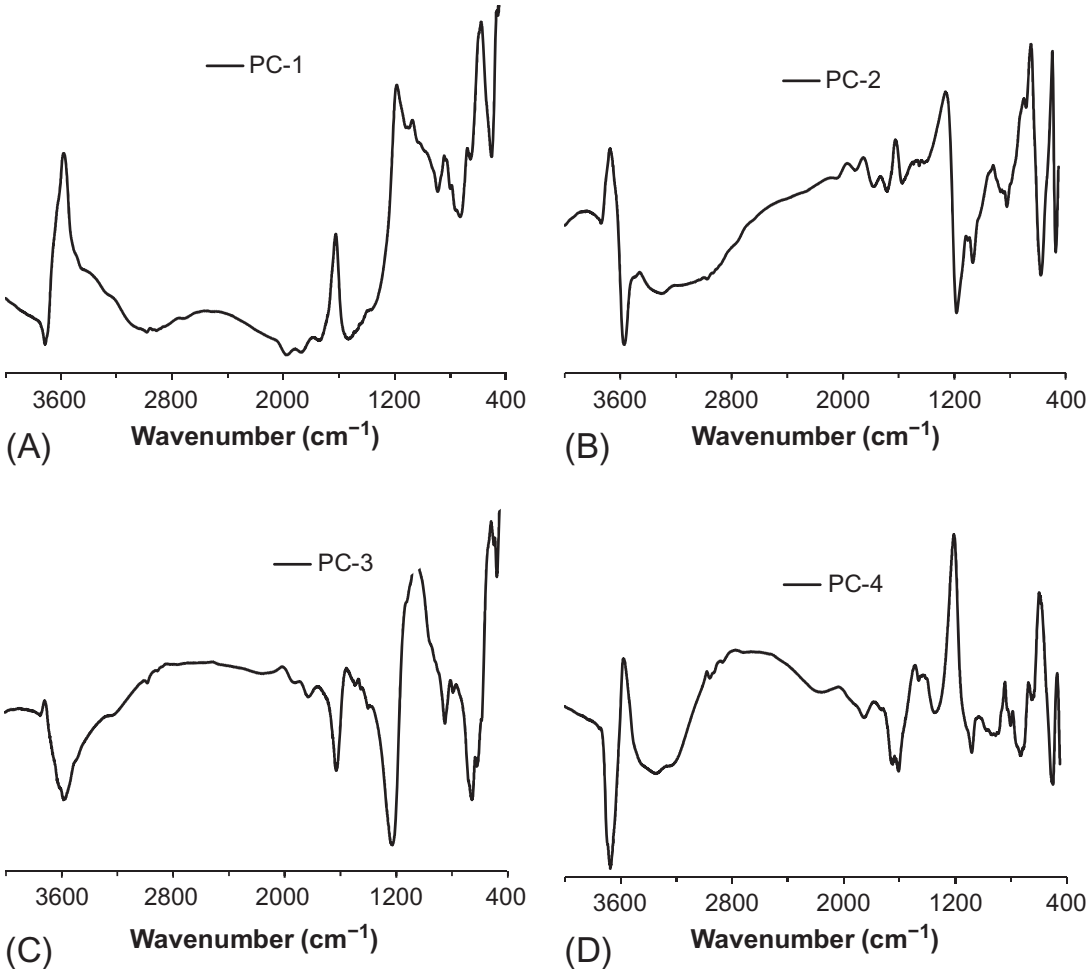



**FIG. 4.10** Spectral loading plots of the first four components (A: PC1; B: PC2; C:PC3 and D: PC4) computed from vector-normalised MIR spectra of 32 dioctahedral smectites. The plots are all differently and arbitrarily scaled.

PC3 (Fig. 4.10C), which captures the next largest proportion (12%) of the remaining variability, indicates a broader negative feature at 3580 $cm^{-1}$ than PC1, as well as strongly negative features at 1620, 1222, and 650 $cm^{-1}$, and a strongly positive feature at 1055 $cm^{-1}$. PC4 (Fig. 4.10D) amounts to only 5% of the captured variability, and appears to convey issues associated with scattering (broad features centred near 2800 $cm^{-1}$).

Score plots (Fig. 4.11A) showcase what physical parameters may be associated with the calculated components. For the 32-sample dataset, nontronites are arrayed negatively with PC1 and both positively and negatively with PC2, indicating that PC1 captures well the variability in Fe content of the samples. This relationship extends to the ferrian smectites (dashed circle), which are positive in PC1 and negative in PC2. Beidellites and montmorillonites are arrayed positively with PC1, but interspersed with each other and mostly positively, in PC2. PC2, which only captures 13% of the variability of the dataset, appears to be associated with the magnesium content of the samples.

The locus of the 32 samples in the score plot of Fig. 4.11A is maintained with small deviations upon performing a leave-two-out internal validation. About one-third of the samples deviate slightly along PC2 and of these, most deviate positively. One sample (circled) is poorly validated along PC2. Such a sample is a candidate outlier and should be checked thoroughly using, for example, a Hotelling statistic (Hotelling, 1931).

The percent explained variance of all the components calculated is shown in Fig. 4.11B, where the validation set returns a lower percentage than

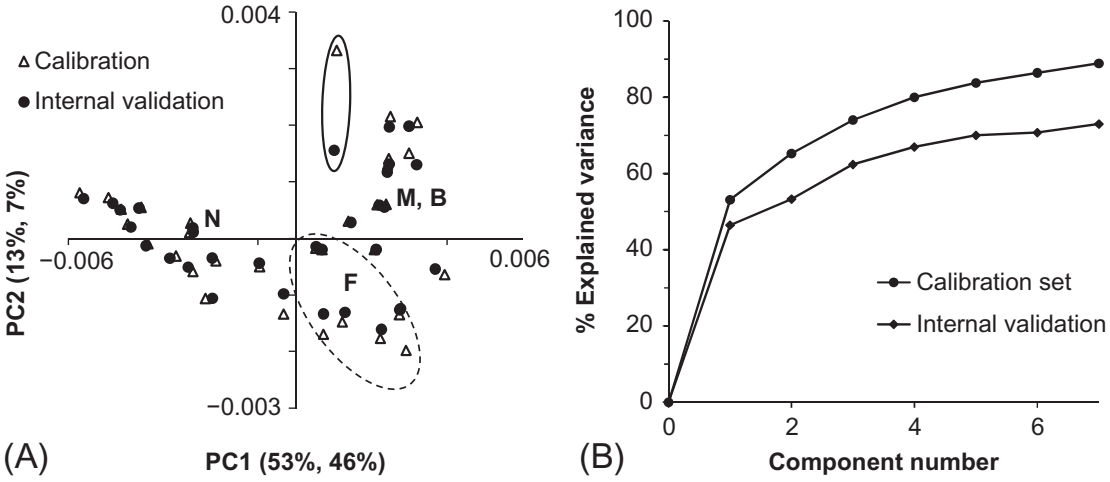



**FIG. 4.11** (A) Scatter plot of the first two principal component scores (which account for 65% of the explained variance) resulting from PCA determined on the 32-sample training set. The internal validation was performed by applying the PCA model iteratively to the data with two randomly selected samples omitted (leave-two-out). The pair of values circled indicates the sample with the strongest deviation of the validation from the calibration. (B) Percentage of explained variance of components (based on Pearson's R) for the 32-sample calibration set and the randomly assigned leave-two-out internal validation. Symbols: *N*, nontronite; *F*, ferruginous smectite; *M*, montmorillonite; *B*, beidellite. The ferruginous smectite grouping is encircled by a dashed line.

the calibration set, without showing a maximum at a certain optimum rank (cf. Fig. 4.8). This indicates that considerable variability remains uncaptured by PCA.

*PCR model validation.* Armed with an example description of the system in terms of its principal components, it is now possible to examine whether the principal components calculated by PCA regress, by least squares, against known chemical or physical properties of the samples. Validations will be presented on the basis of a representative and randomly chosen 8-sample internal test set, calibrated against the remaining 24 spectra. Both calibration and validation sets have identically measured property sets and were subject to the same pretreatments. For the following PCR model development, the PCA components derived from vector-normalised datasets (Fig. 4.10) were used without any other corrections, and applied to the octahedral Fe content given in Gates (2005).

For PCR applied to the octahedral Fe content, PC1 and PC2 component loadings describe the same features of the dataset as PCA (Fig. 4.12), and thus indicate that these components remain the most influential in describing octahedral Fe for this particular dataset. However, the first two components are inverted in importance with respect to PCA (not shown). Among the 24 calibration samples, PC1 accounts for about 11% of the explained variance in the regression on octahedral Fe whereas PC2 accounts for 83% (Fig. 4.13), resulting in a cumulative capture of 94%. Subsequent principal components have no influence on the resulting regressions. Note also that the 8-sample validation set tracks very well the 24-sample calibration set, amounting to 94% cumulative explained variance (16% from PC1 and 78% from PC2).

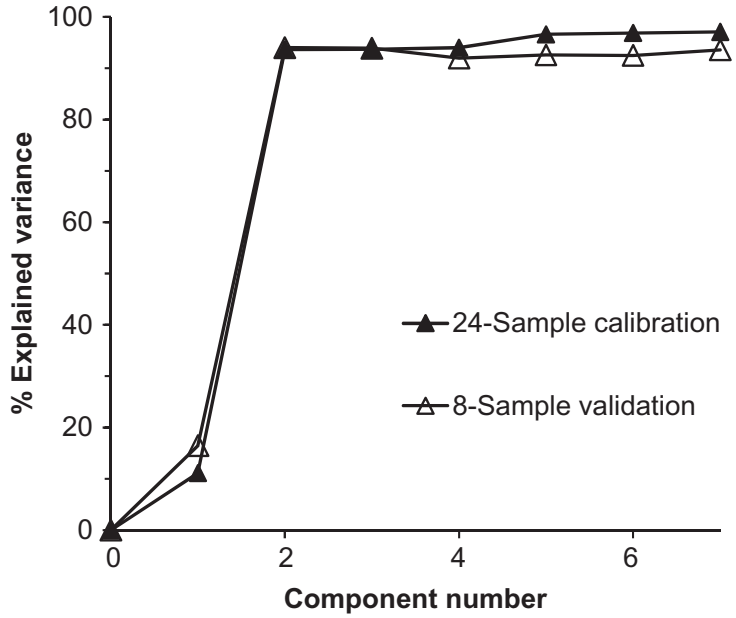



**FIG. 4.12** Percentage of explained variance in the PCR treatment of 32 dioctahedral smectites. The internal validation consisted of a randomly selected set of 8 samples from the original 32-sample dataset used.

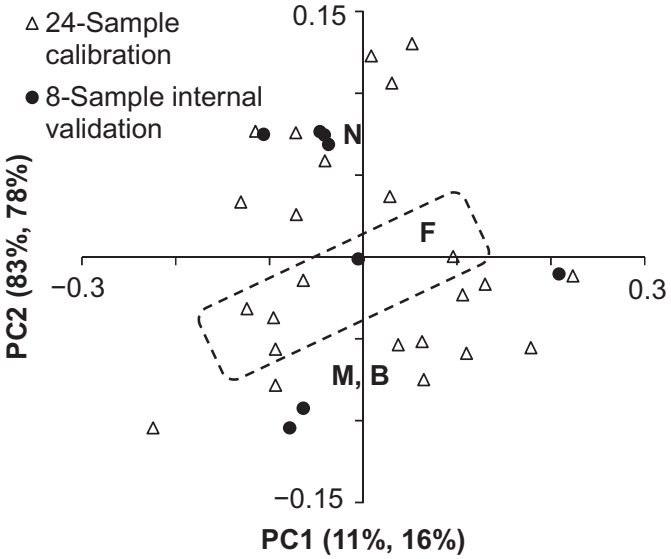



**FIG. 4.13** Scatter plots of the first two principal component scores resulting from PCR on octahedral Fe determined on the vector-normalised, 24-sample training set *(open symbols)*. The internal validation was performed by applying the PCR model to the octahedral Fe contents of eight randomly selected samples comprising the validation set *(filled symbols)*. Percentage values for each PC correspond to, respectively, the PCA calibration and the PCR internal validation. Symbols: *N*, nontronite; *F*, ferruginous smectite; *M*, montmorillonite; *B*, beidellite. The ferruginous smectite grouping is encircled by a dashed line.

A scatter plot for the scores resulting from regression of the first two principal components onto octahedral Fe content (Fig. 4.13) reveals that the samples distribute uniformly along both PC1 and PC2, and do not show the strong distribution along PC1 that was observed previously in Fig. 4.11A. The different clay minerals (nontronite, ferrian (ferruginous) smectite, montmorillonite, and beidellite) are still largely preserved, but with a different orientation in comparison to Fig. 4.11A. In fact, the locus of the groups indicates that PC2 captures most of the passage between montmorillonite-beidellite, ferrian smectite and nontronite, that is, most of the variability in octahedral Fe content. The 8-sample validation set, composed of randomly selected nontronites (4), montmorillonites (2), beidellite (1) and ferrian smectite (1) are well dispersed within the calibration set, indicating that any predictions resulting for this PCR model will be robust.

The 24-sample calibration predictions and the 8-sample validation predictions are depicted in Fig. 4.14 for octahedral Fe content. The prediction of octahedral Fe by PC1 is poor (Fig. 4.14A), which is not surprising given that it accounts for 11% of the variability. Prediction based solely on PC2 highly underestimates octahedral Fe (Fig. 4.14B), but correlates well with the data. The PCR model that combines linearly the orthonormal PC1 and PC2 loadings (Fig. 4.14C) significantly improves the prediction of octahedral Fe, again decreasing significantly the root mean square error (RMSE) and also increasing the resulting coefficient of correlation ($R^2$). As expected, the 8-sample validation set tracks the calibration in the same fashion.

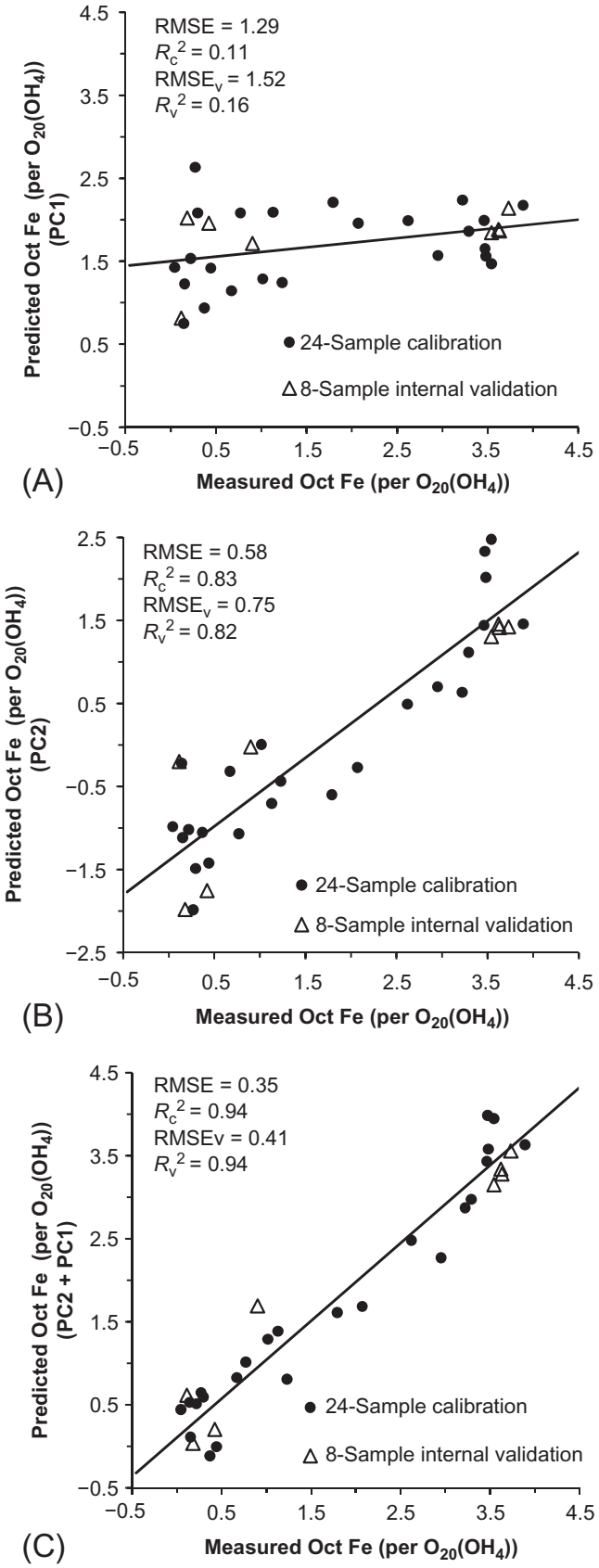



**FIG. 4.14** PCR regression results for (A) PC1, (B) PC2 and (C) a two-component (PC2+PC1) PCR model to predict octahedral Fe content. Root mean square error (RMSE) correlation coefficient ($R^2$) reported is reported for the regression on the calibration and internal validation set for each model.

In summary, the PCA and PCR treatment on the MIR spectra of a series of well-studied dioctahedral smectites indicates the capacity of this multivariate analytical approach to produce valid and useful predictions of their physical or chemical properties. Increasing the number of samples in the calibration set could capture better the latent structure of their spectra over the same, or broader, ranges. It could also enable a more thorough validation, adding to the robustness of the predictive models and offering an appealing alternative for the high-throughput analysis of similar samples.

### 4.4.5 PLS Chemometrics for Clay Mineral Processing Applications

This section highlights the application of PLS chemometrics as a platform technology for quantitative analysis at the clay mineral processing plant. The perspective is both appealing and challenging: Assuming that good-quality spectra can be collected at suitable points of the processing line, these can be used to produce instantaneous predictions on several prevalidated properties of the sample. The latter can then be used for optimising subsequent processing steps or for the quality assurance of the final products. In addition, this approach would create over time a spectral record of all processed materials at the selected sampling points. If properly managed, such detailed records of production can be revisited at later stages for additional data mining regarding the identification of unknown components, for troubleshooting, or for setting up new methodologies.

There are several issues that are critical for the implementation of such projects and require close collaboration and mutual understanding between the spectroscopist and the end user. These issues concern the selection of sampling points, the choice of the spectroscopic technique, the properties to be targeted for fitting among the characterisation data that can be independently available, and the operational maintenance of the application. In addition, they concern the way the spectroscopic chemometric tool will be developed, tested and installed in a 'transparent' way that does not obstruct or delay normal operation. Few of these aspects are detailed in the peer-reviewed literature (but see Goetz et al., 2009; Konrad et al., 2015), because each application is custom made for a specific deposit and plant. Thus this section provides general guidelines and examples rather than detailed accounts of specific applications.

A deposit is mined selectively on the basis of exploratory data, for example on drill-core samples. A typical clay mineral processing plant accepts the mined ore in coarsely crushed rock form, with near-natural moisture content and with variable mineralogical composition. What follows is a series of physical (mixing, drying, milling, beneficiation), and sometimes chemical (ion exchange, calcination) modification steps that are necessary for producing the final product(s). All these steps from ore mining to the final

product(s) eliminate progressively the original variability of the material, resulting in a product with a relatively narrow range of compositions and consistent specified properties. Each of these steps down the process line provides a potential sampling point, with its own requirements and limitations.

The overall geological structure of the deposit may be diverse, but the samples that are isolated from drill cores can be relatively uniform at the medium (100 g–5 kg) scale. In addition they can be available dry after equilibration to ambient conditions and in homogenised powder or bulk rock forms. Such samples are relatively easy to measure by many vibrational techniques, although their potentially large number calls for high-throughput data acquisition (Herrmann et al., 2001; Yang et al., 2001). The opposite holds for the agglomerate rock that is accepted as raw material for further processing, usually after coarse crushing. Critical screening decisions may need to be made at this early stage of production. The material, however, can be very diverse on the medium scale of single lumps. Their surface is created by fracture and cannot be assumed to be identical to the bulk, and their moisture content can vary locally. Such material needs to be sampled and measured in a manner that represents large volumes, of the order of many tons (e.g. Goetz et al., 2009). Subsequent processing steps typically produce finer and more uniform material but may require measurements under dynamically changing conditions, for example, during the evolution of equilibration to ambient reactions and other diffusion-dependent processes. In the final products the properties may require precise measurement over narrow ranges of variability.

Among the various vibrational sampling techniques discussed previously in Chapter 3, diffuse reflectance in the NIR is perhaps the most suitable in fulfilling the majority of the diverse sampling requirements of the processing plant and in providing high-quality and throughput spectra that are suitable for chemometric modelling (see Section 3.3.5). Optical fibres, integrating spheres and external illumination diffuse reflectance probes (see Fig. 3.7) have progressively increasing sampling spot sizes which are useful for averaging the signal from the surface of heterogeneous samples. Fibres and spheres are more suitable for laboratory, off-line applications, whereas external illumination probes allow additionally for the contactless, quasicontinuous, real-time measurement of large or moving samples, for example, during belt transportation (see Fig. 3.8).

Besides noninvasive sampling, NIR spectroscopy has additional advantages that are specific to the study of clay-containing ores. Due to its sensitivity to vibrational modes involving the hydrogen atom, NIR can provide spectra that may be diagnostic for natural samples that contain significant amounts of nonhydrous associated minerals. Further, due to the activity of O—H stretching-bending combination modes, and also due to anharmonicity effects, NIR typically offers better resolution between structural OH and $H_2O$ in the sample than MIR. On the other hand, ATR-MIR is superior to NIR in

modelling the nonhydrous components (Müller et al., 2014), the content of which can be among the critical specifications of the final product. Robust, industrial-grade FT instruments are preferable due to their higher resolution and accuracy, but simpler instruments can be employed if portability and/or the cost associated with the number of sampling points are the dominant considerations (Herrmann et al., 2001; Yang et al., 2001).

The next decision concerns the choice of properties that need to be validated for the chemometric modelling of the vibrational spectra (see Section 4.4.2). A good (and often the only possible) starting point is the modelling of the properties that are already implemented in the plant and used for decision making. With some precautions, these established quality control operations can be used for providing matching calibration spectra and independent property data. The advantage is that sampling and on-site analytics are already available and based on standardised procedures. The usual drawback is that, due to time and cost considerations, the on-site materials characterization can be coarse in terms of fundamental understanding. It may, therefore, be lacking the specificity in terms of structure and composition that would be desirable for correlation with vibrational spectroscopic data. As an example, industrial determinations of smectite content may be based on titrations with a cationic dye (Lagaly, 1981). Such methods can provide the average CEC or layer charge but are unsuitable for distinguishing the type and content of the various smectite species in the sample. Besides that, they can be sensitive to the aggregation of the dye molecule (Budják et al., 2002). As another example, several problems can be anticipated when the exchange of the interlayer cations of the parent material is required in order to develop performance. First, the effect of the nature of interlayer cations in the NIR (or MIR) spectra of clay minerals is indirect and weak. Second, the ion-exchange routines applied on site may be lacking the accuracy of those performed in an academic laboratory. More seriously, the chemometric model may be required to predict the anticipated performance of the raw material, prior to the processing that is needed for the development of this performance.

Despite these concerns, properly validated PLS chemometrics can perform very well at the clay mineral processing plant. The predictions shown in Fig. 4.15 are selected among the results of two independent industrial feasibility projects on dioctahedral bentonite and palygorskite-smectite systems. The spectra have been collected by FT NIR instruments in the diffuse reflectance mode with an external illumination unit and a powder probe, respectively. In both cases, the samples were in a $<250$ μm homogenised powder form, dried at ambient conditions without further treatments. They were selected from company depositories of reference production and deposit exploration samples, respectively, and their properties were determined by in-house standardised procedures applied after a common preparation step. PLS chemometrics were developed with the QUANT2 package of the OPUS software by Bruker Optics. In most cases, spectral preprocessing involved Savitzky–Golay

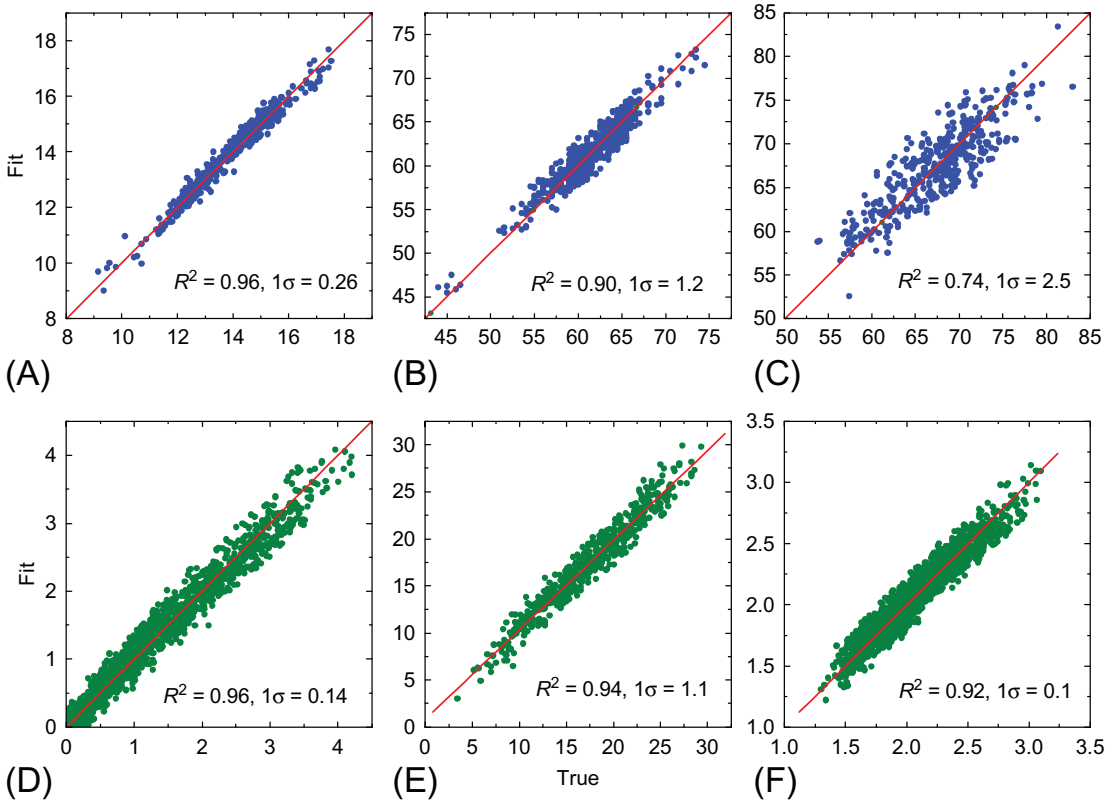



**FIG. 4.15** True-fit performance of NIR-based PLS chemometrics on industrial clay mineral sample sets obtained by cross-validation leaving out 10% of the original samples. The upper series (A–C) is from a dioctahedral bentonite system represented here by ~400 samples, and the lower series (D–F) is from a palygorskite-smectite system (~1200 samples). Property data cannot be disclosed, but some details are provided in the text.

second derivatives, for filtering out variable baselines or broad vibrational bands, and enhancing the resolution of sharp features (see Section 4.3.2). The methods were typically cross-validated by applying a leaving-10%-out method after randomising the order of the spectra in $\boldsymbol{X}$ to remove possible time-series systematics. The properties shown in Fig. 4.15 (arbitrary units) include moisture content (A) estimated gravimetically, 'smectite content' (B, E) estimated by titration methods, palygorskite content (D) estimated from the intensity of the $d_{110}$ XRD reflection at ~10.4 Å as in Gionis et al. (2007), and two proprietary sets of data on anticipated performance (C, F). At this level of development, PLS chemometrics can complement or substitute seamlessly the existing laboratory quality control procedures on site. In addition, a number of knowledge-based parameters, typically based on peak positions or relative intensities (e.g. Post and Noble, 1993; Yang et al., 2001; Chryssikos et al., 2009; Stathopoulou et al., 2011) can be extracted from the spectra and included in the characterisation report of each sample as additional proxies of crystal structure and composition.

As discussed in Section 4.4.2, the successful development of any chemometric correlation method relies on the similarity between the calibration samples and the anticipated unknowns. This is because predictions should be

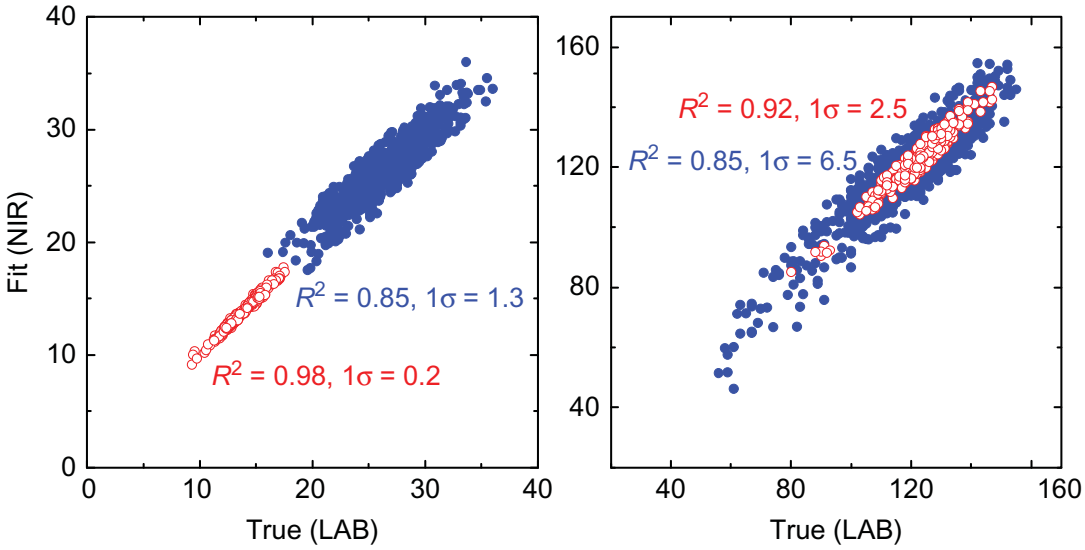



**FIG. 4.16** Comparison of cross-validated PLS predictions at two different sampling points of the same plant during the same time period. Solid symbols (blue in online version) correspond to real-time NIR measurements of the wet, coarse material over the conveyor (~800 samples). Open symbols (red in online version) represent laboratory spectral measurements on <250 μm powders dried at ambient conditions (~300 samples) performed with the same diffuse reflectance head. The left panel is on moisture content validated against the results of an industrial gravimetric method; the right panel represents smectite content proxied by a titration (arbitrary units).

based on interpolation rather than extrapolation, but also because diffuse reflectance techniques depend strongly on the particle size of the sample. For these reasons, the chemometric monitoring of the same material at a different sampling point along the process pipeline requires in most cases independent calibration and validation. In Fig. 4.16 are compared two such calibrations for moisture and smectite content (arbitrary units) based on the incoming wet, coarse material from the conveyor belt against laboratory measurements on powdered samples equilibrated at ambient conditions (as in Fig. 4.15). The two sets of data are obtained from the same production line over the same time period of several months. The data should, therefore, represent the same distributions of the essential properties of the material (except moisture content and particle size), which is in agreement with the results shown in Fig. 4.16. On the other hand, the RMSE of the predictions based on wet, coarse calibrants is clearly inferior to that based on dry powders. The reason for this must be sought in the way calibration and property data ($\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively, see Section 4.4.2) are obtained. For practical and safety reasons, the characterisation of the coarse material may require taking a sample from the conveyor, sealing it to avoid loss of natural moisture, and measuring it at a later time in the laboratory by a sampling system that simulates spectral acquisition over the belt (e.g. a rotating platform as in Goetz et al., 2009). Measuring the independent properties of the same sample by the usual standard methods requires sample drying and homogenisation. Obtaining such $\boldsymbol{X}$ and $\boldsymbol{Y}$ data on samples that exhibit local variations in

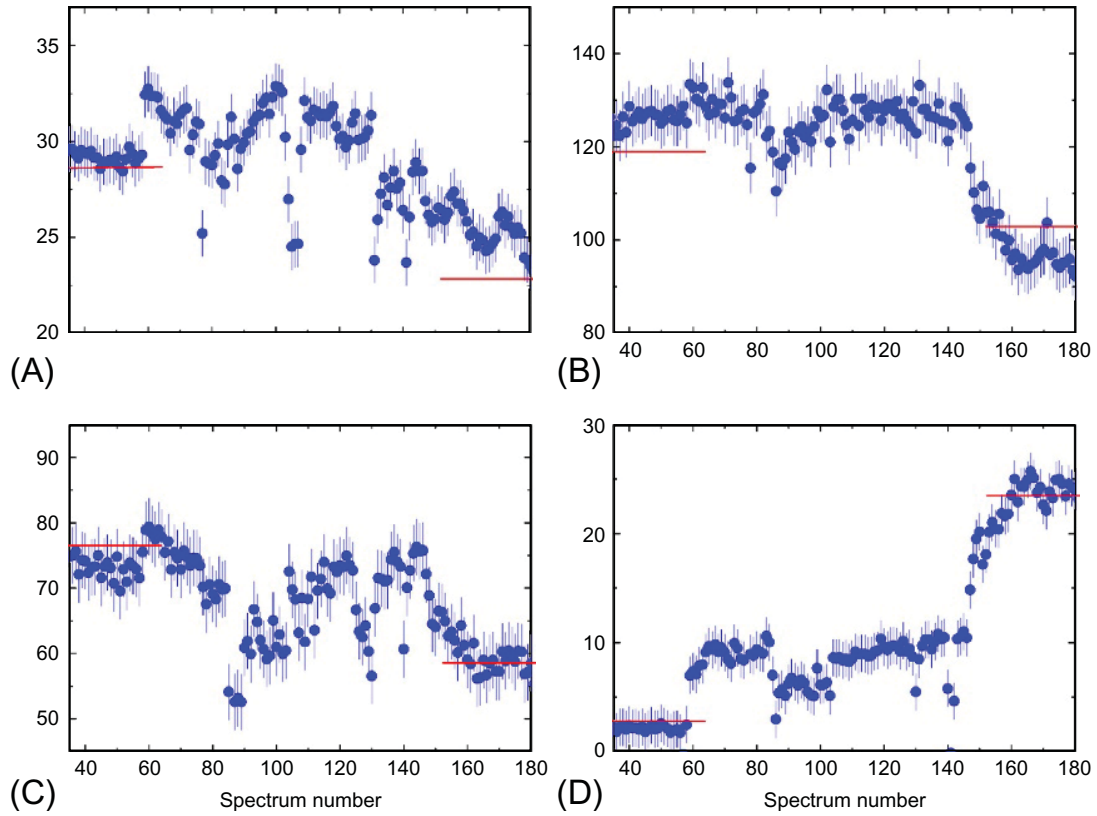



**FIG. 4.17** Real-time PLS chemometric predictions of four different properties during the monitoring of wet, coarse clay-mineral ore during transportation on an open-air conveyor. The spectral time series, similar to that shown in Chapter 3 (Figure 3.8) was collected by an external illumination diffuse reflectance probe. The predictions (arbitrary units) concerned the moisture (A), smectite (B), interlayer (C) and one common impurity (D) contents of a 2-min average sample, over a period of ~5 h of continuous operation. Lines on the left and right side of each panel (red in online version) correspond to the results of independent laboratory analysis from the routine periodic sampling of the transported material.

composition is by no means trivial. Calibration is prone to higher sampling errors in both the spectral and property measurements, as well as in matching exactly the samples of the two sets.

Overcoming these obstacles leads to chemometric tools that can be applied on-line to characterise the incoming raw material over, for example, a conveyor belt in real time. This approach is illustrated in Fig. 4.17. It is based on a representative time series of diffused reflectance NIR data (similar to that in Chapter 3, Fig. 3.8) and a set of suitable chemometric validations as in Fig. 4.16. In this manner, the time dependence of several sample properties can be obtained simultaneously. These properties can feed-back decisions on the potential use of the incoming materials and, if applicable, on the parameters of subsequent treatments. The data are equispaced in time but can be converted to tonnage by considering the time-dependent loading capacity of the conveyor. The time resolution of the graphs in Fig. 4.17 is sufficiently sensitive to sudden changes, but average predictions can be obtained over longer periods, if needed. The performance of each prediction algorithm can be

accessed in real time by the quality figures of the PLS fit, and can also be externally evaluated (and validated) by existing protocols for the regular sampling and characterisation of the raw material (horizontal ticks in each panel of Fig. 4.17).

It needs to be emphasised that any quality control process based on chemometric correlations requires regular maintenance. Modern spectrometers have internal validation standards that can be used regularly for detecting and correcting drifts with time. Systematic changes regarding the state of the material at the sampling points need to be accounted for. More importantly, it cannot be assumed that the unknown incoming samples will remain within the calibrated range of latent variables as mining advances with time. The problem is typically addressed by the periodic update of the sample and property datasets and their use for the revalidation of the methods. It has been proposed that the necessary adjustments and chemometric updates can be performed in an unsupervised manner (Li et al., 2000; Benndorf, 2015). Including the exploration samples as a separate stage of chemometric assessment can provide an early external validation warning about anticipated material changes in production. The need for method maintenance implies that the in-house conventional quality control procedures should not be abandoned. Chemometrics should be employed for increasing the number of characterised samples by orders of magnitude, also for providing an independent assessment of the material, but not as a full substitute of conventional characterisation.

## 4.5 CONCLUDING REMARKS

When applying mathematical treatments to interrogate the data, it is important to keep in mind that statistics do not generate information. Instead, they just help extract relevant information from signals that contain noninformative content. The less noninformative content, the greater is the accuracy of information extraction. Emphasis should therefore be placed on the quality of the raw data *per se*, and this implies decisions regarding the most suitable sampling technique, the spectral range, the optical and digital resolution, the acquisition time, the elimination of (or compensation for) spectral artifacts. With appropriate adaptations, the same holds for the various types of independent data that are considered for chemometric correlation with the spectral data.

Deciding what part(s) of the spectral information is(are) informative is question dependent: It is determined by what information is sought from the spectra. Vibrational spectroscopy is extremely rich in information concerning the composition, structure and bonding of clay minerals and related samples. Yet one should be prepared to encounter situations where the information content of the spectra that is relevant to the specific question of interest is below the detection limit of a particular (or any) vibrational spectroscopic technique. Further, the threshold between significance and triviality can be vague and should only be approached statistically by validation procedures.

Vibrational spectra are fingerprints of samples (Farmer and Russell, 1964, 1967). From the perspective of multivariate analysis, spectra (or samples) are pixels of a more general picture and acquire most of their significance as specific parts of that general picture. Thus a single spectrum of a specific montmorillonite has little to offer if observed outside the context of the IR spectra of other montmorillonites or smectites in general. The situation calls for a large number of observations that are needed to create the general picture.

Multivariate analysis is a toolbox designed for identifying and analyzing systematics within large sets of observations, or among large sets of observations obtained by different 'sensors'. It is in their latter capacity that multivariate analysis and especially chemometrics can bridge between disciplines and open the way towards scientific discovery.

Yet it should always be remembered that the human brain has amazing multivariate analytic capabilities in handling and rationalising observations. The visual inspection of large numbers of spectra on screen, subjected to various pretreatments (derivatives, normalisations, etc.) has always been rewarding in identifying important trends and subtle systematics.

A final remark is deserved regarding multivariate analysis and chemometrics: Their performance can be properly judged during application, but its reporting in the literature is often received with scepticism. This is because the models are built from datasets that may not be available for inspection, even less for review, but also because the physical meaning of the latent variables is rarely straightforward. Further, it is easy to use commercial software and modest computing resources for producing solutions of no value. Poorly validated 'results' appear more impressive the more mathematically trivial they are. For these reasons, multivariate analysis and chemometric modelling should not be developed, nor reported as 'black box' applications. They should, instead, be discussed in their real context, as tools for classifying, identifying and predicting the behaviour of individual objects that are part of a complex, poorly understood cosmos.

DEVELOPMENTS IN CLAY SCIENCE 8
SERIES EDITOR: F. BERGAYA

# Infrared and Raman Spectroscopies of Clay Minerals

Edited by
W. P. Gates, J.T. Kloprogge,
J. Madejová, and F. Bergaya

An up-to-date systematic review of spectroscopic theory, methods, and techniques for the study of clay minerals by infrared and Raman spectroscopies.

- Includes a systematic review of spectroscopic methods
- Covers the theory of infrared and Raman spectroscopies and instrumentation
- Features a series of chapters, each covering either a particular technique or an application

*Infrared and Raman Spectroscopies of Clay Minerals*, Volume 8 in the Developments in Clay Science series, is an up-to-date overview of the spectroscopic techniques used in the study of clay minerals. The methods include infrared spectroscopy, covering near-IR (NIR), mid-IR (MIR), far-IR (FIR) and IR emission spectroscopy (IES), as well as FT-Raman spectroscopy and Raman microscopy. This book complements the succinct introductions to these methods described in the original *Handbook of Clay Science (Volumes 1, 1st Edition and 5B, 2nd Edition)*, offering greater depth and featuring the most important literature since the development and application of these techniques to clay science. No other book covers such a wide variety of vibrational spectroscopic techniques in a single volume for clay and soil scientists.

**W. P. Gates**, Institute for Frontier Materials, Deakin University, Victoria, Australia

**J. T. Kloprogge**, Department of Chemistry, College of Arts and Sciences, University of the Philippines Visayas, Miagao, Philippines

**J. Madejová**, Slovak Academy of Sciences, Institute of Inorganic Chemistry, Bratislava, Slovakia

**F. Bergaya**, CNRS, Interfaces, Confinement, Matériaux et Nanostructures (ICMN) Orléans, France

GEOLOGY

ISBN 978-0-08-100355-8

9 780081 003558

ELSEVIER

elsevier.com/books-and-journals