Infrared Reflection-Absorption Spectrometry and Chemometrics for Quantitative Analysis of Trace Pharmaceuticals on Surfaces

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Chemistry

at the

University of Canterbury

Christchurch
New Zealand

Benjamin B. Perston
September 2006

Abstract

Cleaning validation, in which cleaned surfaces are analysed for residual material, is an important process in pharmaceutical manufacturing and research facilities. Current procedures usually consist of either swab or rinse-water sampling followed by analysis of the samples. The analysis step is typically either rapid but unselective (conductivity, pH, total organic carbon, etc.), or selective but time-consuming (HPLC). This thesis describes the development of an *in situ* surface-spectroscopic analysis that removes the need for swab sampling and is both rapid and selective. This method has the potential to complement existing analyses to increase the efficiency of cleaning-validation protocols.

The spectrometric system consists of a Fourier-transform infrared (FTIR) spectrometer coupled to a fibre-optic grazing-angle reflectance probe, and allows the measurement of infrared reflection-absorbance spectra (IRRAS) from flat surfaces in $\sim$10 s. Multivariate chemometric methods, such as partial least squares (PLS) regression, are used to exploit the high information content of infrared spectra to obtain selective analyses without physical separation of the analyte or analytes from whatever interfering species may be present.

Multivariate chemometric models require considerably more effort for calibration and validation than do traditional univariate techniques. This thesis details suitable methods for preparing calibration standards by aerosol deposition, optimising and validating the model by cross- and test-set validation, and estimating the uncertainty by resampling and formula-based approaches.

Successful calibration models were demonstrated for residues of acetaminophen, a model active pharmaceutical ingredient (API), on glass surfaces. The root-mean-square error of prediction (RMSEP) was $\sim$0.07 $\mu g\,cm^{-2}$. Simultaneous calibration for acetaminophen and aspirin, another API, gave a similar RMSEP of 0.06 $\mu g\,cm^{-2}$ for both compounds, demonstrating the selectivity of the method. These values correspond to detection limits of $\sim$0.2 $\mu g\,cm^{-2}$, well below the accepted visual detection limit of $\sim$1–4 $\mu g\,cm^{-2}$.

The sensitivity of the method with a stainless steel substrate was found to depend strongly on the surface finish, with highly polished surfaces giving more intense IRRAS. RMSEP values of 0.04–0.05 $\mu g\,cm^{-2}$ were obtained for acetaminophen on stainless steel with three different finishes. For this system, severe nonlinearity was encountered for loadings $\gtrsim$1.0 $\mu g\,cm^{-2}$.

From the results presented in this thesis, it is clear that IRRAS has potential utility in cleaning validation as a complement to traditional techniques.

Acknowledgments

First, I would like to thank my supervisors, Bryce Williamson and Peter Harland, both for their invaluable guidance and encouragement during the course of this research and for their friendship.

My time at Canterbury has been a very enjoyable one, and this is due in no small part to the people I have been fortunate enough to work with. Thanks to the past and present denizens of (or frequent visitors to) Room 732: Michelle Hamilton, Josh Lehr and James Bull for the stimulating (relevant or otherwise) discussions; and Brett, Dave and Sam for teaching me (by example, of course) how to lose gracefully at cards.

Thanks to Danny Leonard in the Mechanical Workshop for building the variable-angle reflectance probe, and to Mary Thomson and Peter Melling at Remspec for their assistance during the early phase of this work.

Financial assistance in the form of a University of Canterbury Doctoral Scholarship and a Vice Chancellors' Committee of New Zealand William Georgetti Scholarship is gratefully acknowledged, as is the provision of the spectrometer, software and travel funds by Remspec.

I owe a lot to my family for the support of all kinds (gestational, habitational, gustatory…) they have given me over the last quarter-century; in particular, I would like to thank my parents for their tolerance during the final stages of thesis preparation.

And finally, I would like to thank Katy for her patience!

Contents

List of abbreviations

ADC Analog-to-digital converter

API Active pharmaceutical ingredient

a. u. Arbitrary units

CLS Classical least squares

EPDM Ethylene propylene diene monomer

ESA Electrostatic self-assembly

FTIR Fourier transform infrared

HPLC High performance liquid chromatography

ILS Inverse least squares

IRRAS Infrared reflection-absorption spectrometry

LOD, LOQ Limit of detection, quantitation

MCT Mercury cadmium telluride

MLR Multiple linear regression

(R)MSE(C, P, CV) (Root) mean square error (of calibration, prediction, cross-validation)

NIPALS Nonlinear iterative partial least squares

NP, OE, WV Chemometric pre-treatments: no pre-treatment, offset elimination, water vapour spectrum subtraction (respectively)

OLS Ordinary least squares

P, S, R Polished, smooth, rough (surface finishes in Chapter 8)

PAH Poly(allylamine hydrochloride)

PCA Principal component analysis

PCR Principal component regression

PDF Pseudo-degrees of freedom

PLS(R) Partial least squares (regression)

PMMA Poly(methylmethacrylate)

PPCR Polynomial principal component regression

PSS Poly(styrene sulfonate)

PTFE Poly(tetrafluoroethene)

Q-PLS Quadratic PLS

RA Reflection-absorbance

RAL Acceptable residual limit

RMS Root mean square

RSD Relative standard deviation

SDS Sodium dodecyl sulfate

SEM Scanning electron microscope

SNR Signal-to-noise ratio

SVD Singular value decomposition

ZPD Point of zero displacement

Chapter 1

Introduction

This thesis concerns the use of grazing angle infrared reflection-absorption spectrometry (IRRAS) as a quantitative analytical technique. IRRAS is a specular reflection method in which the spectrum of a substrate coated with a film of analyte is compared to the spectrum of a clean substrate. If the experimental conditions are chosen appropriately, IRRAS can be a very sensitive method, enabling detection and characterisation of sub-monolayer films [1].

When combined with the sampling flexibility afforded by fibre optics and the analytical power of modern chemometric regression methods, IRRAS has promise as a complement to existing techniques in determination of trace surface contamination in applications such as pharmaceutical cleaning validation. This introduction briefly discusses current cleaning-validation methodology and outlines the role that IRRAS could play.

1.1 Pharmaceutical cleaning validation

Cleaning validation has received a lot of attention in the pharmaceutical community since the publication, in 1993, of the USFDA's "Guide to Inspections: Validation of Cleaning Processes" [2]. This document outlines the requirements that US pharmaceutical firms must meet with respect to validating their cleaning processes. Cleaning validation is a complex topic, particularly where one piece of equipment is used to process several different drugs or drug ingredients. This section is intended to give an overview of the validation process to provide context for the analytical methods that underpin it. In addition to the FDA document, articles that provide additional background include those by Zeller [3], Jenkins and Vanderwielen [4], and Amer and Deshmane [5], as well as the guidelines published by the Active Pharmaceutical Ingredients Committee [6].

1.1.1 Determining acceptable residual limits

The purpose of cleaning validation is to ensure that contamination of pharmaceutical products due to residues left on the equipment by the previous process is minimised. However, the sensitivity of modern analytical methods is such that extremely small amounts of material can be detected: in many cases far less than could lead to any significant contamination of the final product. For this reason, it is not required that equipment be shown to be absolutely clean within the detection limit of the analysis. Rather, an acceptable residual limit (RAL) is determined, based on various criteria, and the analysis result must be less than this limit (with a specified level of certainty) for the equipment to be deemed clean.

The methodology introduced by Fourman and Mullen [7] for determining RALs is best explained by example. Compound X is produced in a reactor with surface area A, which is then cleaned, leaving behind a residual loading (mass per unit area) L_{X}. The same reactor is then used to produce a mass m_{Y} of compound Y. The per-dose contamination (with mass dimensions) of Y with X is given by

$$C_{\mathrm{XY}} = L_{\mathrm{X}} \times A \frac{d_{\mathrm{Y}}}{m_{\mathrm{Y}}} \tag{1.1}$$

where d_{Y} is the dose mass of Y. The criteria Fourman and Mullen give for C_{XY} are that a) no more than 0.001 of a dose of X will appear in the maximum daily dose of Y; and b) no more than 10 ppm of X will appear in Y. Thus, $C_{\mathrm{XY}}^{\mathrm{max}}$ is given by the smaller of $0.001 d_{\mathrm{X}}$ and $d_{\mathrm{Y}}/10^5$. This value in turn implies the RAL for X, $L_{\mathrm{X}}^{\mathrm{max}}$. Their final restriction is that no residue may be visible on the equipment surface, even though the presence of visible contamination does not guarantee that the calculated RAL would be exceeded. Fourman and Mullen state that this last criterion places an upper limit on L for most materials of pharmaceutical interest of about $4\,\mu\mathrm{g\,cm^{-2}}$. Some variations on this general procedure are presented by LeBlanc [8, 9].

1.1.2 Sampling

Once the RAL has been established, an analytical method must be found that has a low enough limit of quantitation (LOQ) to reliably measure loadings around $L_{\mathrm{X}}^{\mathrm{max}}$. These methods are either direct or indirect. The FDA guidelines prefer methods that sample the surface directly: these include swab sampling and visual inspection, as well as spectroscopic methods (although the latter two are not necessarily endorsed by the FDA). Indirect methods include rinse-water sampling and placebo sampling. Placebo sampling is mostly applicable to finished products. After the equipment has been cleaned, a second batch of the product is made, minus the active ingredient. The placebo product is then analysed for the

active ingredient. This method is discouraged, because the placebo may dilute the active ingredient to the point where it is difficult to detect [10]. Rinse sampling involves analysing the rinse water from the final cleaning step; details are given by LeBlanc [11]. In general, rinse sampling is suitable for monitoring the cleaning process (if the rinse water is dirty, the equipment must also be dirty), but is not regarded as suitable for cleaning validation except in certain situations (since clean rinse water does not imply clean equipment).

The most popular sampling method is swab-sampling [12]. This procedure consists of swabbing a region of the equipment surface, extracting the residues from the swab with solvent, then analysing the solution by some suitable method, commonly high-performance liquid chromatography (HPLC). To ensure reproducibility, detailed instructions regarding swab materials, swabbing technique and extraction conditions must be documented [13]. Swabbing cannot, in general, be relied upon to retrieve all material from the sampled area, even if the swab is moistened with solvent. Consequently, it is also necessary to determine the average recovery percentage.

Recently, some authors have suggested that visual inspection can be used as a cleaning validation method, provided that the RAL is greater than the lowest visibly detectable level. LeBlanc [14] argues that, provided care is taken in determining the visible detection limit (which depends on the analyte, the surface and the viewing conditions), visual inspection is a valid analytical method. Forsyth et al. tested the method in both manufacturing [15] and research [16] contexts and concluded that it is a reasonable approach. In Ref. 16, they list visible residue limits for a number of compounds. These limits vary from $\sim 0.1\ \mu g\,cm^{-2}$ to $\sim 3\ \mu g\,cm^{-2}$. They emphasise the speed and convenience of visual inspection, which takes minutes, compared to the hours or days for analysis of many samples by HPLC. However, there has been no endorsement of this method yet by regulatory bodies; in fact, it is specifically excluded by the FDA guidelines except for validation between two batches of the same product.

1.1.3 Analysis

Once a sample has been obtained by swabbing or rinse-water sampling, it can be analysed by one or several of a variety of methods. Kaiser and Minowitz [17] compared the commonly used analytical methods. In general, analytical techniques are evaluated on the basis of their sensitivity, selectivity and cost, the last of which includes a time component. The selectivity of a method is its ability to discriminate for the analyte in the presence of other species. Broadly, methods can be characterised as being selective or non-selective; the latter group includes techniques such as total organic carbon (TOC) analysis, conductivity and pH measurements, and will not be considered here.

Most selective methods employed in cleaning validation involve a chromatographic separation,

with HPLC being a particularly powerful and versatile example. Appropriate choice of the mobile and stationary phases can usually ensure separation of the analytes, and a variety of detectors are available. Compounds without a UV chromophore, such as many surfactants, may be detected with an evaporative light scattering detector (ELSD); however, HPLC excels when used in conjunction with a UV diode array detector (DAD) for the analysis of UV-absorbing species. In addition to having excellent sensitivity, this detector allows measurement of the UV absorbance spectrum (rather than the absorbance at a single wavelength), which provides additional security against interferences. HPLC with a UV detector or DAD appears to be the standard method for analysing active pharmaceutical ingredient (API) residues, and the sensitivity seems more than adequate. Mirza et al. [18] give a conservative estimate of the LOQ of 0.2 $\mu g\,cm^{-2}$ for meclizine hydrochloride, for which the RAL they determined was 0.5 $\mu g\,cm^{-2}$. Klinkenberg et al. [19] reported an analysis of amlodipine in which the LOQ of the HPLC method is 0.08 $\mu g\,mL^{-1}$. With the 20 cm^2 swabbing area and 10 mL extraction volume they used, this limit corresponds to 0.04 $\mu g\,cm^{-2}$ (the RAL they determined was 0.76 $\mu g\,cm^{-2}$). More examples are cited in Ref. 17.

The main problem with HPLC is that it is not rapid: while the time required for an analysis depends strongly on the conditions, times of tens of minutes to over an hour are not uncommon. Since a single cleaning validation may require measuring several samples for each of several pieces of equipment, this is a significant burden.

A second problem is that all swab-based methods are inherently somewhat indirect since the material must be removed from the surface. Not only does this further increase the time required for cleaning validation, it also introduces the possibility of interferences from the swab material; further, the recovery percentage may be subject to considerable variability. An *in situ* surface-spectroscopic method, on the other hand, can be truly direct.

1.2 Infrared reflection-absorption spectrometry

There has been some application of IRRAS in cleaning validation, particularly in the determination of cleaning compounds, which can be difficult to analyse by HPLC [20, 21, 22]. However, these studies were limited by the sampling restrictions of in-compartment reflectance accessories. In 1993, LeBlanc stated [21]

> Such surface methods as Fourier transform infrared spectroscopy (FTIR) are possible if the equipment can be disassembled, if removable coupons can be evaluated in the system, or if flexible probes can access internal surfaces.

The maturation of infrared fibre-optic technology in the decade since then has led to vast improvements in the performance of the "flexible probes" he mentions. The fibre-optic system used in this thesis (described in detail in Chapter 4) is a prototype version of a hand-held reflectance probe coupled to an FTIR spectrometer by a ~1 m optical cable (cable lengths up to a few metres are possible). The probe can be freely moved about, and sampling requires merely holding the probe on the surface to be analysed for ~10 s. The area sampled by the present system is ~20 cm^2, but larger or smaller sampling regions are possible. Provided that IRRAS can be shown to be sufficiently selective, sensitive and reliable, it has a few obvious advantages over the prevailing swab/HPLC technique:

- IRRAS is a truly direct method: sampling and analysis are combined into a single step.
- Rapidity: a spectrum can be measured and analysed in ~30 s or less.
- Applicability to a wide range of analytes: almost all organic molecules (including surfactants) have bands in the infrared with high absorptivity.

Of course, there are also some disadvantages:

- Sampling is much less flexible than swabs: at present, only reasonably large, flat areas can be analysed, and they must be fairly accessible. Cable lengths of more than a few metres are impractical with current technology.
- If unanticipated interfering species that were not included in the calibration are present, they are very likely to confound the analysis. This is also a risk with HPLC, but somewhat less so since the interfering compound must have both a similar retention time and spectroscopic overlap with the analyte. However, the diagnosis of this situation is straightforward with chemometric tools, so it would be obvious that something had gone awry.
- IRRAS, being a new method, has not been tested by regulatory bodies.

The first of the disadvantages above is significant enough that IRRAS is unlikely ever to displace the swab/HPLC method. A complementary role for it can easily be imagined, however: after cleaning, the initial analysis would be by IRRAS. If this returned a "clean" verdict, swab samples would be taken and sent for analysis; if the IRRAS verdict were "dirty", cleaning could be repeated without the need for time-consuming HPLC measurements. Production of the next product could begin before the HPLC results were returned, depending on the confidence in the IRRAS method and the relative costs of equipment down-time and (in the case that the HPLC analysis disagreed with the IRRAS) wasted materials.

The work carried out in this thesis was initiated and supported by Remspec Corporation (Sturbridge, Massachusetts; website: http://www.remspec.com). Remspec has supplied a range of mid-infrared spectroscopic systems using fibre optics since 1993. The work described in this thesis is part of a collaboration with the University of Canterbury to investigate the feasibility of *in situ* fibre-optic IRRAS for pharmaceutical cleaning validation

1.3 Structure of this thesis

Since IRRAS combines elements of both reflection and transmission spectroscopies, the theoretical treatment is somewhat complicated. Chapter 2 describes the optical theory of isotropic, layered media and presents the equations required for simulation studies.

To utilise the selectivity inherent in the multivariate IRRAS measurements, a "chemometric" approach to the analysis of the spectra is required. Chapter 3 provides details of the regression methods and error analysis techniques used in this thesis.

The IRRAS system is described in Chapter 4, where some background to Fourier transform spectrometry and infrared fibre optics is also given. Methods for preparing calibration standards are also characterised and compared.

The main results are presented in Chapters 5–8. Chapter 5 is a detailed exploration of the thin-film model described in Chapter 2, aimed at finding the ideal instrument parameters (incidence angle and state of polarisation of the infrared radiation) for a variety of substrate and film-thickness combinations. This chapter also details the construction of a fibre-optic IRRAS probe with variable incidence angle, and compares some experimental results with the theory. Chapter 6 concerns the applicability of IRRAS to a glass surface contaminated with a single API, demonstrating the feasibility of the method with dielectric substrates. Chapter 7 extends these results to mixtures of two APIs on glass, demonstrating the selectivity possible with IRRAS. Chapter 8 investigates the effect of the roughness of a stainless-steel surface.

Chapter 2

Thin-film optics

2.1 Introduction

Infrared reflection-absorption spectroscopy (IRRAS) is a method for studying films on solid or liquid substrates. The spectrum is obtained by comparing the specular reflectance spectrum of a coated substrate (R) with that of a clean one (R_0). Reflection-absorbance[1] (RA) is defined as

$$\mathrm{RA} = -\log_{10}\frac{\mathrm{R}}{\mathrm{R_0}} \tag{2.1}$$

and is, in general, a complicated function of the film and substrate optical constants, the film thickness, and the incidence angle and polarisation of the radiation. As described in this chapter, for certain systems and under certain conditions, the RA behaves very similarly to the transmission-mode absorbance.

Early experimental work, such as that by Francis and Ellison [23] and by Greenler [24, 25, 26, 27], was limited to very thin films on metallic substrates: in this case, the RA resembles an absorption spectrum and is proportional to the film thickness. The proliferation of Fourier transform spectrometers and advances in infrared detector technology in the 1970s and 1980s allowed the IRRAS technique to be applied to films on reflective dielectrics, such as silicon [28], and even to monolayer films on weakly reflective substrates, such as water [29, 30].

The theoretical treatment for isotropic films was presented by Greenler in 1966 [24] and, more thoroughly, by Hansen in 1968 [31]. Since monolayer films often have a preferred molecular orienta-

[1] Some authors use the differential reflectance, given by

$$\frac{R_0 - R}{R_0} = 1 - \frac{R}{R_0}$$

rather than the RA. These quantities are related: the differential reflectance is proportional to the first term in the power series expansion of the reflection-absorbance.

tion, optically anisotropic films are also important. A variety of approximations based on the isotropic theory have been used [32], and the rigorous extension of the isotropic theory was first presented by Yeh [33]. Parikh and Allara [34] described a practical method for using Yeh's theory in conjunction with structural models and reflection-absorption spectroscopy to determine the molecular orientation in monolayer films. In this thesis, however, only the isotropic theory is necessary.

The first section of this chapter introduces the physical model that forms the basis for the calculations and provides some necessary background material. The second section presents the theory for calculating the reflectance of a stratified medium, while the third describes how to calculate the electric field intensities at any point in the system. The fourth section is a brief numerical exploration of the behaviour of some relevant systems. In the final section, the consequences of these results for the applications presented elsewhere in this thesis are discussed briefly.

2.1.1 Physical model

The physical model that forms the basis for the calculations is illustrated in Figure 2.1. From a transparent medium with refractive index N_0, light is incident at an angle θ_0 to the surface normal on a stack of q films supported on a substrate. Each film is parallel-sided and optically isotropic (extension to anisotropic films is possible [34, 35], but not required for this work). Under these assumptions, film j is characterised by its refractive index $N_j = n_j + ik_j$ (see Section 2.1.4 below) and its thickness d_j. The films are considered to be thin in comparison with the radiation wavelength (meaning that they show interference effects), while the substrate on which the stack of films rests is taken to be semi-infinite, meaning that no light is returned from the far side of the substrate. While the incident medium is taken to be transparent, the films and the substrate may be absorbing. There are q films, so there are $q + 2$ media (including the incident medium and the substrate) and $q + 1$ interfaces. The incident medium is denoted by a subscript 0 (N_0). The film indices increase from the incident side to the substrate, which is denoted by a subscript s (N_s), corresponding to the index $q + 1$. The interfaces are labelled from 1 to $q + 1$.

A Cartesian coordinate system is defined as shown in Figure 2.1a. The z axis is normal to the surface and the plane $z = 0$ is the interface between the incident medium and the first film; the substrate is at a depth of $z = \sum_{j=1}^{q} d_j$. The plane of incidence is defined by the propagation vector of the incident radiation and the surface normal, and x is chosen to lie in this plane. Since the radiation is represented by infinite plane waves, the origin can be placed at any point on the xy plane.

The incident light is plane-polarised: if results are derived for two polarisations then any polarisation can be obtained by superposition. The most convenient polarisations to use are called p-

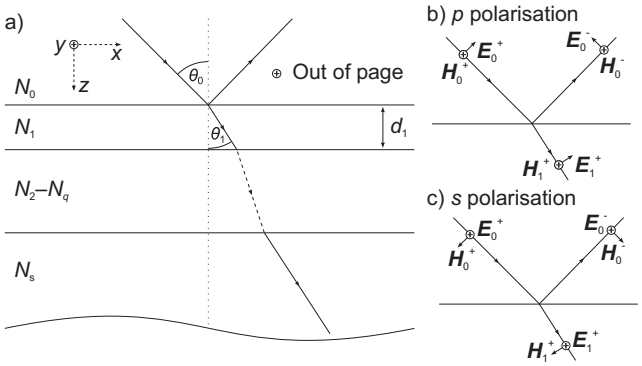

Figure 2.1: Model for calculating thin-film reflectance. (a) Depiction of a stack of q films on a substrate. The Cartesian coordinate frame is offset for clarity; the plane $z = 0$ is the interface between the incident medium and the first film. (b) Electric and magnetic field orientations for p-polarised light. (c) Electric and magnetic field orientations for s-polarised light.

polarisation, in which the electric vector is parallel to the plane of incidence (the xz plane); and s-polarisation, in which the electric vector is perpendicular to the plane of incidence. Sometimes p-polarised light is called transverse magnetic (TM), while s-polarised light is called transverse electric (TE) (see Figures 2.1c and d).

The quantities measured in an IRRAS experiment are the reflectances of the bare and coated substrates superimposed on the instrument response. IRRAS is calculated from these single-beam spectra in the same way as an absorbance spectrum (Equation 2.1) so that the instrument response is cancelled. The subsequent sections derive expressions for the reflectance at a simple boundary (R_0) and from a film stack supported on a substrate (R). The case of a single film can be treated more simply, but for completeness the method for a stack of any number of films is also given.

2.1.2 Plane harmonic waves

The motion of light through an uncharged, current-free medium with refractive index $N = n - ik$ can be described by a plane harmonic wave solution to Maxwell's equations. The complex amplitude of the electric field vector at any point in such a wave can be written [36]

$$E = \mathcal{E} \exp i\left((2\pi N/\lambda)\,\mathbf{s}\cdot\mathbf{r} - \omega t\right) \tag{2.2}$$

where $\mathcal{E}$ is a scalar, λ is the vacuum wavelength, ω is the circular frequency, $\mathbf{s}$ is a unit vector giving the direction of propagation and $\mathbf{r} = (x, y, z)$ is the Cartesian coordinate vector. A relative phase, ϕ, can

be included in the expression by allowing the amplitude term $\mathcal{E}$ to become complex:

$$E = \mathcal{E}\exp i\left((2\pi N/\lambda)\,\mathbf{s}\cdot\mathbf{r} - \omega t + \phi\right) \tag{2.3}$$

$$= \mathcal{E}\exp i\phi\exp i\left((2\pi N/\lambda)\,\mathbf{s}\cdot\mathbf{r} - \omega t\right) \tag{2.4}$$

$$= \hat{\mathcal{E}}\exp i\left((2\pi N/\lambda)\,\mathbf{s}\cdot\mathbf{r} - \omega t\right) \tag{2.5}$$

Substituting $N = n + ik$ into Equation 2.2 and taking the wave travelling in the positive z direction,

$$E = \mathcal{E}\exp i\left((n+ik)(2\pi/\lambda)z - \omega t\right) \tag{2.6}$$

$$= \mathcal{E}\exp\left(-2\pi kz/\lambda\right)\exp i\left(2\pi nz/\lambda - \omega t\right) \tag{2.7}$$

it can be seen that the imaginary part of the refractive index leads to an exponential decay in the amplitude with increasing z.

2.1.3 Snell's law

Snell's law (which can be derived from Fermat's least time principle [37]) relates the angles of incidence and refraction when light is incident on a boundary between two transparent media with different refractive indices.

$$N_1\sin(\theta_1) = N_2\sin(\theta_2) \tag{2.8}$$

When the second medium is absorbing, N_2 is complex and no longer simply represents the ratio of the speed of light in a vacuum to its speed in the medium. Snell's law still holds, however, if the angle θ_2 is allowed to be complex, in which case it no longer represents the physical angle between the direction of propagation and the surface normal. Snell's law will be used later in this chapter to calculate cosines of the complex propagation angles.

2.1.4 Optical constants in the infrared

The optical constants of a medium are the real and imaginary parts of the complex refractive index, $N = n + ik$. An introduction to optical constants concentrating on infrared radiation is given by Bertie [38]. Of relevance to this work are the optical constants of metals, solid inorganic materials such as glass, and solid organic materials (the former two as substrates and the latter as analytes/film materials). All media are assumed to be isotropic.

The behaviour of the optical constants through an organic vibrational resonance is illustrated in Figure 2.2 for liquid benzene, where the absorption band is due to a ring stretch mode [39]. The k

curve resembles an absorption band, while n has a derivative shape: from high wavenumber towards the band centre n decreases, then across the band centre it increases sharply before decreasing again. This behaviour of the refractive index is called anomalous dispersion ("normal" dispersion is the slow decrease in n with decreasing wavenumber seen in regions of the spectrum far from any absorption bands).

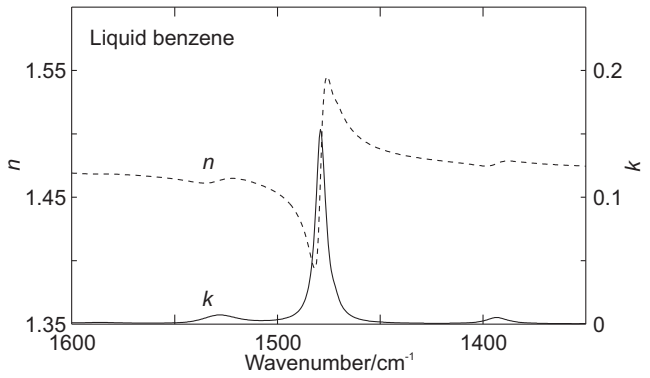

Figure 2.2: Real and imaginary parts of the refractive index of liquid benzene at 25 °C, from Ref. 39.

Inorganic materials often have very strong absorption bands, as illustrated in Figure 2.3 by the Si–O stretch in silica glass [40]. The stronger the absorption, the greater the dispersion of the refractive index.

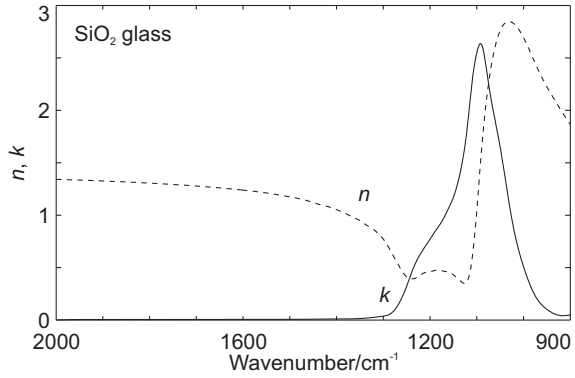

Figure 2.3: Real and imaginary parts of the refractive index of silica glass, from Ref. 40.

The optical constants can be used to calculate any other intensity quantity. By inspection of Equation 2.7 and using the relation that the intensity of radiation is proportional to the square of the amplitude, the intensity of radiation that has passed through a distance l of an absorbing material is given by

$$I = I_0 \exp(-4\pi k \bar{\nu} l) \tag{2.9}$$

where $\bar{\nu} = 1/\lambda$ is the wavenumber and I_0 is the initial intensity of the radiation. This leads to the definition of the linear absorption coefficient $\alpha = 4\pi k\bar{\nu}$, which relates to the Napierian absorbance A_e:

$$A_e = \alpha l = \ln(I_0/I) \tag{2.10}$$

The decadic molar absorption coefficient E_m is defined by

$$E_m = \frac{\alpha}{c \ln 10} \tag{2.11}$$

where c is the molar concentration. The decadic absorbance (usually just called the absorbance) is given by

$$A_{10} = lcE_m = \log_{10}(I_0/I) \tag{2.12}$$

2.2 Reflection and transmission

2.2.1 The simple boundary

The Fresnel equations give the amplitude ratios (Fresnel coefficients) for reflection and transmission at an interface. They are derived by solving Maxwell's equations given the boundary conditions that the tangential (x and y) components of the electric and magnetic fields must be continuous across the boundary [37]. When light is incident from medium 0 on medium 1 (as in Figure 2.1b–c),

$$r_{p1} = \frac{E_0^-}{E_0^+} = \frac{N_1\cos\theta_0 - N_0\cos\theta_1}{N_1\cos\theta_0 + N_0\cos\theta_1} \tag{2.13}$$

$$t_{p1} = \frac{E_1^+}{E_0^+} = \frac{2N_0\cos\theta_0}{N_1\cos\theta_0 + N_0\cos\theta_1} \tag{2.14}$$

$$r_{s1} = \frac{E_0^-}{E_0^+} = \frac{N_0\cos\theta_0 - N_1\cos\theta_1}{N_0\cos\theta_0 + N_1\cos\theta_1} \tag{2.15}$$

$$t_{s1} = \frac{E_1^+}{E_0^+} = \frac{2N_0\cos\theta_0}{N_0\cos\theta_0 + N_1\cos\theta_1} \tag{2.16}$$

The subscript on the electric field amplitude E denotes the medium, while a superscript plus sign denotes a wave travelling in a positive z direction and a superscript minus sign denotes a wave travelling in a minus z direction. Thus, E_0^+ is the amplitude of the incident wave, E_1^+ is for the transmitted wave and E_0^- is for the reflected wave. If the refractive indices are real, then the Fresnel coefficients will also be real.[2] If either N_0 or N_1 is complex then the Fresnel coefficients will be complex, representing

[2] Unless $n_1 > n_2$ and $\theta_0 > \sin^{-1}(n_1/n_2)$; these are the conditions for total internal reflection, in which no light is transmitted and r becomes complex (see Section 4.2).

changes in both the amplitude and phase of the wave.

The Fresnel coefficients are related to the reflectance and transmittance as follows:

$$R = r^2 \tag{2.17}$$

$$T = \frac{N_1 \cos\theta_1}{N_0 \cos\theta_0} t^2 \tag{2.18}$$

The reflectance and transmittance always add to unity even if N_1 is complex, since the boundary has no thickness and cannot absorb energy.

When light is incident on the boundary from the opposite direction, several useful identities can be derived [36] from Equations 2.13–2.16:

$$r' = -r \tag{2.19}$$

$$t' = t \tag{2.20}$$

$$tt' = 1 - r^2 \tag{2.21}$$

where a prime symbol indicates that the light is incident from medium 1 rather than medium 0.

2.2.2 Reflectance of a single film supported on a substrate

When there is only a single film, an expression for the reflectance can be derived [36] by adding all the multiply reflected rays, as illustrated in Figure 2.4, and allowing them to interfere with one another.

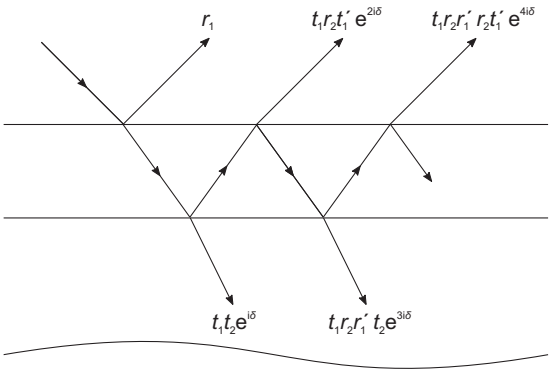

Figure 2.4: Multiple-reflection model for the reflectance of a single thin film on a substrate, after Ref. 36. The overall Fresnel reflection coefficient is the sum of the geometric series of terms due to a given number of internal reflections, a few of which are listed at the top of the figure.

The Fresnel coefficients for the air/film/substrate system, r_{tot} and t_{tot}, are calculated in terms of the Fresnel coefficients for the two boundaries and the phase change, δ_1, that occurs on traversal of the

film. The phase change is given by

$$\delta_1 = \frac{2\pi}{\lambda} N_1 d_1 \cos\theta_1 \tag{2.22}$$

If N_1 is complex, δ_1 will also be complex, with the imaginary part representing absorption by the film. To calculate $\cos\theta_1$ (which is also needed for the calculation of the Fresnel coefficients for the second interface), Snell's law can be used, as shown below in Equation 2.25.

The amplitude of the reflected wave is the sum of the amplitudes of the upwards-going waves. These waves have undergone $0, 1, 2, \ldots$ reflections from the substrate, and sum of their amplitudes is a geometric series:

$$\begin{aligned} r_{\text{tot}} &= r_1 + t_1 t_1' r_2 e^{2i\delta_1} - t_1 t_1' r_1 r_2^2 e^{4i\delta_1} + \ldots \\ &= r_1 + \frac{t_1 t_1' r_2 e^{2i\delta_1}}{1 + r_1 r_2 e^{2i\delta_1}} \\ &= \frac{r_1 + r_2 e^{2i\delta_1}}{1 + r_1 r_2 e^{2i\delta_1}} \end{aligned} \tag{2.23}$$

where the last equality uses the identity from Equation 2.21. This equation is valid for either polarisation. Additionally, the transmitted amplitude is

$$t_{\text{tot}} = \frac{t_1 t_2 e^{i\delta_1}}{1 + r_1 r + 2e^{2i\delta_1}} \tag{2.24}$$

The reflectance and transmittance can be calculated as described below (Equations 2.34–2.37).

2.2.3 Matrix method for multiple films

When there is more than one film, the summation method described above becomes impractical. In the matrix method [31, 36, 41, 42], each film is represented by a characteristic matrix that relates the tangential fields at the boundaries on either side of the film. The matrix representing the complete system relates the tangential fields in the incident medium at the first interface to those in the substrate at the final interface and is obtained by multiplying the individual matrices. A more detailed derivation is given by Macleod [43]. Here, the relevant results from Hansen's treatment are presented.

The notation is simplified by introducing the angle-dependent refractive index term $\eta = N\cos\theta$.

This quantity can be calculated for film j from Snell's law:

$$\begin{aligned} N_0 \sin\theta_0 &= N_j \sin\theta_j \\ N_0^2 \sin^2\theta_0 &= N_j^2(1-\cos^2\theta_j) \\ N_j \cos\theta_j &= (N_j^2 - N_0^2 \sin^2\theta_0)^{1/2} \end{aligned} \tag{2.25}$$

Which root to take depends on the phase conventions in use. If the time dependence of the field is $\exp(-i\omega t)$ and the imaginary part of the refractive index is positive, then both the real and imaginary parts of $N_j \cos\theta_j$ must be positive [31].

The characteristic matrix for layer j is given by

$$\mathbf{M}_j^p = \begin{bmatrix} \cos\delta_j & -\frac{i}{\eta_j}\sin\delta_j \\ -i\eta_j \sin\delta_j & \cos\delta_j \end{bmatrix} \tag{2.26}$$

for p-polarisation and by

$$\mathbf{M}_j^s = \begin{bmatrix} \cos\delta_j & -\frac{iN_j^2}{\eta_j}\sin\delta_j \\ -\frac{i\eta_j}{N_j^2}\sin\delta_j & \cos\delta_j \end{bmatrix} \tag{2.27}$$

for s-polarisation.

The characteristic matrix for the system, $\mathbf{M}_{\text{tot}}$, is the product of the characteristic matrices for all the interfaces. This relates the tangential field amplitudes at the first boundary with those at the final boundary.

$$\begin{bmatrix} U_1 \\ V_1 \end{bmatrix} = \prod_{j=1}^{q} \mathbf{M}_j \begin{bmatrix} U_s \\ V_s \end{bmatrix} \tag{2.28}$$

For p-polarised light, $U = H_y$ and $V = E_x$, while for s-polarised light, $U = E_y$ and $V = H_x$ (see Figure 2.1). From Equation 2.28 the Fresnel coefficients for the system can be derived:

$$r_s = \frac{E_{y,1}^-}{E_{y,1}^+} = \frac{\left(m_{11}^s + m_{12}^s\eta_s\right)\eta_0 - \left(m_{21}^s + m_{22}^s\eta_s\right)}{\left(m_{11}^s + m_{12}^s\eta_s\right)\eta_0 - \left(m_{21}^s + m_{22}^s\eta_s\right)} \tag{2.29}$$

$$r_p = \frac{H_{y,1}^-}{H_{y,1}^+} = \frac{\left(m_{11}^p + m_{12}^p\eta_s/N_s^2\right)\eta_0/N_0^2 - \left(m_{21}^p + m_{22}^p\eta_s/N_s^2\right)}{\left(m_{11}^p + m_{12}^p\eta_s/N_s^2\right)\eta_0/N_0^2 - \left(m_{21}^p + m_{22}^p\eta_s/N_s^2\right)} \tag{2.30}$$

$$t_{Hp} = \frac{H_{y,N}^+}{H_{y,1}^+} = \frac{2\eta_1/N_1^2}{\left(m_{11}^p + m_{12}^p\eta_s/N_s^2\right)\eta_0/N_0^2 - \left(m_{21}^p + m_{22}^p\eta_s/N_s^2\right)} \tag{2.31}$$

$$t_{Ep} = \frac{N_0}{N_s} t_{Hp} \tag{2.32}$$

$$t_{Es} = \frac{E^+_{y,N}}{E^+_{y,1}} = \frac{2\eta_0}{\left(m^s_{11} + m^s_{12}\eta_s\right)\eta_0 - \left(m^s_{21} + m^s_{22}\eta_s\right)} \tag{2.33}$$

where m_{ij} is the ijth element of $\mathbf{M}_{\text{tot}} = \prod_{j=1}^{q} \mathbf{M}_j$.

From the Fresnel coefficients, the reflectance and transmittance are given by

$$R_s = |r_s|^2 \tag{2.34}$$

$$R_p = \left|r_p\right|^2 \tag{2.35}$$

$$T_s = \frac{\Re\left(\eta_s\right)}{\eta_0} |t_{Es}|^2 \tag{2.36}$$

$$T_p = \frac{\Re\left(\eta_s / N_s^2\right)}{\eta_0 / N_0^2} \left|t_{Hp}\right|^2 \tag{2.37}$$

where $\Re$ is the real-part operator.

2.3 Electric field calculations

The rate of energy absorption is proportional to the mean square electric field $\langle E^2 \rangle$ [44]. From quantities determined previously, it is possible to calculate $\langle E^2 \rangle$, relative to the incident mean square amplitude, at any point along the z axis. Equations are derived by Hansen for both the single-film and the many-film case, but only the single-film equations are reproduced here. Separate equations are provided for each field component (y for s-polarisation; x and z for p-polarisation) and each of the three media.

For the incident medium ($z < 0$),

$$\langle E^2_{y0} \rangle = 1 + R_s + 2R_s^{1/2} \cos\left[\phi^r_s - (4\pi/\lambda)\,\eta_0 z\right] \tag{2.38}$$

$$\langle E^2_{x0} \rangle = \cos^2\theta_0 \left[1 + R_p - 2R_p^{1/2} \cos\left(\phi^r_p - (4\pi/\lambda)\,\eta_0 z\right)\right] \tag{2.39}$$

$$\langle E^2_{z0} \rangle = \sin^2\theta_0 \left[1 + R_p + 2R_p^{1/2} \cos\left(\phi^r_p - (4\pi/\lambda)\,\eta_0 z\right)\right] \tag{2.40}$$

where ϕ is the phase change on reflection or transmission and is given by the argument of the corresponding complex Fresnel coefficient,

$$\phi^r = \tan^{-1} \frac{\Im r}{\Re r} \tag{2.41}$$

where $\Im$ is the imaginary-part operator.

Within the film itself ($0 < z < d$), direct expressions for the mean square amplitude are unwieldy, and it is easier to calculate the complex field amplitude numerically and then calculate its mean square

value. The fields are given by

$$E_{y1} = \exp i\left(\frac{2\pi}{\lambda} n_0 \sin\theta_0 x - \omega t\right)\left[(1+r_s)\cos\left(\frac{2\pi\eta_1 z}{\lambda}\right) + i\frac{\eta_0}{\eta_1}(1-r_s)\sin\left(\frac{2\pi\eta_1 z}{\lambda}\right)\right] \tag{2.42}$$

$$E_{x1} = \exp i\left(\frac{2\pi}{\lambda} N_1 \sin\theta_1 x - \omega t\right)\left[\cos\theta_0\left(1-r_p\right)\cos\left(\frac{2\pi\eta_1 z}{\lambda}\right) + i\frac{\eta_1}{N_1^2} n_0\left(1+r_p\right)\sin\left(\frac{2\pi\eta_1 z}{\lambda}\right)\right] \tag{2.43}$$

$$E_{z1} = -\exp i\left(\frac{2\pi}{\lambda} N_1 \sin\theta_1 x - \omega t\right)\left[\frac{N_1 \sin\theta_1}{N_1^2} n_0\left(1+r_p\right)\cos\left(\frac{2\pi\eta_1 z}{\lambda}\right) + \frac{N_1 \sin\theta_1}{\eta_1}\cos\theta_0\left(1-r_p\right)\sin\left(\frac{2\pi\eta_1 z}{\lambda}\right)\right] \tag{2.44}$$

(Remember that $N_1 \sin\theta_1 = n_0 \sin\theta_0$.)

The mean square amplitude is calculated by setting $x = 0$ and taking the time average:

$$\langle E^2 \rangle = \frac{1}{2} E E^* \tag{2.45}$$

Finally, for the substrate ($z > d$),

$$\langle E_{ys}^2 \rangle = |t_{Es}|^2 \exp\left[-\frac{4\pi}{\lambda}\Im\eta_s (z-d)\right] \tag{2.46}$$

$$\langle E_{xs}^2 \rangle = \left|\frac{\eta_s}{N_s} t_{Ep}\right|^2 \exp\left[-\frac{4\pi}{\lambda}\Im\eta_s (z-d)\right] \tag{2.47}$$

$$\langle E_{zs}^2 \rangle = \left|\frac{n_0 \sin\theta_0}{N_s} t_{Ep}\right|^2 \exp\left[-\frac{4\pi}{\lambda}\Im\eta_s (z-d)\right] \tag{2.48}$$

2.4 Numerical exploration of the thin film model

The aim of this section is to investigate the general features of IRRAS of organic compounds on metallic and dielectric substrates. Liquid benzene is chosen as a model film material, since high-quality optical constant spectra in the mid-infrared are available [39]. Aluminium [45] and silica glass [40] are chosen as substrates.

2.4.1 Mean square electric field amplitudes

The electric field amplitudes in the incident medium (Equations 2.38–2.40) are determined by the reflectance and the phase change on reflection. If the reflectance is zero, then $\langle E_{y0}^2 \rangle = 1$, the intensity of the incident radiation. As the reflectance increases, a standing wave pattern develops due to interference between the incident and reflected waves. The phase of the standing wave (that is, whether there is a

node or an antinode or some intermediate value at the interface) depends on the phase change, ϕ^r, on reflection (see Figure 2.5). If ϕ^r is near zero, the interference at $z = 0$ will be constructive and there will be an antinode at the interface. If it is near 180°, destructive interference will result in a node at the interface. The behaviour of the x and z components is similar, but a phase change of $\phi^r_p = 180°$ is required for an antinode in $\langle E^2_{x0} \rangle$ at $z = 0$. The x and z components also include a geometric factor, since the intensity of a p-polarised wave must be divided between its x and z components. At normal incidence the z component vanishes, and at grazing incidence the x component vanishes.

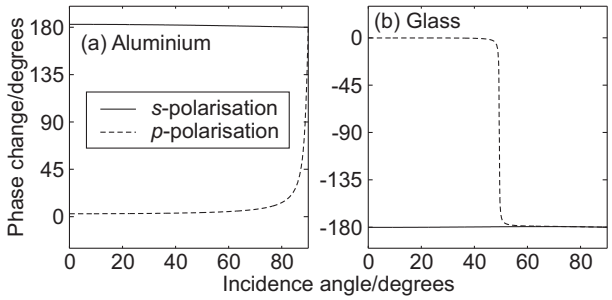

Figure 2.5: Phase change on reflection as a function of incidence angle for a 10 nm film of benzene on (a) aluminium and (b) glass at $\bar{\nu} = 1478\,\mathrm{cm}^{-1}$. The optical constants are $n_0 = 1$, $N_1 = 1.50 + 0.15i$, (a) $N_s = 13 + 64i$ (b) $N_s = 1.16 + 0.008$.

In the substrate, there is no reflected wave, so the mean square amplitude is constant unless the substrate is absorbing, in which case it falls off exponentially.

If the film is very thin, the mean square field amplitude will be essentially constant across it and will depend on its optical constants [46]. In this case, the film has very little effect on the fields in either of the adjacent phases.

A plot of the mean square field amplitudes as a function of z (Figure 2.6) illustrates a number of important phenomena. Firstly, in both cases, the x component is very small. This is because θ_0 is near grazing. For the metallic substrate, the high reflectance and phase change near 180° for s-polarisation results in a node at $z = 0$ for $\langle E^2_{y0} \rangle$. The phase change near 0° for p-polarisation leads (approximately) to a node for the x component and an antinode for the z component. This makes it clear why s-polarisation is ineffective for IRRAS with a metallic substrate, and also leads to the "surface selection rule" [1]: vibrational transition dipole moments oriented parallel to the surface cannot absorb light. When an adsorbed species has a preferred orientation, the relative intensities of the bands in its spectrum are influenced by the orientation of the corresponding dipoles.

For the glass substrate, the phase change is near −180° for both polarisations, so there are nodes for $\langle E^2_{y0} \rangle$ and $\langle E^2_{z0} \rangle$ at $z = 0$, and an antinode for $\langle E^2_{x0} \rangle$. However, the low reflectivity compared to the metal ensures that the amplitudes are still appreciable for all components. An interesting observation

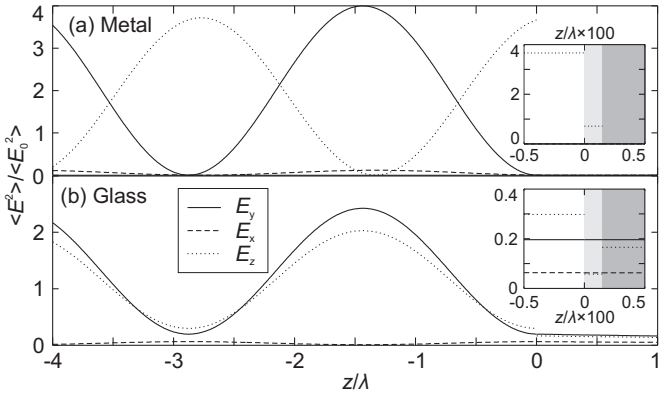

Figure 2.6: Mean square electric field amplitudes as a function of distance from the incident medium/film interface for a 10 nm film of benzene on (a) aluminium and (b) glass. The incidence angle is $\theta_0 = 80°$ and the wavenumber is $\bar{\nu} = 1478\,\mathrm{cm}^{-1}$. The optical constants are $n_0 = 1$, $N_1 = 1.50 + 0.15i$, (a) $N_s = 13 + 64i$ (b) $N_s = 1.16 + 0.008i$. The inset plots expand the abscissa around $z = 0$; note the discontinuities in the z component of the field.

is that while the z component is strongest in the incident phase, the y component is actually stronger in the film. This indicates that, as mentioned by Mielczarski [46], it is important to consider the field inside the film rather than just in the incident medium. Based on the field amplitudes, the s-polarised IRRAS should be more intense for this incidence angle.

The dependence of the electric fields on the incidence angle is illustrated in Figure 2.7. For the metallic substrate, only the z component has appreciable intensity: $\langle E_z^2\rangle$ increases from zero at $\theta_0 = 0°$ (at which angle the propagation vector of the wave is parallel to z so there can be no z component of the electric field) to a maximum at $\theta_0 \approx 80°$ before decreasing sharply to zero at $\theta_0 = 90°$.

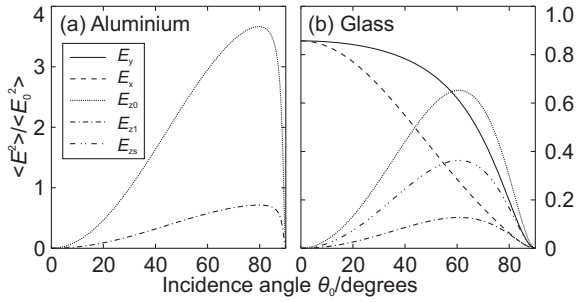

Figure 2.7: Mean square electric field amplitudes as a function of incidence angle for a 10 nm film of benzene on (a) aluminium and (b) glass at $\bar{\nu} = 1478\,\mathrm{cm}^{-1}$. The optical constants are $n_0 = 1$, $N_1 = 1.50 + 0.15i$, (a) $N_s = 13 + 64i$ (b) $N_s = 1.16 + 0.008i$. The fields are calculated at at the interface ($z = 0$) for the incident and film media and at $z = d = 10$ nm for the substrate. $\langle E_y^2\rangle$ and $\langle E_x^2\rangle$ are effectively constant over the film so are not plotted independently for each medium. In (a) $\langle E_x^2\rangle$, $\langle E_y^2\rangle$ and $\langle E_{zs}^2\rangle$ are indistinguishable from zero.

When the substrate is dielectric, all three field components have appreciable amplitudes in all media. The continuity of the x and y components across the interfaces and the small scale of the distances

involved relative to the wavelength means that they are essentially the same in all the phases. For normal incidence, the x and y components are indistinguishable and the z component is zero. As the incidence angle is increased, the z component increases at the expense of the x component. The phase change for s-polarisation is near $-180°$ for all angles, so the y component decreases as the reflectance increases with increasing θ_0.

2.4.2 Substrate reflectance

While the actual property of interest is the reflection-absorbance, the reflectance of the substrate is important because of its effect on the signal to noise ratio. The reflectances of aluminium and glass at 80° in the mid-infrared are plotted in Figure 2.8. The reflectance of the metal is very nearly one for s-polarised light at all wavelengths, and increases with decreasing wavenumber from about 0.90 to about 0.94 for p-polarised light. The reflectance of unpolarised light is the arithmetic mean of the reflectances of the two polarisations, $R = \left(R_s + R_p\right)/2$. For glass, the reflectance is reasonably constant from 4000 to about 2000 cm^{-1}, where it starts to decrease, reaching a minimum at about 1400 cm^{-1}. This minimum is due to the real part of the refractive index crossing the refractive index of the incident medium (air; $n \cong 1$) while the imaginary part remains small (Figure 2.3). To the red of the minimum, the reflectance increases sharply then decreases again.

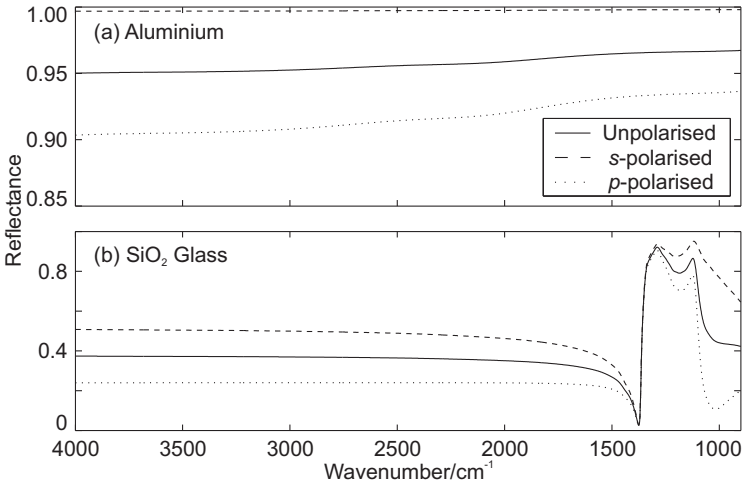

Figure 2.8: Reflectance of aluminium [45] and silica glass [40] in the mid-infrared at an incidence angle of 80°.

The behaviour of the reflectance of glass as the incidence angle is varied is illustrated in Figure 2.9 for several wavelengths, demonstrating the strong wavenumber dependence of the optical properties of this material. In the high-wavenumber region, the traditional reflectance plot is obtained [37]. For s-polarised light, R_s increases monotonically as the incidence angle is increased from normal incidence

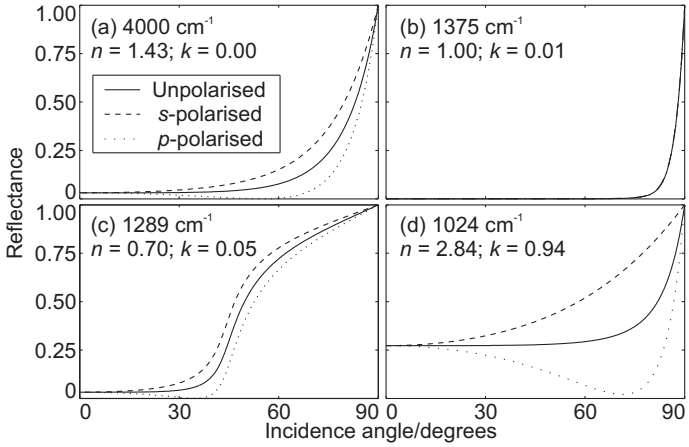

Figure 2.9: Reflectance of glass as a function of incidence angle for several wavelengths. (a) 4000 cm^{-1}, in a region of no absorption. (b) 1375 cm^{-1}, at the 80° reflectance minimum. (c) 1289 cm^{-1}, at the 80° reflectance maximum. (d) 1024 cm^{-1}, at the second 80° reflectance minimum for *p*-polarisation. The optical constants at the selected wavelengths are indicated.

to grazing incidence. For *p*-polarised light, however, R_p decreases as θ is increased, reaching 0 at about 55° (the Brewster angle [47]). At this angle, since the reflectivity for *p*-polarised light is zero, the light reflected from the surface is completely *s*-polarised. At the wavelength corresponding to the minimum in the reflectance in Figure 2.8b, the reflectance is very small except when θ is very near to grazing incidence, and there is little difference between the two polarisations. If the ordinate is greatly expanded, similar behaviour to that in Figure 2.9 a can be seen, with R_p reaching a minimum at $\theta = 45°$. At the wavelength corresponding to the maximum in the reflectance in Figure 2.8, the Brewster angle is about 35°. At this wavelength, R is greater than 0.7 for incidence angles greater than about 60°. At 1024 cm^{-1}, the large value of n causes a more marked difference in the reflectivity for *p*-polarised and *s*-polarised light. Because of the large value of k, there is no angle at which $R_p = 0$, but the minimum in R_p occurs at $\theta = 71°$ (where $R_p = 0.02$); this is called the pseudo-Brewster angle.

2.4.3 Reflection-absorbance: metallic substrate

Figure 2.10 compares the absorbance and reflection-absorbance (aluminium substrate, $\theta = 80°$) spectra of a 10 nm film of liquid benzene. The RA spectrum is about 3–4 times as intense as the absorption spectrum. For this thin a film, there is no obvious band shape distortion or peak shift. As explained by Greenler [24], and expected from the electric field calculations above, the reflection-absorption of *s*-polarised light is very small when the substrate is metallic, in this case not exceeding 1.5×10^{-6}. To quite a good approximation, therefore, the unpolarised RA is half the magnitude of the *p*-polarised reflection-absorbance.

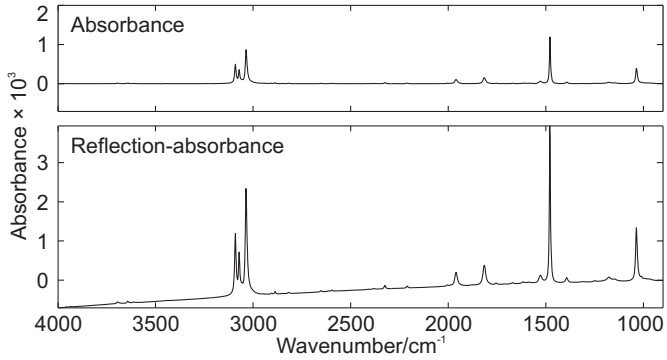

Figure 2.10: Absorbance and reflection-absorbance spectra of a 10 nm film of liquid benzene. The reflection-absorbance calculation is for an aluminium substrate and unpolarised light with an incidence angle $\theta = 80°$. The ordinate scales are the same.

Figure 2.11 illustrates the behaviour of the reflection-absorbance with increasing film thickness. For thin films ($d \ll \lambda$; panels a and c), the reflection-absorbance increases linearly and then falls off: in this example, at $d \approx 150$ nm. For thicker films (panels b and d), complex interference effects can be seen. However, only thin films are expected to be important in this work.

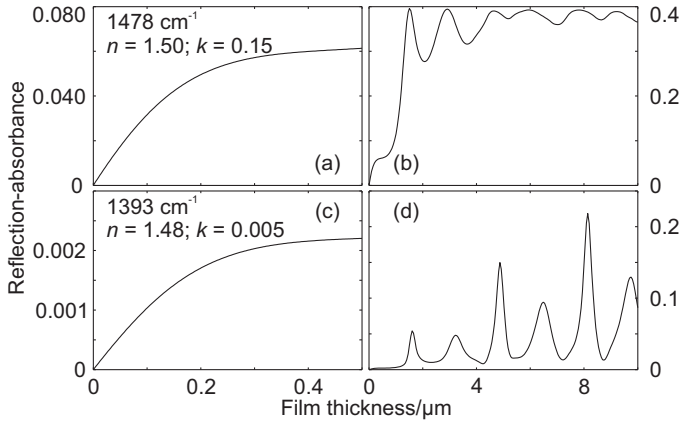

Figure 2.11: Unpolarised RA as a function of film thickness for benzene films on aluminium. The top row of plots corresponds to the maximum of a reasonably strong band; the bottom row to a weak band. The plots on the left are for relatively thin films, while those on the right are for thick films (the wavelength is $\lambda \approx 7$ μm and the substrate optical constants are $n \approx 14$; $k \approx 66$). The spectra were baseline corrected, so the absorbance is relative to a nearby region where benzene has $k = 0$.

2.4.4 Reflection-absorbance: glass substrate

The reflection-absorption spectra of species supported on a dielectric substrate are generally more complicated than when the substrate is metallic, particularly when the substrate material has absorption resonances of its own. In this case, features due to the substrate will appear in the spectrum. This is due to the film modifying the interaction between the light and the substrate. This effect is illustrated

for glass in Figure 2.12. In this figure reflection-absorbance spectra for 10 nm films of benzene and a hypothetical transparent material with refractive index $n = 1.47$ (the average of the real part of the refractive index of benzene over the plotted wavelength range) are compared. The strongest features in the spectra are due to the substrate.

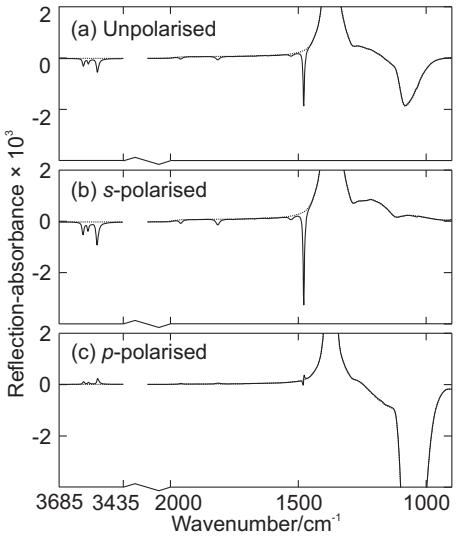

Figure 2.12: Reflection-absorbance spectra ($\theta = 80°$) of a 10 nm film of benzene on silica glass. The dotted lines correspond to a 10 nm film with refractive index $n = 1.47$; the solid lines are for a 10 nm film of benzene. The truncated peak at about 1374 cm^{-1} reaches about 0.10 in (a), 0.14 in (b) and 0.06 in (c). The negative truncated band in (c) descends to about -0.01.

As with the metallic substrate, the polarisation is very important: however, its effect is more complicated. The *s*-polarised spectrum has negative peaks arising from the benzene film. The bands in the *p*-polarised spectrum are much weaker, and exhibit a variety of shapes. The bands at high wavenumber (where the substrate is transparent) are positive and resemble absorbance bands. The band at about 1480 cm^{-1}, which is near the minimum in the substrate reflectance, is very weak and has a derivative-like shape. Since this band is so weak, it does not cause appreciable distortion in the unpolarised spectrum.

Figure 2.13 illustrates the change in shape of the substrate band with increasing film thickness. For thin films, the band is positive and centred at about 1374 cm^{-1}. As the film thickness increases, a negative lobe appears to the blue of the positive band. For thick films, again, more complex effects are evident.

Reflection-absorbance spectra of thin films of benzene on glass are plotted in Figure 2.14. Two groups of bands are plotted: the C–H stretching bands at about 3070 cm^{-1} and another band due to a ring stretching mode at about 1480 cm^{-1}. The former group is in a region where the substrate is transparent, while the latter band is near the Si–O resonance. The depths of the bands relative to the

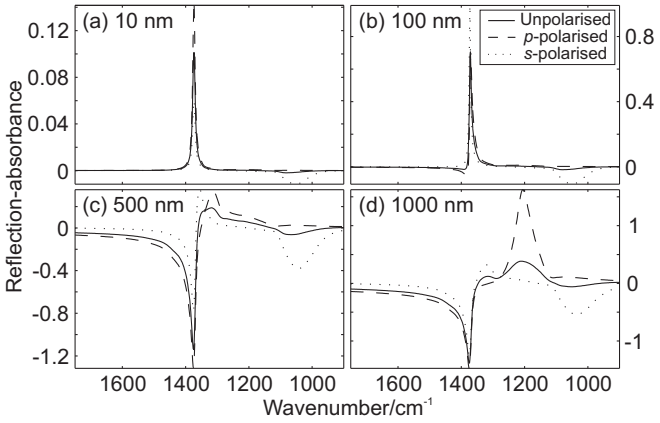

Figure 2.13: Reflection-absorbance spectra ($\theta = 80°$) of films with refractive index $n = 1.47$ on silica glass. Solid line: unpolarised light; dashes: *s*-polarisation; dots: *p*-polarisation.

baseline increase linearly with increasing thickness. For thicker films (Figure 2.15), distortion of the band shape becomes apparent, with a positive lobe appearing to the blue of the band. This effect occurs in the *s*-polarised spectrum as well, so it is not due to the dispersive shape of the *p*-polarised band.

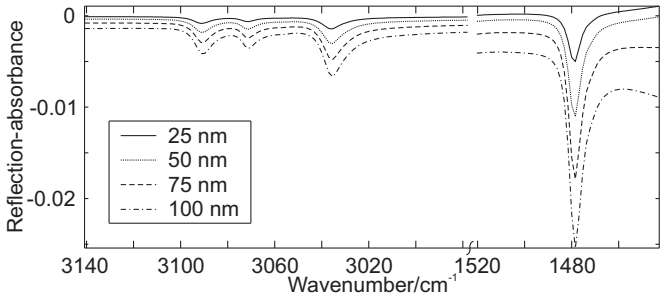

Figure 2.14: Unpolarised reflection-absorbance spectra ($\theta = 80°$) of benzene films on silica glass.

2.5 Applicability of IRRAS to cleaning validation

The assumptions in the above theory are quite restrictive:

1. The substrate and film must be optically isotropic.

2. The substrate and the film must be smooth.

3. The film must be continuous and of equal thickness everywhere.

In the context of pharmaceutical cleaning validation (discussed generally in Chapter 1), the "film" is likely to have been deposited by evaporation of contaminated solvent. As such, it is very unlikely to

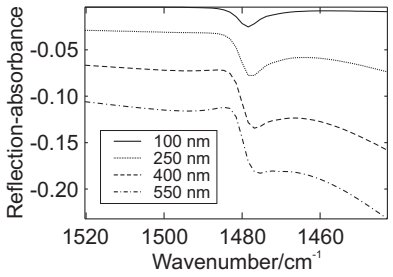

Figure 2.15: Unpolarised reflection-absorbance spectra (θ = 80°) of benzene films on silica glass.

obey the second or third of the conditions above. Still, provided that the film is everywhere thin enough for the proportionality between thickness and RA to hold, the method should be viable.

The purpose of this chapter is not to enable quantitative comparisons between theory and experiment, but to explain qualitatively and illustrate the general features of the IRRAS of organic films on metallic and dielectric layers. An additional goal is to use calculations based on approximate optical constants to determine, semi-quantitatively, the optimum incidence angle for various substrates. This topic is treated in detail in Chapter 5.

Some general observations can be made at this point, however. The nature of the substrate is extremely important. In particular, the biggest division is between metallic and semiconductor or dielectric substrates:

- Metal substrates contribute to strong RA with *p*-polarisation and, for thin films, negligible RA with *s*-polarisation.
- Dielectric substrates exhibit more complicated behaviour, with both polarisations contributing to the spectrum.
- Substrates featuring absorption resonances of their own are particularly interesting: baseline features are seen with all substrates, but in the region of a substrate absorption band, they can dominate the spectrum.

For very thin films, the RA is proportional to the film thickness; for thicker films, complex interference patterns are seen. The likely consequence of this phenomenon, in terms of cleaning validation, is that, in addition to a detection limit determined by the signal-to-noise ratio, there will be an upper limit of quantitation determined by this nonlinear behaviour. The significance of this limit depends on the particular application: it can only be considered a severe limitation of the method if the acceptable residual limit is much higher, or if very heterogeneous contamination is expected.

Chapter 3

Chemometric methods

3.1 Introduction

Chemometrics has been loosely defined by Svante Wold [48], one of the pioneers in the field, as

> How to get chemically relevant information out of measured chemical data, how to represent and display this information, and how to get such information into data.

The goals of chemometrics can thus be summarised as analysis of experimental data, visualisation of the data and the analysis, and design of experiments so as to maximise the information content of the data.

The complexity of each of these tasks increases as the amount of information afforded by the experiment increases. A major distinction can be drawn between univariate and multivariate analyses:

- Univariate:
  - Single concentration variable, single measured signal.
  - Standard linear or nonlinear univariate regression.
  - Experimental data completely represented by a single, two-dimensional plot.
- Multivariate
  - k concentration variables, $m \geq k$ measured signals.
  - Multivariate regression (many different approaches).
  - Visualisation of experimental data difficult.

In univariate analyses, the measured variable must be highly selective for the property of interest. For example, if only one compound in a mixture absorbs at a particular wavelength, then the absorbance

of the mixture at that wavelength is proportional (assuming Beer's law) to the concentration of the compound. If there are two absorbing compounds, then the absorbance at that wavelength is no longer proportional to the first concentration.

If there are multiple absorbing species, there are two ways to proceed. The first is to restore the selectivity of the univariate method by physically separating the analyte from the other species (by chromatography or complexation, for example). The second is to make a series of partially selective measurements (such as absorbance at a number of wavelengths; a spectrum) and use multivariate regression to find a linear combination of the measurements that is fully selective for the analyte. Typically, the first of these options is by far the more time-consuming. The second, however, is more mathematically complicated, and constitutes the first aspect of the definition of chemometrics given above: most of this chapter is devoted to describing some of these mathematical tools.

The basic processes in predictive chemometric modelling are outlined in Figure 3.1. The goal is to build an empirical model that relates certain input data (such as spectra), called the X block, to certain output data (such as concentrations), called the Y block. This is always achieved through a set of data, called the calibration or training set, for which both the inputs and outputs are known. The model can then be used to estimate the Y values for new X data. To obtain an estimate of the uncertainty of new predictions, the model is validated using an additional data set (the test set) with known X and Y blocks.

The remainder of this section is devoted to mathematical preliminaries. The following sections treat chemometric methods relevant to this thesis. Several multivariate regression methods are described, along with tools for model optimisation. The effects of errors in the measured data are also discussed, and a variety of approaches for uncertainty estimation are presented. The treatment given here is not based on any particular text, but good introductions to the basic ideas are given by Martens and Næs [49], by Beebe et al. [50] and by Kramer [51]. A recent IUPAC report [52] gives a good current overview of multivariate calibration in chemistry, as do reviews by Kalivas [53], Hopke [54], and Geladi [55].

3.1.1 Conventions

Scalars are denoted by lower-case italic characters (n), column vectors by lower-case bold Roman characters ($\mathbf{b}$), and matrices by upper-case bold Roman characters ($\mathbf{X}$). The matrix transpose is indicated by a superscript "T" or prime ($\mathbf{X}' \equiv \mathbf{X}^{\mathrm{T}}$). Row vectors are written as transposed column vectors ($\mathbf{b}^{\mathrm{T}}$). Where necessary, measured and estimated quantities are distinguished from their underlying "true" values by a tilde ($\tilde{\mathbf{x}}$) or circumflex ($\hat{\mathbf{x}}$), respectively.

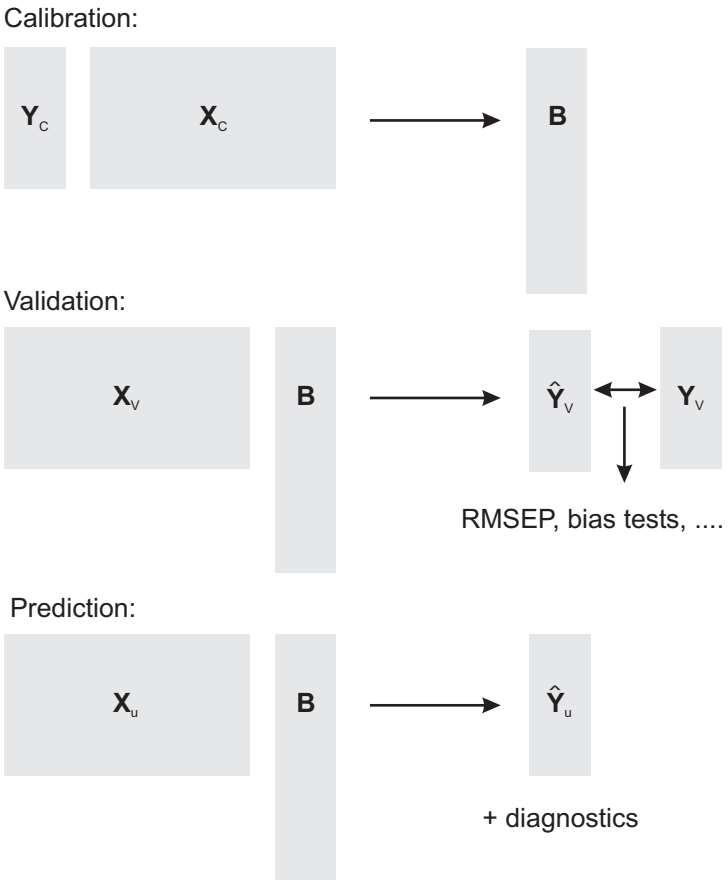

Figure 3.1: Schematic depicting the basic steps in chemometric modelling. In the calibration step a model **B** is generated by some algorithm from calibration (training) spectra $\mathbf{X}_C$ with associated concentrations $\mathbf{Y}_C$. The validation step consists of applying the model to validation set spectra $\mathbf{X}_V$, giving estimated concentrations $\hat{\mathbf{Y}}_V$: these are compared with the known concentrations $\hat{\mathbf{Y}}_V$ to determine the effectiveness of the model. After validation, the model can be applied to genuine unknown samples $\mathbf{X}_U$ to estimate their concentrations. Uncertainty information can also be calculated.

3.1.2 Matrix representation of data

A digitised spectrum is a set of measurements of some photometric property (such as intensity, absorbance, transmittance, reflectance) associated with values of some energy scale (such as wavelength, wavenumber, frequency). Each point along the abscissa is an x variable. When the spectrum itself is of interest, the discrete measured spectrum is thought of as an approximation to the continuous underlying spectrum. For concentration calibration purposes, however, the abscissa value is just a label for a variable rather than a variable itself.

A set of n spectra each comprising values for m variables can be stored as an $n \times m$ or $m \times n$ matrix. The former convention, in which each spectrum is a row of the matrix, will be adopted here. Associated with each spectrum is a set of values of k physical properties[1] pertaining to the sample from which the spectrum was taken. The $\mathbf{Y}$ matrix will therefore have dimensions $n \times k$. Commonly there will only be a single y variable, and $\mathbf{Y}$ will be an $n \times 1$ vector.

3.1.3 Beer's law in matrix form

Beer's law states that the absorbance due to each compound in a solution of several compounds is proportional to the path length of the light through the solution and the concentration of the solution, and independent of the concentrations of the other species [56]. Consequently, the absorbance of the mixture is the sum of the absorbances of the components. For a single wavelength, this can be written

$$A = l \sum_{i=1}^{n} \epsilon_i c_i \tag{3.1}$$

where l is the path length, ϵ is the molar absorptivity and c is the concentration. The absorbance of multiple mixtures at multiple wavelengths and for unit path length can be expressed as a single matrix equation

$$\mathbf{A} = \mathbf{C}\mathbf{S}^{\mathrm{T}} \tag{3.2}$$

where $\mathbf{C}$ is a matrix of concentrations and $\mathbf{S}$ is a matrix of molar absorptivity spectra.

3.1.4 Forward and inverse regression

The difference between forward and inverse regression is in the definition of the dependent and independent variables [52, 57]. In forward regression the concentrations are the independent variables

[1] Typically, these properties will be concentrations of chemical species, or other quantities related to concentrations, such as pH, octane number or flavour. However, any property that influences the spectrum of a sample, such as temperature, is a candidate for calibration.

and the spectroscopic responses are the dependent variables, while for inverse regression the roles are reversed. Since most of this chapter will focus on inverse methods, the spectroscopic response matrix will be denoted **X** and the concentration matrix **Y** (or **y** when there is only a single concentration variable). The principal advantage of inverse regression methods is that they can implicitly model interfering species (absorbing species whose concentrations are unknown) [52]. In forward regression, the concentrations of all absorbing species must be known.

All regression methods produce a regression vector which, ideally, is orthogonal to the spectra of all components except the analyte. This means that the dot product of the regression vector with a measured spectrum is proportional to the concentration of the analyte and completely independent of the concentrations of the other species. Forward and inverse regression methods differ in how the regression vector is estimated. In forward regression, the spectra of all absorbing species are estimated or measured directly, and the regression vector is derived from these spectra in a second step. Inverse regression methods find the regression vector directly.

3.1.5 Software implementation

MATLAB [58] is an environment and language for numerical mathematics and programming that provides a wide array of built-in functions. The language makes it extremely easy to work with arrays. As part of this work a suite of MATLAB programmes for chemometrics has been written using algorithms from the literature. Code listings and instructions are given in Appendix C. Almost all of the programmes are also compatible with GNU Octave [59], a freely available interpreted language that is mostly compatible with MATLAB.

3.2 Regression methods

3.2.1 Classical least squares (CLS) regression

Classical least squares is a forward regression method and may be applied directly to Beer's law (Equation 3.2):

$$\mathbf{X} = \mathbf{Y}\mathbf{K}^{\mathrm{T}} + \mathbf{E} \tag{3.3}$$

where **X** and **Y** are the spectroscopic and concentration data, **K** is the matrix of pure-component spectra and **E** is a matrix of errors. Because CLS is a forward regression method, **Y** must contain concentrations for all components (absorbing species) in all samples. In the calibration step the goal is to find **K**, the

least-squares solution of which is given by

$$\hat{\mathbf{K}}^{\mathrm{T}} = \left(\mathbf{Y}^{\mathrm{T}}\mathbf{Y}\right)^{-1}\mathbf{Y}^{\mathrm{T}}\mathbf{X} \tag{3.4}$$

The quantity $\left(\mathbf{Y}^{\mathrm{T}}\mathbf{Y}\right)^{-1}\mathbf{Y}^{\mathrm{T}}$ is called the pseudoinverse or Moore-Penrose matrix inverse [60] and is denoted $\mathbf{Y}^{+}$. It exists only if $\mathbf{Y}$ has at least as many rows as columns, which corresponds to the experimental requirement that at least as many mixtures be prepared as there are components. Each row of $\mathbf{K}$ is the least-squares estimate of the molar absorptivity spectrum of the compound whose concentrations are in the corresponding column of $\mathbf{Y}$. As an alternative to the above process, the pure component spectra can be measured individually and used to construct $\mathbf{K}$ directly (some authors call this procedure "direct CLS", while Equation 3.4 represents "indirect CLS").

The prediction step (calculation of concentrations for new measurements $\mathbf{X}_{\mathrm{u}}$) consists of solving, in the least-squares sense, Equation 3.3 for $\mathbf{Y}$:

$$\begin{aligned}\hat{\mathbf{Y}}_{\mathrm{u}} &= \mathbf{X}_{\mathrm{u}}\left(\mathbf{K}^{\mathrm{T}}\right)^{+} \\ &= \mathbf{X}_{\mathrm{u}}\left(\mathbf{K}^{+}\right)^{\mathrm{T}} \\ &= \mathbf{X}_{\mathrm{u}}\mathbf{K}\left(\mathbf{K}^{\mathrm{T}}\mathbf{K}\right)^{-1}\end{aligned} \tag{3.5}$$

A property of the pseudoinverse is that

$$\mathbf{K}^{+}\mathbf{K} = \left(\mathbf{K}^{\mathrm{T}}\mathbf{K}\right)^{-1}\mathbf{K}^{\mathrm{T}}\mathbf{K} = \mathbf{I}$$

so that the ith row of $\mathbf{K}^{+}$ is orthogonal to all the columns of $\mathbf{K}$ except the ith (with which its dot product is unity), and therefore constitutes a regression vector for the ith component (that is, $\boldsymbol{\beta}^{\mathrm{T}} \equiv \mathbf{K}^{+}$). A consequence of this property is that only the portion of the ith analyte spectrum that is orthogonal to all the other pure component spectra is used. If species have highly correlated spectra, then only a small portion of the spectrum is actually available for calibration, so the effects of noise are greatly magnified. In this work, CLS regression was used in conjunction with UV colorimetry to determine concentrations of species in solution (see Section 4.3.3).

3.2.2 Inverse least squares (ILS)

Inverse least squares is sometimes called the P-matrix, multiple linear regression (MLR), or ordinary least squares (OLS) method. The model equation is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E} \tag{3.6}$$

in which $\boldsymbol{\beta}$ is a matrix of regression coefficients and $\mathbf{E}$ is a matrix of residuals.

The calibration process consists of finding the least-squares solution to Equation 3.6:

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= \left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \\ &= \mathbf{X}^{+}\mathbf{Y} \end{aligned} \tag{3.7}$$

Concentrations can be predicted from new measurements $\mathbf{X}_{\mathrm{u}}$ by

$$\hat{\mathbf{Y}}_{\mathrm{u}} = \mathbf{X}_{\mathrm{u}}\boldsymbol{\beta} \tag{3.8}$$

In order for $\mathbf{X}^{+}$ to exist, $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ must be invertible, so $\mathbf{X}$ must have at least as many rows as it has columns (i.e. there must be at least as many samples as there are wavelengths). This restriction is not usually met with raw data, since spectra often have hundreds or thousands of variables and data sets usually have fewer samples than this. Consequently, it is usually necessary to reduce the number of X variables prior to regression. This is most simply achieved by choosing a set of variables to use and discarding the rest of the spectroscopic variables, but new variables can be created by, for example, curve fitting or band integration.

3.2.3 Principal component regression (PCR)

Principal component decomposition

Principal component decomposition [49] is a method by which the matrix $\mathbf{X}$ is expressed as the product of two matrices:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} \tag{3.9}$$

The matrix $\mathbf{T}$ has dimensions $n \times m$ and is called the scores matrix; $\mathbf{P}$ ($m \times m$) is called the loadings matrix. Both $\mathbf{T}$ and $\mathbf{P}$ have orthogonal columns. In addition, the columns of $\mathbf{P}$ are normalised to unit

length. These two matrices are most commonly found by a singular value decomposition (SVD) [61]:

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^{\mathrm{T}} \tag{3.10}$$

where $\mathbf{U}$ and $\mathbf{V}$ have orthonormal columns and $\mathbf{S}$ is diagonal. The relationship between the SVD and the principal component decomposition is: $\mathbf{T} = \mathbf{U}\mathbf{S}$ and $\mathbf{P} = \mathbf{V}$.

For noiseless measurements of a system obeying Beer's law, Equation 3.2 dictates that the rank of $\mathbf{X}$ is k, the number of chemical components. In practice, noise ensures that (formally) $\mathbf{X}$ has full rank. However, only the first k columns of $\mathbf{T}$ and of $\mathbf{P}$ contain genuine information; the remaining columns of the matrices describe only noise. The number of significant factors is sometimes termed the pseudorank. The similarity between Equation 3.9 and Equation 3.2 gives a clue to the meanings of $\mathbf{T}$ and $\mathbf{P}$. The columns of $\mathbf{T}$ are linear combinations of the columns of $\mathbf{C}$ (sometimes called abstract concentration profiles [62]), while the rows of $\mathbf{P}$ are linear combinations of the rows of $\mathbf{S}$ (abstract spectra). Features of the data such as baseline variations and certain kinds of nonlinearities (spectra changing shape with concentration) result in additional components. The term "principal component" can refer either to a single column of either $\mathbf{T}$ or $\mathbf{P}$, or to the two collectively.

The consequence of this rank deficiency is that $\mathbf{X}$ can usually be regenerated to within the noise level by using only a few of the principal components. In fact, this has the effect of filtering noise from the data [62], since noise that is orthogonal to the principal components is discarded completely. This regeneration is written

$$\mathbf{X}_A = \mathbf{T}_A\mathbf{P}_A^{\mathrm{T}} \tag{3.11}$$

where $\mathbf{T}_A$ and $\mathbf{P}_A$ consist of the first A columns of $\mathbf{T}$ and $\mathbf{P}$, respectively.

Principal component regression

Principal component decomposition can be used to reduce the dimensionality of $\mathbf{X}$ to provide new, uncorrelated variables (the columns of $\mathbf{T}_A$) to use with ILS, with the combination being called principal components regression (PCR). The first step is to calculate the principal component decomposition (Equation 3.9) and to decide on a value of A, the number of significant factors. This can be achieved by statistical tests based on the magnitude of the singular values (the diagonal elements of $\mathbf{S}$) [63, 64]; however, for real data these tests can be difficult to interpret correctly [65], and resampling approaches like cross-validation (described below in Section 3.3.1) are usually preferred. Once A has been deter-

mined, the regression is performed between $\mathbf{Y}$ and $\mathbf{T}_A$:

$$\mathbf{Y} = \mathbf{T}_A\mathbf{B}_A + \mathbf{E} \tag{3.12}$$

$$\hat{\mathbf{B}}_A = \left(\mathbf{T}_A^{\mathrm{T}}\mathbf{T}_A\right)^{-1}\mathbf{T}_A^{\mathrm{T}}\mathbf{Y} \tag{3.13}$$

The invertibility of $\mathbf{T}_A\mathbf{T}_A^{\mathrm{T}}$ is guaranteed because the columns of $\mathbf{T}$ are orthogonal. $\mathbf{B}_A$ is a $k \times A$ matrix of regression coefficients interrelating the concentrations and the scores.

The first step in prediction for new spectra $\mathbf{X}_{\mathrm{u}}$ is calculating the scores matrix for the new spectra; that is, projecting the new spectra onto the basis formed by the loading vectors generated from the calibration set spectra:

$$\mathbf{T}_{A,\mathrm{u}} = \mathbf{X}_{\mathrm{u}}\mathbf{P}_A \tag{3.14}$$

Since the columns of $\mathbf{P}_A$ are normalised to unit length, this corresponds to finding the orthogonal projection of each new spectrum (row of $\mathbf{X}_{\mathrm{u}}$) onto each of the loading vectors (columns of $\mathbf{P}_A$). Now the regression vector estimated earlier is used to predict the concentrations from the new scores:

$$\hat{\mathbf{Y}}_{\mathrm{u}} = \mathbf{T}_{A,\mathrm{u}}\mathbf{B}_A \tag{3.15}$$

These two steps can be combined into a single matrix multiplication by defining a new regression vector $\boldsymbol{\beta}_A$ in terms of the original spectroscopic variables:

$$\hat{\boldsymbol{\beta}}_A = \mathbf{P}_A\left(\mathbf{T}_A^{\mathrm{T}}\mathbf{T}_A\right)^{-1}\mathbf{T}_A^{\mathrm{T}}\mathbf{Y} \tag{3.16}$$

$$\hat{\mathbf{Y}}_{\mathrm{u}} = \mathbf{X}_{\mathrm{u}}\boldsymbol{\beta}_A \tag{3.17}$$

The MATLAB code for PCR is very simple, and can be found inside the `CROSSVAL` programme in Appendix C.

Polynomial principal component regression (PPCR)

PCR is the application of ILS to data that have been compressed by principal component analysis. Other regression methods can also be used. For example, if the underlying relationship between $\mathbf{X}$ and $\mathbf{y}$ is nonlinear, so will be the relationship between $\mathbf{T}_A$ and $\mathbf{Y}$, and some function other than a straight line might provide a better fit. Linear regression can still be used provided that the model is linear in the parameters ($\mathbf{B}$) [66]. To include higher-order terms, all that is necessary is to carry out the regression

against $\mathbf{T}_A^*$, defined as

$$\mathbf{T}_A^* = \left[\mathbf{1} \;\; \mathbf{T}_A \;\; \mathbf{T}_A^2 \ldots\right] \tag{3.18}$$

where $\mathbf{T}_A^2$ is obtained by squaring each element of $\mathbf{T}_A$ and $\mathbf{1}$ is an optional vector of ones, allowing for a constant term in the model. It should be noted that the number of predictor variables in polynomial PCR is equal to $pA + 1$, where p is the polynomial order: so a polynomial PCR model consumes many more degrees of freedom than a first-order one.

Prediction is achieved in exactly the same way as for first-order PCR, but with the matrix $\mathbf{T}_{A,\mathrm{u}}^*$ in place of $\mathbf{T}_{A,\mathrm{u}}$. MATLAB code for polynomial PCR is also listed as part of `CROSSVAL`.

3.2.4 Partial least squares (PLS) regression

In PCR, the new variables are found by an orthogonal decomposition of $\mathbf{X}$, and then the relationship between the new variables and the concentration variables is estimated in a separate regression step. In the decomposition step, each successive factor explains the maximum possible variance in $\mathbf{X}$. In PLS, each successive factor explains the maximum possible covariance between $\mathbf{X}$ and $\mathbf{Y}$ [67]. Consequently, for a given number of factors less than the pseudorank of $\mathbf{X}$, PLS will tend to have better predictive performance than PCR (while PCR will tend to give smaller residuals in the $\mathbf{X}$ block). Often, there is little difference between the optimum results for the two methods [68]. Despite this, PLS is vastly more popular. A discussion of the relative merits of PCR and PLS is given by Hasegawa [69].

Another consequence of the dependence of PLS on $\mathbf{Y}$ is that when there are multiple response variables, there are two ways to proceed. Calibration can be performed for one analyte at a time (PLS-1) or for several analytes together (PLS-2). It has generally been found that PLS-1 performs slightly better [70], and it also allows more flexibility, since different numbers of factors may be chosen for each response variable. It should be noted that for PCR, because the decomposition is independent of $\mathbf{Y}$, these two approaches are equivalent. However, it may still be preferable to build separate models so that different parameters (pretreatments, wavelength ranges) may be chosen.

In common with PCR, PLS has score ($\mathbf{T}$) and loading ($\mathbf{P}$) matrices, but their definitions are somewhat different, as explained below. In addition, there is an $m \times A$ weights matrix $\mathbf{W}$ and score and loading matrices $\mathbf{U}$ and $\mathbf{Q}$ for the Y block. $\mathbf{T}$ and $\mathbf{W}$ are each orthogonal, but $\mathbf{P}$ is not. There are several equivalent PLS algorithms, of which the classical NIPALS (nonlinear iterative partial least squares) algorithm is the most straightforward. It is presented below (and in diagrammatic form in Figure 3.2). In this algorithm $\mathbf{W}$, $\mathbf{T}$ and $\mathbf{P}$ are constructed by appending a column vector ($\mathbf{w}_a$, $\mathbf{t}_a$, $\mathbf{p}_a$) for each factor a.

1. Mean-centre and/or scale $\mathbf{X}$ and $\mathbf{Y}$ (optional; see Section 3.3.3)

2. Define $\mathbf{E}_0 = \mathbf{X}$, $\mathbf{F}_0 = \mathbf{Y}$

3. For a in $\{1 \dots A\}$

   (a) Initialise the Y score vector $\mathbf{u}_a$ as a column of $\mathbf{F}_{a-1}$

   (b) Calculate the ath weight vector: $\mathbf{w}_a^{\mathrm{T}} = \mathbf{u}^{\mathrm{T}}\mathbf{E}_{a-1}/\mathbf{u}^{\mathrm{T}}\mathbf{u}$

   (c) Normalise the weight vector: $\mathbf{w}_a \leftarrow \mathbf{w}_a/\|\mathbf{w}_a\|$

   (d) Calculate the score vector (latent variable): $\mathbf{t}_a = \mathbf{E}_{a-1}\mathbf{w}$

   (e) Calculate the Y loading vector: $\mathbf{q}_a^{\mathrm{T}} = \mathbf{t}^{\mathrm{T}}\mathbf{F}_{a-1}/\mathbf{t}_a^{\mathrm{T}}\mathbf{t}_a$

   (f) Calculate the new Y score vector: $\mathbf{u}_a = \mathbf{F}_{a-1}\mathbf{q}_a/\mathbf{q}_a^{\mathrm{T}}\mathbf{q}_a$

   For PLS-2, iterate steps 3b–3f until convergence is reached (no change in $\mathbf{u}_a$ beyond rounding error).

   (g) Calculate the X loading vector: $\mathbf{p}_a^{\mathrm{T}} = \mathbf{t}_a^{\mathrm{T}}\mathbf{E}_{a-1}/\mathbf{t}_a^{\mathrm{T}}\mathbf{t}_a$

   (h) Update the X matrix: $\mathbf{E}_a = \mathbf{E}_{a-1} - \mathbf{t}_a\mathbf{p}_a^{\mathrm{T}}$

   (i) Update the Y matrix: $\mathbf{F}_a = \mathbf{F}_{a-1} - \mathbf{u}_a\mathbf{q}_a^{\mathrm{T}}$

This algorithm is iterative for PLS-2, but for PLS-1 it can be thought of as a series of least-squares regression steps [70]. Step 3b above is equivalent to a CLS calibration for a single-component system (in PLS-1 Step 3a simply sets $\mathbf{u}_1 = \mathbf{y}$). Subsequent weight vectors are calculated in the same way, but working with the residual matrices $\mathbf{E}$ and $\mathbf{F}$.

Step 3d is equivalent to a CLS prediction step in which the concentrations in the calibration set are predicted using the spectrum, $\mathbf{w}_a$, estimated in the previous step. Each element of $\mathbf{t}_a$ indicates how much of $\mathbf{w}_a$ is present in the corresponding row of $\mathbf{X}$.

In Step 3e the concentration vector is regressed on the scores vector. This is similar to the regression step in PCR, in which the concentrations are related to the intensities in the principal component coordinate system 3.12.

The purpose of Step 3g is to ensure orthogonal $\mathbf{t}$ vectors. The vector $\mathbf{p}_a$ is the least-squares solution in the model $\mathbf{E}_{a-1} = \mathbf{t}_a\mathbf{p}_a^{\mathrm{T}}$. Unlike for PCR, this vector is not optimal in the sense of explaining as much variance in $\mathbf{X}$ as possible. It accounts for as much variation in $\mathbf{X}$ as possible while being correlated with $\mathbf{t}$, which is an approximation for $\mathbf{y}$.

Finally, the PLS approximations to $\mathbf{E}_{a-1}$ and $\mathbf{F}_{a-1}$ are subtracted from them, and the process is repeated for the next factor.

For PLS-2, the interpretation is not quite so straightforward, since the vector $\mathbf{u}_a$ no longer represents a single component but a linear combination of all the species being calibrated for. In this case, the algorithm becomes iterative: $\mathbf{u}_a$ starts out as a column of $\mathbf{F}_{a-1}$ but is recalculated iteratively until it converges to the latent variable (analogous to $\mathbf{t}_a$) that accounts for the largest portion of the covariance between $\mathbf{X}$ and $\mathbf{Y}$.

The steps for the prediction of concentrations $\mathbf{Y}_u$ from new spectra $\mathbf{X}_u$ in the classical algorithm are as follows:

1. Mean-centre and/or scale $\mathbf{X}_\mathrm{u}$ using parameters from calibration.
2. Define $\mathbf{E}_{\mathrm{u},0} = \mathbf{X}_\mathrm{u}$; $\mathbf{Y}_{\mathrm{u},0} = \mathbf{0}$
3. For a in $\{1 \ldots A\}$
   (a) Calculate the score vector: $\mathbf{t}_{\mathrm{u,a}} = \mathbf{E}_{\mathrm{u,a}-1}\mathbf{w}_a$
   (b) Update the concentration matrix: $\mathbf{Y}_{\mathrm{u,a}} = \mathbf{Y}_{\mathrm{u,a}-1} + t_{\mathrm{u,a}}\mathbf{q}_a^\mathrm{T}$
   (c) Update the spectroscopic matrix: $\mathbf{E}_{\mathrm{u,a}} = \mathbf{E}_{\mathrm{u,a}-1} - \mathbf{t}_{\mathrm{u,a}}\mathbf{p}_a$
4. Add the mean concentrations from calibration and/or apply scaling factors to $\mathbf{Y}_{\mathrm{u,A}}$, as appropriate.

Alternatively, prediction can be achieved by calculating the regression vector (matrix, for PLS-2) and postmultiplying $\mathbf{X}_\mathrm{u}$ with it:.

$$\boldsymbol{\beta} = \mathbf{W}_A \left(\mathbf{P}_A^\mathrm{T}\mathbf{W}_A\right)^{-1} \mathbf{Q}_A^\mathrm{T} \tag{3.19}$$

$$\hat{\mathbf{Y}}_\mathrm{u} = \mathbf{X}_\mathrm{u}\boldsymbol{\beta} \tag{3.20}$$

The NIPALS algorithm has been superseded by faster (but equivalent) algorithms such as the kernel algorithm [71] and variations thereof [72]. An alternative formalism is SIMPLS by De Jong [73] in which each loading vector is defined with respect to the original spectroscopic matrices rather than residual matrices. SIMPLS with one Y variable is equivalent to PLS-1, but when there is more than one Y variable it produces slightly different results from PLS-2. Since PLS-2 is rarely used and SIMPLS has certain other advantages relating to notation [74], SIMPLS has become the standard algorithm. In this thesis the "improved kernel algorithm #1" by Dayal and MacGregor [72] has been used, as informal benchmarking revealed it to have a significant speed advantage over the standard NIPALS and kernel algorithms for the relevant sizes of data matrices.

Since PLS-2 is not used in this thesis, the concentration data in any discussion of PLS will be represented by a vector, $\mathbf{y}$.

3.3 Model optimisation

For any model, there will be a number of choices to make. For example: which X variables should be used? What is the best number of factors? What pretreatments are optimal? The aim is to choose the model that has the best "predictive ability" for future samples. Predictive ability is usually quantified by one or several measures based on the fit of the model to training and validation data sets.

3.3.1 Measures of predictive ability

Generally, there will be a limited number of standards (objects i for which both $\mathbf{x}_i$ and $\mathbf{y}_i$ have been measured) available. With these standards, a model must be built and optimised, and an estimate of its quality obtained. A quality metric often used in regression problems is the mean squared error of calibration, MSEC:

$$\text{MSEC} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - df} \tag{3.21}$$

where n is the number of standards in the calibration set and the $\hat{y}_i$ are their predicted (or fitted) concentrations, i.e. the elements of the vector $\hat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}$. The term df in the denominator represents the number of degrees of freedom used by the model. For ordinary linear regression (MLR) this is equal to the number of model parameters. For PCR, df is equal to the number of principal components used (mean-centring uses an additional degree of freedom). For PLS, however, the regression vector uses information from $\mathbf{Y}$ as well as $\mathbf{X}$, so each factor uses more than one degree of freedom (unless very many factors are used: since there are only n degrees of freedom in total, later factors use less than a degree of freedom each). Van der Voet [75] pointed out that in fact the problem is deeper: since the PLS estimator of the regression vector is nonlinear, the term "degree of freedom" is undefined. By considering the relationship between predictive ability and degrees of freedom he derived an expression for the "pseudo-degrees of freedom" (PDF) in a model, based on the results of a leave-one-out cross-validation (see below for a discussion of cross-validation).

$$\text{MSEP}_{\text{rs}} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 / n \tag{3.22}$$

$$\text{MSECV} = \sum_{i=1}^{n} (y_i - \hat{y}_{i,\text{cv}})^2 / n \tag{3.23}$$

$$\text{PDF} = n(1 - \sqrt{\text{MSEP}_{\text{rs}}/\text{MSECV}}) \tag{3.24}$$

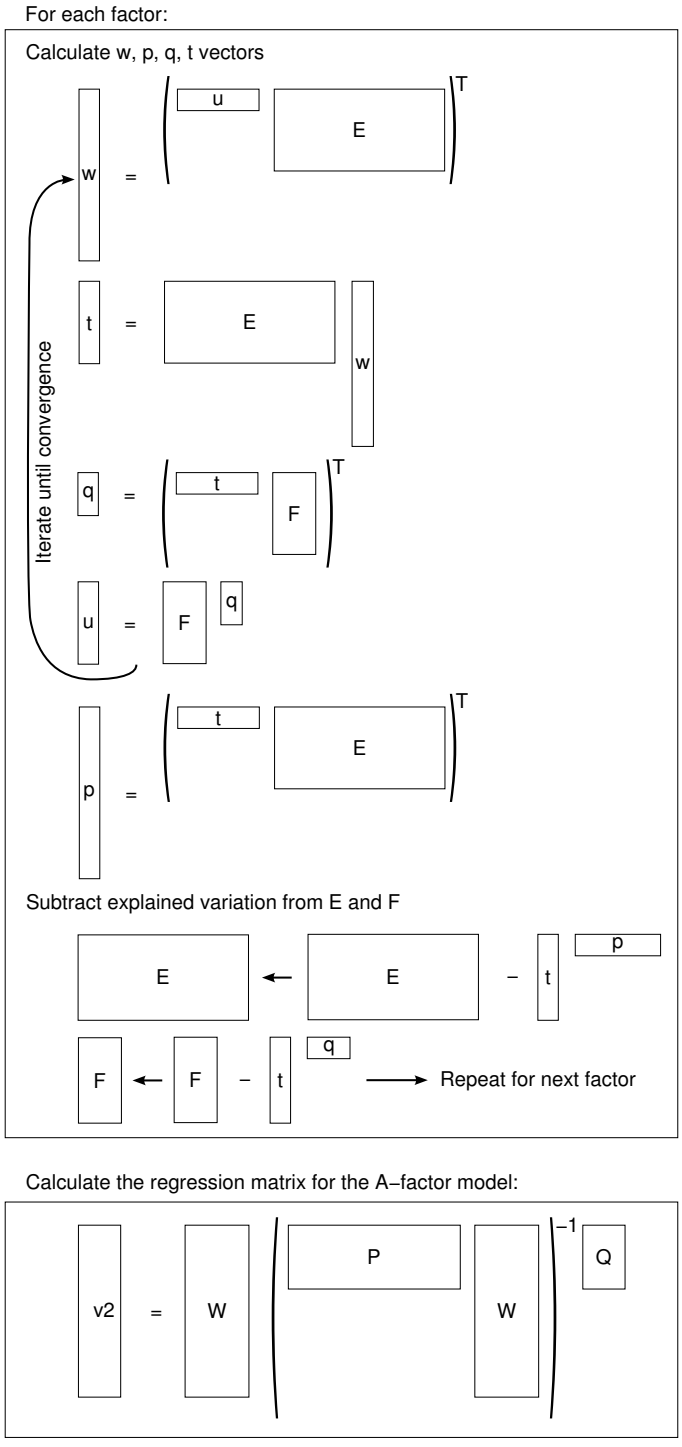

Figure 3.2: The NIPALS algorithm for PLS regression. For clarity, normalisations have been omitted (see in the text).

MSEP_{rs} is the mean squared error for resubstitution (the $\hat{y}_i$ are calculated by applying the model to the calibration spectra), and is like the MSEC but without the degrees-of-freedom correction. MSECV is the mean squared error of cross-validation (see below).

If this estimate for the degrees of freedom is inserted into the equation for the MSEC, the following relationship can be derived:

$$\text{MSEC} = \frac{\sum(y_i - \hat{y}_i)^2}{n - \text{PDF}} \tag{3.25}$$

$$= \sqrt{\text{MSEP}_{\text{rs}}\text{MSECV}} \tag{3.26}$$

Test-set validation

Because it represents the lack of fit of the model to the data used to create the model, rather than to independent data, the MSEC does not always provide a useful estimate of the model's predictive ability [51]. To avoid this problem, it is common to divide the available standards into a calibration set and a test set. The calibration set is used to generate the model, and the test set is used to evaluate its quality. This is quantified by the mean square error of prediction, MSEP, which differs from MSEC only in that the summation is carried out over items in the test set rather than the calibration set.

$$\text{MSEP} = \frac{\sum_{i=1}^{n_{\text{test}}} (y_i - \hat{y}_i)^2}{n_{\text{test}}} \tag{3.27}$$

In this case there is no need to correct for the number of degrees of freedom, since the predictions are made for independent data. MSEP is an unbiased estimate of the average squared prediction error, and therefore a reasonable indicator of the quality of the model. However, there are several problems with this approach. First, the reduction in size of the calibration set will generally decrease the quality of the model. Second, the variance of MSEP may be very high, depending on the size of the test set (see Appendix A.4): and if this is to be reduced, more samples must be removed from the calibration set, further decreasing its predictive ability. Finally, an estimate of the model's predictive ability is required both for optimising the model and, once a model has been selected, for evaluating its usefulness. If the same test set is used for both these purposes, selection bias will be introduced, so that the MSEP is no longer unbiased.[2] An unbiased estimate would necessitate an additional "verification" set, again at the expense of the calibration set.

[2] As an uncertainty measure, MSEP is highly variable, particularly if the test set is small: any calculated MSEP has an error associated with it (see Appendix A). Since the best model is chosen on the basis of having the lowest MSEP, the error term for that MSEP value is more likely to be negative than if the model were chosen at random from the available models. Clearly, the relevance of this concern increases with the number of potential models considered.

Cross-validation

A more efficient approach (in terms of laboratory effort) is to use resampling. The simplest form of this is cross-validation [76], in which the calibration set is divided into k subsets, each of which in turn is designated as a temporary test set. The model is built from the remaining calibration data, and predictions $\hat{y}_{i,\mathrm{cv}}$ are made for the "left-out" samples. The mean square error of cross validation (MSECV) is calculated as

$$\mathrm{MSECV} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_{i,\mathrm{cv}})^2}{n} \tag{3.28}$$

The most frequently used form of cross-validation is "leave-one-out" or "full" cross-validation, in which each calibration object is left out individually (i.e., $k = n$).

It should be noted that a leave-one-out cross-validation requires n models to be built to obtain a single MSECV. Depending on the size of the dataset and the number of variable subsets and pretreatments to be considered, the amount of computation required may be significant. This has not been a problem in the present work since the datasets are not unduly large.

A frequent compromise between efficient use of available standards and the elimination of selection bias is to use a cross-validation for optimisation and a test-set validation for independent evaluation. It is important that MSEP is only calculated once model optimisation is complete, to ensure that it is not contaminated by selection bias.

Other resampling methods

In Monte-Carlo cross-validation, the basic idea is to leave out a large portion (half or more) of the calibration standards during each iteration of the cross-validation. The samples to be left out are chosen randomly, and enough iterations are performed to obtain stable results. This procedure counters the tendency of leave-one-out cross-validation to indicate the use of too many factors [77, 78]. However, an upward bias is introduced to the estimated MSEP since in each step the model is built from many fewer standards than will be used for the final model [79].

Bootstrapping [80] is a general statistical technique based on the idea of generating many new data sets by sampling with replacement (so that a given object may be present several times in a new set) from the pool of available objects. Bootstrapping is very versatile, and can be used to estimate almost any desired statistic. A description of the use of the bootstrap to estimate confidence intervals in prediction is given below in Section 3.4.2.

3.3.2 Choosing the optimum rank

In biased regression methods such as PLS and PCR, there is a trade-off between bias and variance (both of which contribute to the MSEP: algebraically, MSEP = bias^2 + variance [81]). The more factors used, the lower the bias and the higher the variance: the optimum model has the minimum MSEP. The bias-variance tradeoff is discussed by Faber [81], and is illustrated schematically in Figure 3.3. Using too few or too many factors is called underfitting or overfitting, respectively. Faber asserts that, in practice, the bias associated with underfitting decreases rapidly as a function of model complexity, while the variance due to overfitting increases only slowly. These trends result in the minimum MSEP corresponding to a model with very small bias. Intuitively, one would expect the behaviour of the bias term to depend (in the Beer's law example) on the number and magnitude of independent contributions to the spectrum: bias would drop off quickly with increasing number of factors if there were only a few species absorbing appreciably, and more slowly if the analyte absorbance were buried amongst strong absorbances due to many interfering species.

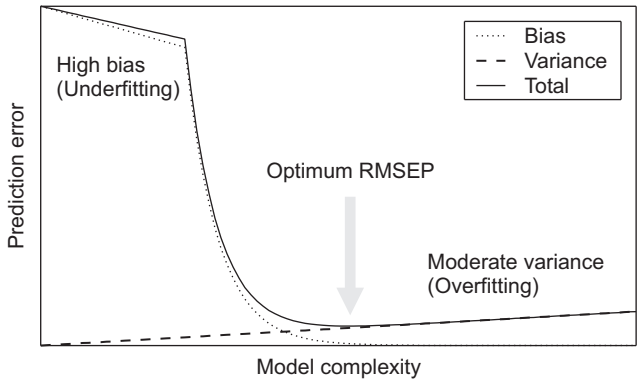

Figure 3.3: Schematic representation of the bias-variance tradeoff, after Ref. 81. Bias falls off quickly with increasing model complexity (rank), while variance increases only slowly: consequently, the bias is small for the optimum model.

A common first step is to conduct a leave-one-out cross validation and examine a plot of MSECV *vs* rank. When insufficient factors are used, the predictions will be strongly biased, leading to a high MSECV (underfitting). As more factors are added, the bias decreases, but this is offset by an increase in variance (overfitting) [74]. Ideally, there will be a clear minimum, the model giving the best compromise between bias and variance. However, it is common for the minimum MSECV to occur at a large number of factors, when a "smaller" model will perform as well or better in practice. Several authors have commented on this tendency of leave-one-out cross-validation to suggest overfitted models [70, 77, 82, 83].

Another important point relates to model robustness. If model errors (such as those associated with

transferring a calibration to a new instrument) dominate random errors due to measurement noise, a robust model is required. Models with fewer factors are generally regarded as being more robust [81]. So, if model errors are anticipated to be important, the actual penalty associated with using additional factors may be much greater.

A common rule of thumb is to select the first local minimum in MSECV, or the first factor at which the MSECV starts to plateau [74]. It is useful to have an algorithmic method to decide upon the rank, since this removes subjectivity from the decision and allows the choice to be automated. Several methods garnered from the literature are presented below.

F-test on MSECV

This method was suggested by Haaland and Thomas [70]. The rank of the model with the minimum MSECV is denoted A^* and all models with rank $A \leq A^*$ are considered as candidates. For each model, the test statistic $F_A = \text{MSECV}(A)/\text{MSECV}(A^*)$ is compared to $F_{\alpha,n,n}$, the $1-\alpha$ percentile of the F distribution with n and n degrees of freedom. The smallest A such that $F_A < F_{\alpha,n,n}$ is chosen as the ideal model. The choice of the significance level α is arbitrary; Haaland and Thomas recommend 0.25 based on their experience.

The major problem with this test is that, since the prediction errors for one model are very likely to be correlated to those for another (since they represent the same samples), the normal probabilistic interpretation of the F-test does not hold. The effect of this correlation is to suggest models with too few factors. The relatively high value of α they recommend presumably reflects this tendency.

Adding a penalty function to RMSECV

Martens and Dardenne [82] recommend an approach based on adding a linear function to the RMSECV[3] curve and then selecting the minimum. They look for the minimum in the function

$$C(A) = \text{RMSECV}(A) + sA \times \text{RMSECV}(0) \tag{3.29}$$

where RMSECV(0) is calculated for a "zero-factor" model, in which the mean of the calibration set is used for all predictions. For a leave-one-out cross-validation,

$$\text{RMSECV}(0) = \sqrt{\frac{\sum_{i=1}^{n}\left(y_i - \frac{\sum_{j\neq i} y_j}{n-1}\right)}{n}} \tag{3.30}$$

[3] An "R" prepended to an abbreviation for a mean-squared-error quantity indicates the square root be taken: RMSECV is the root-mean-square error of cross-validation, $\sqrt{\text{MSECV}}$.

and the constant s is chosen beforehand (they suggest $s = 0.05$). The interpretation of this criterion is that for an additional factor to be included, it must contribute an improvement of at least s times the amount of variation present in $\mathbf{Y}$, as measured by RMSECV(0). This method is appealing because it is very simple and formalises the standard "inspection of MSECV *vs* rank" method.

It should also be noted that RMSECV(0) is very closely related to the standard deviation of y, and constitutes a useful benchmark value to which to compare the optimal RMSECV. If the RMSECV is of similar magnitude to RMSECV(0), the model is only of marginal utility, since simply using the calibration-set mean for every new prediction would give equally good results.

Other methods

Denham [84] compares a variety of methods for choosing the number of factors. His methodology is based around the definition of the optimal rank as giving the lowest MSEP for future samples: if the complete pool of future samples were available with known concentrations, then the ideal rank could be determined by evaluating the MSEP at each rank and choosing the minimum. In reality the search for the optimal rank is equivalent to the search for a good estimate of the future MSEP obtainable with data available in the calibration step. He considers several methods, including cross-validation, bootstrapping, and several formulae based on a linearisation of the regression coefficient (it is the nonlinearity of the regression coefficient in PLS that makes uncertainty estimates difficult [74]). He concludes that for data sets where the number of variables exceeds the number of observations (common in spectroscopy), the resampling approaches work best. He states that, in cross-validation, the number of factors can be chosen to correspond to either the global minimum in MSECV or the first local minimum, but does not elaborate on which is to be preferred.

Green and Kalivas [83] describe a number of graphical diagnostics. They divide them into measures that reflect mostly bias and those that reflect mostly variance. MSEC, MSECV and the determination coefficient R^2 are given as examples of the former. They present several measures of variance. One is the "A-value", defined as

$$A\text{-value} = tr(\mathbf{X}_A^{\mathrm{T}}\mathbf{X}_A)^{+} \tag{3.31}$$

where $\mathbf{X}_A$ is the $\mathbf{X}$ matrix rebuilt with A PCR or PLS factors and tr is the matrix trace operator. Minimising the A-value minimises the maximum variance of a predicted value in the calibration set. Another useful measure of the model variance is the Euclidian norm of the regression vector, $||\boldsymbol{\beta}||$. This measure is related to the variance of the regression coefficients and of predicted concentrations [74]. Green and Kalivas also mention Van der Voet's pseudo-degrees of freedom, as an indicator of model complexity,

which is associated with variance. Their advocated approach is to look for "corners" in plots of a variance measure against a bias measure, since a corner represents a region where neither bias nor variance is excessively high (and from which moving increases either bias or variance).

It seems that MSECV should include a variance component as well as a bias component, since the prediction is of a sample not included in the calibration set. It is therefore subject to the variance in prediction that arises from overfitting. Green and Kalivas' approach may be accounting twice for the variance if MSECV is plotted against a variance measure, potentially leading to underfitted models.

Conclusions

It is apparent that, as yet, there is no clear-cut method for selecting the optimal rank on the basis of cross-validation or test-set validation results. In this work a combination of subjective examination and, where automation is required, Martens' method with $s \approx 0.02$ has been used.

3.3.3 Pre-processing

Pre-processing is any treatment of the data matrices prior to regression analysis. The idea is generally to remove variation that is not associated with the property of interest, such as baseline drift. Pre-processing is used extensively in near infrared (NIR) calibrations to correct for baseline variations and scattering.

Mean centring and variance scaling

A variable is mean-centred by subtracting its mean value: the mean of a mean-centred variable is zero. Variance scaling is dividing a variable by its standard deviation so that it has unit variance, and is often necessary when variables in different units are being combined into a single response profile: it is not normally used for spectroscopic variables and will not be discussed further.

When mean centring is used during calibration, the mean of each variable is subtracted from all instances of that variable:

$$\mathbf{X}_0 = \mathbf{X} - \mathbf{1}\bar{\mathbf{x}}^{\mathrm{T}} \tag{3.32}$$

$$\mathbf{Y}_0 = \mathbf{Y} - \mathbf{1}\bar{\mathbf{y}}^{\mathrm{T}} \tag{3.33}$$

where $\mathbf{1}$ is a vector of ones. The model is calculated using $\mathbf{X}_0$ and $\mathbf{Y}_0$. The mean values $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ must

be saved to apply during the prediction step:

$$\mathbf{X}_{0,\mathrm{u}} = \mathbf{X}_{\mathrm{u}} - \mathbf{1}\bar{\mathbf{x}} \tag{3.34}$$

$$\mathbf{Y}_{\mathrm{u}} = \bar{\mathbf{y}} + \mathbf{X}_{0,\mathrm{u}}\boldsymbol{\beta} \tag{3.35}$$

Alternatively, mean centring can be accomplished by a projection, using the relation $\bar{\mathbf{x}}^{\mathrm{T}} = \left(\mathbf{1}\mathbf{1}^{\mathrm{T}}/n\right)\mathbf{X}$:

$$\begin{aligned}\mathbf{X}_0 &= \mathbf{X} - \left(\mathbf{1}\mathbf{1}^{\mathrm{T}}/n\right)\mathbf{X} \\ &= \left(\mathbf{I} - \mathbf{1}\mathbf{1}^{\mathrm{T}}/n\right)\mathbf{X} \\ &= \mathbf{P}^{\perp}\mathbf{X}\end{aligned} \tag{3.36}$$

When cross-validating a mean-centred model, there are two ways to proceed: either carry out a cross-validation using the mean-centred data $\mathbf{X}_0$ and $\mathbf{Y}_0$, or make the mean centring a part of the cross validation; that is, recalculate $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ for each set of left-out objects. For leave-one-out cross-validation with reasonably large data sets, there is unlikely to be much difference unless some samples represent extreme outliers for some variables. Where mean centring has been used in this work, the second cross-validation approach has been used.

Background correction

It is common for various background features to contribute to the complexity of spectra without adding any information about the analyte. While these features can be modelled implicitly by the inclusion of extra factors, it may be preferable to remove them from the data beforehand. This is often achieved by fitting low-order polynomials through regions of the spectra free of absorption bands.

Smoothing and differentiation

Smoothing and differentiation of spectra can be accomplished by the Savitzky-Golay method [85, 86, 87]. The smoothed or differentiated value for each point in the spectrum is calculated by fitting a polynomial through the neighbouring points. Smoothing can enhance the signal-to-noise ratio (SNR) of a spectrum at the expense of resolution: as a band becomes smoother, it also becomes broader. Differentiation is useful for removing baseline terms (taking the first derivative of a spectrum will remove a constant offset; the second derivative will remove a sloping baseline) and for enhancing features such as shoulders.

3.3.4 Outliers

An outlier is an object in the calibration or validation sets (or in prediction) that does not fit in with the others. Outliers may be due to mistakes or instrument errors or may simply be unusual samples, due to, for example, having a particularly high concentration of the analyte or some interferent. Outliers arising from errors should be removed or the error corrected, but those that represent valid measurements on atypical samples may be valuable, particularly in a setting where the properties of the standards cannot be varied at will. In any case, it is important that outliers be identified. Outliers that influence the model strongly usually negatively impact its performance for more typical samples, so should generally be removed. Outliers that have little influence but very large residuals do not harm the model much, but can increase the MSECV dramatically; again, these should be removed. It is important that outliers are also considered in prediction, since much larger errors may be associated with their predicted concentrations than the error variance formula (Section 3.4.2) would suggest.

Leverage

Leverage is a measure of the distance of a sample from the origin of the model space. For factor-based methods (PCR and PLS), the leverage is defined as [49]

$$h_i = 1/n + \mathbf{t}_i^{\mathrm{T}} \left(\mathbf{T}_A^{\mathrm{T}} \mathbf{T}_A \right)^{-1} \mathbf{t}_i \tag{3.37}$$

where $\mathbf{T}$ is the A-factor scores matrix for the calibration set, n is the number of calibration standards and $\mathbf{t}_i$ is the score vector for the sample i. The term $1/n$ is due to mean centring and should be omitted if the data are not mean-centred prior to regression. Furthermore, this term is not always considered part of the leverage: when using formulas that require the leverage it is important to check whether the $1/n$ term is incorporated into the formula explicitly or expected to be present in the leverage. In particular, the prediction variance formula given later (Equation 3.56) accounts for the $1/n$ term independently of the leverage. The leverage is a squared Mahalanobis distance [88] with $\mathbf{T}_A^{\mathrm{T}} \mathbf{T}_A$ as the weighting covariance matrix.

The variation of the leverage with concentration depends on whether the data are mean-centred. If they are not, the origin of the model space corresponds to a vector of zeros; if they are, the origin is $\bar{\mathbf{x}}$, the mean of the calibration spectra. For mean-centred data, samples with particularly low or high concentrations of any component will tend to have higher leverage, while for non-mean-centred data, low concentrations always correspond to low leverage. This phenomenon is illustrated for a one-component system in Figure 3.4. In this simple case, $\mathbf{t} = \mathbf{y}$ and $h_i = y_i^2 / \mathbf{y}^{\mathrm{T}} \mathbf{y}$, where $\mathbf{y}$ is understood to

be mean-centred if mean centring is being used. Since, as shown later, the prediction variance increases with increasing leverage, it could be argued that for trace analysis applications, where precision at the low end of the calibrated range is more important, mean centring may be harmful. Of course, h depends on all species contributing to the spectrum, not just the analyte, so this may be a minor concern.

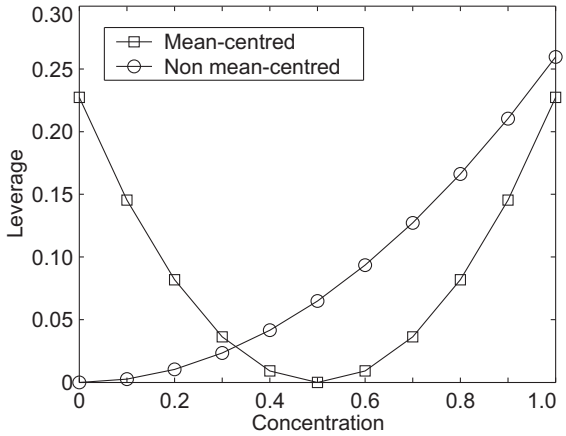

Figure 3.4: Leverage as a function of concentration for a one-component system with and without mean centring. The mean-centring offset $1/n$ is not included.

For objects in the calibration set, the maximum leverage is 1 and the sum of the leverages is $\sum_{i=i}^{n} h_i = A$. If a sample has leverage near one, that means that nearly an entire factor is devoted to the description of that sample. A large leverage does not guarantee a poor standard, simply a unique one: standards with unusually high leverage should be inspected to establish the cause of the high leverage and decide whether the standard should remain in the calibration set.

Outliers from spectroscopic residuals

The spectroscopic residual vector for a sample i is given by

$$\mathbf{e}_i = \mathbf{x}_i - \left(\mathbf{t}_i^{\mathrm{T}}\mathbf{P}_A^{\mathrm{T}}\right)^{\mathrm{T}} \tag{3.38}$$

where $\mathbf{t}_i^{\mathrm{T}}$ is the score vector for the sample, given by $\mathbf{x}^{\mathrm{T}}\mathbf{P}_A$ for PCR and by $\mathbf{x}^{\mathrm{T}}\mathbf{W}_A\left(\mathbf{P}_A^{\mathrm{T}}\mathbf{W}_A\right)^{-1}$ for PLS [49]. Inspection of the residual vector can be informative: in principle, if all the systematic variation in the spectra is modelled, the residuals should be only noise. Structure in the residuals indicates unmodelled variation, which may (for new samples) indicate a new source of variability that was not included in the calibration set. The residual vector can be summarised by its estimated variance (assuming zero mean):

$$s_{e_i}^2 = \frac{\sum_{j=1}^{m} e_{ij}^2}{m - df/n} \tag{3.39}$$

where m is the number of spectroscopic variables and df is the number of degrees of freedom consumed by the model. This quantity can then be compared (via an appropriate F-test; see Appendix A) to the equivalent measure over the entire calibration or validation set:

$$s_e^2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} e_{ij}^2}{nm - df} \tag{3.40}$$

These equations are the same as those given by Martens and Næs [49] but with df being defined as the number of degrees of freedom consumed by, rather than remaining after, the modelling. Since there are correlations between the elements of the residual vector for a given sample, df is actually greater than the expected Am (or $(A + 1)m$ for a mean-centred model); however, Martens and Næs claim that the having a precise estimate of df is not essential for outlier detection. Nevertheless, if df is underestimated, the critical value of the F distribution will be too small and too many objects will be classified as outliers.

Lindgren et al. [89] recommend using $m - A$ and $(m - A)(n - A - 1)$ as the denominators in Equations 3.39, but suggest that these degrees of freedom should be divided by two for calculation of the critical F-value. Without any theoretically justified correction to df, the best approach may be to abandon statistical testing of the F ratios and simply set an arbitrary limit based on the distribution of F ratios observed in the calibration fit or cross-validation. Haaland and Thomas [70] state that, in their experience, for spectra with a few hundred variables, F ratios less than about 3 do not indicate real outliers. Given that $F_{0.95,100,100} = 1.4$ and that the degrees-of-freedom estimates discussed above are likely to be significantly greater than 100 ($n = 20$, $m = 100$ and $A = 5$ are fairly typical values, for example), their experience suggests that the current procedures dramatically overestimate the degrees of freedom associated with the variance estimates.

Outliers from concentration residuals

For the calibration and validation set objects, concentration residuals are also available. The squared residual for any object can be compared to any of the mean-squared error measures (MSEC, MSEP, MSECV) by means of an F-test. The degrees of freedom in this test are 1 (for the individual sample), and an appropriate estimate for the MSE, such as $n - A$ or $n -$ PDF (where PDF is Van der Voet's pseudo-degrees of freedom, given by Equation 3.24) for MSEC; n_{test} for MSEP; or n for MSECV.

Influence plots

A calibration standard with a large residual and high leverage is said to be influential, because its inclusion causes a large change in the model regression coefficients [49]. A standard with high leverage but low residual has relatively little effect on the model; such a standard might represent a higher concentration of the analyte but one still within the linear range. There is an inverse relationship between the leverage and the residual. A useful plot is the "influence plot", in which the sample leverage is plotted against the residual [49] (or, more commonly, the studentised residual; see Appendix A). This plot allows for easy identification of influential samples.

3.3.5 An overall view

It is clear from the preceding sections that model optimisation is a complicated task because of the sheer number of choices involved. Interactions between the various choices make it impossible to isolate them: for example, the optimal set of wavelength ranges may be strongly influenced by an outlier possessing contamination by some compound not present in the other standards. Finally, the large uncertainty in the MSE estimates means that the optimisation process is itself subject to considerable uncertainty. Obviously, a global search over all possible combinations of the various choices is both impractical from a computational point of view and quite likely to recommend a sub-optimal model that just happens, by chance, to have a small value for the validation statistic used.

The process can be simplified by making some choices based on prior knowledge. The general procedure that has been followed in this work is outlined below.

1. Decide on one or a few sensible wavelength ranges, based on what is known about the spectrum of the analyte and other properties of the spectra.

2. For each of several possible pre-treatments:

   (a) Perform a cross-validation (for all ranks up to a sensible limit).

   (b) Look for standards that are persistently identified as outliers, and remove them from the calibration set. If a standard is an outlier in one model but not in another, this may provide a clue as to the reason for it being an outlier: for example, a sample with an unusually large baseline offset may be an outlier with no pre-processing but not an outlier when first-derivative spectra are used. Either correct the problem or remove the outlier for that particular model.

3. Repeat steps 2a–2b until there are no outliers (this should not take more than a few iterations).

4. Choose the best model (wavelength range and pre-treatment) based on it having a small MSECV and a low rank. An F-test can be used to compare the MSECV values for two models to see whether the difference is statistically significant.

5. If a test set is available, validate the best model by test-set validation.

The initial choice of the wavelength ranges can be modified if inspection of the spectroscopic residuals reveals regions of the spectrum that are poorly modelled: these regions may be removed.

3.4 Effects of errors in X and Y

Real data are never free from errors (noise), and it is important to consider the consequences of this in multivariate calibration. Recently, much work on this subject has been published in the chemometrics literature (see, for example, a recent IUPAC review [90]). Two topics will be considered here: the effects of errors in the reference data $\mathbf{Y}$ on the interpretation of validation results; and the propagation of errors in $\mathbf{X}$ and $\mathbf{Y}$ through to the model coefficients $\boldsymbol{\beta}$ and to new predicted concentrations. Two classes of approaches for estimating prediction confidence intervals are discussed: methods based on resampling, and an approximate formula derived recentlyunder the errors-in-variables model.

3.4.1 Errors in y: real and apparent MSEPs

If the method used to generate the reference concentration values is imprecise, the model is trying to find the regression coefficients that match the spectra (assumed error-free) to concentrations that are in error. The most obvious consequence of this is that there will be error associated with the regression coefficients themselves. However, an often more important problem arises during validation. Since the model predictions are being compared to noisy reference values, the calculated MSEP will be larger than the true MSEP [91]. If the model actually predicted the test-set concentrations perfectly (true MSEP = 0), the calculated MSEP would be equal, on average, to the variance in the errors in the reference concentrations. This concept is best illustrated by a simple simulation.

Since the goal is to isolate the effect of errors in $\mathbf{y}$, a single-component system with noise-free $\mathbf{X}$ is used. A vector $\mathbf{y}$ of true concentrations is generated to span the range 0–2 units, and the $\mathbf{X}$ matrix is calculated as in Equation 3.2. Noise (a vector $\Delta\mathbf{y}$ of random numbers drawn from a normal distribution with mean zero and standard deviation $\sigma_{\Delta y}$) is added to $\mathbf{y}$ and a cross-validation conducted, giving a vector of predicted concentrations, $\hat{\mathbf{y}}_{\mathrm{cv}}$. From these predictions, the apparent and true RMSECV values

can be calculated:

$$\mathrm{RMSECV_{app}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i + \Delta y_i - \hat{y}_{i,\mathrm{cv}})^2}{n}} \tag{3.41}$$

$$\mathrm{RMSECV_{true}} = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_{i,\mathrm{cv}})^2}{n}} \tag{3.42}$$

The results are displayed in Figure 3.5. Panel (a), in which the predicted concentrations are plotted against the noisy reference concentrations, depicts a calibration with a moderate amount of scatter about the idea-fit line. However, plotting the predictions against the true concentrations reveals that there is in fact very little scatter. Most of the apparent errors in the plot (a) are due to the errors in the reference y values: the cross-validation predictions are actually closer to the true values than they are to the erroneous ones. The reason for this is that the errors in the reference values are averaged out in the regression procedure [91].

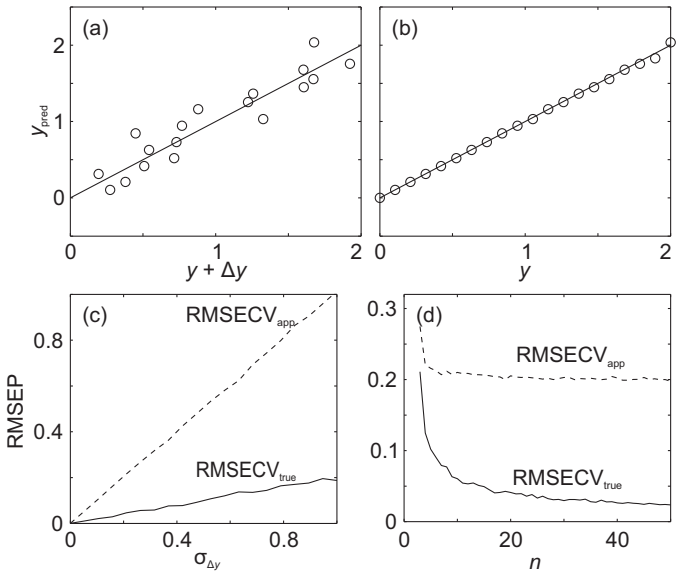

Figure 3.5: Cross-validation statistics for a single-component system with noise-free **X** and homoscedastic errors in **y**. (a) Predicted y plotted against noisy reference values $y + \Delta y$. (b) Predicted y plotted against true y. (c) Variation of the apparent (calculated with respect to $y + \Delta y$) and true (calculated with respect to y) RMSECV with regard to the standard error $\sigma_{\Delta y}$ in y; number of standards $n = 20$. (d) Variation of the apparent and true RMSECV with regard to the number of calibration standards; $\sigma_{\Delta y} = 0.2$. In (a) and (b) $\sigma_{\Delta y} = 0.2$ and $n = 20$.

Plots (c) and (d) show how the RMSECV values are affected by the noise level and the number of calibration standards. The apparent RMSECV increases linearly with increasing standard error of the y values, as shown in panel (c). The true RMSECV also increases linearly, but with a much smaller slope. The ratio of the slopes depends on the number of calibration standards: as seen in panel (d), $\mathrm{RMSECV_{app}}$ decreases as n is increased, asymptotically approaching $\sigma_{\Delta y}$; however, $\mathrm{RMSECV_{true}}$ is

proportional to $1/n$ and continues to decrease below $\sigma_{\Delta y}$.

Because of the potential for the apparent and true RMSECV values to be so different, it is important to obtain an estimate of the uncertainty in the reference values. If the apparent RMSECV is close to the standard error of the reference values, the model is probably performing adequately and the best way to improve the apparent RMSECV is to obtain better reference values for the samples used for validation. Alternatively, if better reference values are not available but an estimate, $\hat{\sigma}_{\Delta y}$, of the reference-method error is, the apparent RMSECV can also be corrected for the effect of the errors in the reference values [74, 91]:

$$\widehat{\text{RMSECV}}_{\text{true}} = \sqrt{\text{RMSECV}_{\text{app}}^2 - \hat{\sigma}_{\Delta y}^2} \tag{3.43}$$

However, the accuracy of this procedure depends on the accuracy of $\hat{\sigma}_{\Delta y}$. If $\hat{\sigma}_{\Delta y}$ and $\text{RMSECV}_{\text{app}}$ are of similar magnitude, the proportional error in $\widehat{\text{RMSECV}}_{\text{true}}$ may be very large; and negative RMS errors are not out of the question. This embarrassment can be avoided by calculating a confidence interval for $\hat{\sigma}_{\Delta y}$ and using the lower bound in Equation 3.43.

3.4.2 Confidence intervals for prediction

For an analysis result to be useful, it must be presented along with an estimate of its uncertainty. The average prediction uncertainty associated with a model is approximated by RMSECV or RMSEP obtained during model validation. However, the prediction uncertainty associated with any particular sample will generally differ from the average value. In particular, samples that lie far from the centre of the model space will have greater uncertainty than the average. The position of a sample in the model space depends on the spectrum as a whole, not just the portion due to the analyte.

There are two basic approaches to estimating the uncertainty involved with a prediction [55]: re-sampling and approximate formulas. Advantages to resampling methods are that they rely less on assumptions about the errors in the data and that they require no additional experimental effort. However, they can involve a considerable amount of computation time, whereas formulas are usually easy to evaluate. Another advantage of formulas is that they can offer more insight into which sources of error predominate.

Jack-knife

The jack-knife [92] is a resampling method very closely related to cross-validation. Essentially, it is a generalisation of the cross-validation methodology enabling estimation of the uncertainty associated

with any parameter. In the specific case of estimating the standard error s_{y_u} associated with a single new prediction, the procedure is:

1. Calculate the regression vector $\boldsymbol{\beta}$ and predicted concentration $\hat{y}_u$ as normal.
2. For every object i in the training set:
   (a) Temporarily delete object i from the calibration set;
   (b) Generate the regression vector $\boldsymbol{\beta}_i$ by MLR, PCR, or PLS;
   (c) Calculate and store the predicted concentration $\hat{y}_{u,i} = \mathbf{x}_u^T \boldsymbol{\beta}_i$;
   (d) Restore object i to the calibration set, increment i, and repeat.
3. The uncertainty is given by

$$\hat{s}_{y_u}^{JK} = \left(\frac{n-1}{n} \frac{\sum_{i=1}^{n} (\hat{y}_{u,i} - \bar{\hat{y}}_u)^2}{n} \right)^{1/2} \tag{3.44}$$

   where $\bar{\hat{y}}_u$ is the mean of the jack-knifed predictions, and can probably be replaced by the actual prediction $\hat{y}_u$ [92].

Bootstrap

There are two versions of the bootstrap commonly used in chemometrics: bootstrapping objects and bootstrapping residuals [93]. The object bootstrap is somewhat similar to the jack-knife, but differs in the way in which the new calibration sets are generated. Rather than deleting standards one at a time from the calibration set, new calibration sets are generated by sampling randomly with replacement from the original calibration set. The number of bootstrap replications, B, is chosen to be large enough to obtain stable estimates, and the estimated standard deviation of the prediction error is just the standard deviation of the bootstrap predictions:

$$\hat{s}_{y_u}^{BS} = \left(\frac{\sum_{i=1}^{B} (\hat{y}_{u,i} - \bar{\hat{y}}_u)^2}{B-1} \right)^{1/2} \tag{3.45}$$

Bootstrapping residuals is quite a different approach:

1. Calculate the regression vector and fit residuals from the full calibration set. The residuals are given by

$$\mathbf{e} = \frac{\mathbf{y} - \mathbf{X}\boldsymbol{\beta}}{1 - df/N} \tag{3.46}$$

   where df is an appropriate estimate of the number of degrees of freedom.

2. For each bootstrap replicate i,
   (a) Generate a bootstrap residual vector $\mathbf{e}_i^{\mathrm{BS}}$ by sampling with replacement from the residual vector $\mathbf{e}$;
   (b) Calculate a bootstrap calibration-set concentration vector: $\mathbf{y}_i^{\mathrm{BS}} = \mathbf{y} + \mathbf{e}_i^{\mathrm{BS}}$;
   (c) Determine the regression vector $\boldsymbol{\beta}_i^{\mathrm{BS}}$ using $\mathbf{X}$ and $\mathbf{y}_i^{\mathrm{BS}}$ as the calibration set;
   (d) Predict and store the concentration for the unknown sample: $\hat{y}_{\mathrm{u,i}}^{\mathrm{BS}} = \mathbf{x}_{\mathrm{u}}^{\mathrm{T}}\boldsymbol{\beta}_i^{\mathrm{BS}}$.
3. The uncertainty is given by

$$\hat{s}_{y_{\mathrm{u}}}^{\mathrm{BS}} = \frac{1}{n}\left(\frac{\sum_{i=1}^{B}(\hat{y}_{\mathrm{u,i}} - \bar{\hat{y}}_{\mathrm{u}})^2}{B-1}\right)^{1/2} \tag{3.47}$$

Noise addition

The noise addition method [93] is extremely similar to the residual bootstrap, except that artificial noise is added to $\mathbf{y}$, rather than random samples of the calibration residuals. The advantage of this approach is that any desired noise structure can be used, including correlation and heteroscedasticity.

Approximate prediction error variance formula

An approximate formula for the prediction error variance in MLR, PCR, and PLS has been derived by Faber and Kowalski [74], building on the earlier work of Höskuldsson [67], Phatak et al. [94] and Denham [95]. The formula takes into account errors in both $\mathbf{X}$ and $\mathbf{y}$, and is derived on the basis of the "error in variables" model. In this section it is important to distinguish between true quantities, measured quantities (indicated by a tilde, e.g. $\tilde{\mathbf{x}}$), and estimated or predicted quantities, (indicated by a circumflex, e.g. $\hat{y}_{\mathrm{u}}$). Mean centring is assumed, but the modifications necessary for non-mean-centred models are straightforward.

The model for the calibration, in terms of the true spectra and concentrations, is written

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{3.48}$$

where $\mathbf{y}$ is a vector of true concentration values, $\mathbf{X}$ is a matrix of true spectra, and $\mathbf{e}$ is a vector of residuals (accounting for error in the model, such as deviations from Beer's law). The measured values for the concentrations and spectra are given by

$$\tilde{\mathbf{y}} = \mathbf{y} + \Delta\mathbf{y} \tag{3.49}$$

$$\tilde{\mathbf{X}} = \mathbf{X} + \Delta\mathbf{X} \tag{3.50}$$

and the model can now be written

$$\tilde{\mathbf{y}} = \tilde{\mathbf{X}}\boldsymbol{\beta} - \Delta\mathbf{X}\boldsymbol{\beta} + \mathbf{e} + \Delta\mathbf{y} \tag{3.51}$$

The regression vector $\hat{\boldsymbol{\beta}}$ is estimated by PLS from the noisy spectra and concentrations, and differs from the true regression vector:

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \Delta\boldsymbol{\beta} \tag{3.52}$$

In the prediction step, a new measured spectrum $\tilde{\mathbf{x}}_u = \mathbf{x}_u + \Delta\mathbf{x}_u$ is multiplied by the regression vector to give a new estimated concentration.

$$\hat{y}_u = \bar{\tilde{y}} + \tilde{\mathbf{x}}_u\hat{\boldsymbol{\beta}} \tag{3.53}$$

The true unknown concentration is

$$y_u = \bar{y} + \mathbf{x}_u\boldsymbol{\beta} + e_u \tag{3.54}$$

and the prediction error is defined as

$$\begin{aligned} \mathrm{PE}_u &= \hat{y}_u - y_u \\ &= \bar{\tilde{y}} - \bar{y} + \tilde{\mathbf{x}}_u\boldsymbol{\beta} - \mathbf{x}_u\boldsymbol{\beta} - e_u \\ &\approx \Delta\bar{y} + \Delta\mathbf{x}_u\boldsymbol{\beta} + \mathbf{x}_u\Delta\boldsymbol{\beta} - e_u \end{aligned} \tag{3.55}$$

where $\hat{\boldsymbol{\beta}}$ is expanded as $\boldsymbol{\beta} + \Delta\boldsymbol{\beta}$ and products of error terms are ignored.

From Equation 3.55 it can be seen that the error in the predicted concentration arises from several terms: the error in the mean centring, the error in the new measured spectrum, and the error in the regression vector. The error in the regression vector, $\Delta\boldsymbol{\beta}$, arises from the error in the calibration set spectra and concentrations, and is calculated by determining how the measurement errors propagate through the calculation of the regression vector. Accurate formulas are possible for MLR and PCR, but because the PLS estimation of the regression vector is nonlinear, no closed-form expression is possible, and an approximation based on a linearisation of the algorithm can be used (called the first-order approximation). A further approximation is just to use the same formula as for PCR, neglecting the nonlinearity entirely (zeroth-order approximation). Details will not be given here: for a thorough treatment see Faber and Kowalski [74].

Equation 3.55 gives the prediction error arising from a given set of errors $\Delta\mathbf{y}$ and $\Delta\mathbf{X}$. Of course, the actual values of the errors are never known: all that can be known is their statistical distribution. Usually, for each noisy variable, the errors are assumed to follow a normal distribution with zero mean

and the variance is estimated by replicate measurements. The variance of the prediction error, $V(\mathrm{PE})_\mathrm{u}$, can then be estimated.

The zeroth-order approximation to the variance $V(\Delta\boldsymbol{\beta})$ in the regression vector and the assumption of homoscedastic noise in both $\mathbf{X}$ and $\mathbf{y}$ lead to the following expression for the prediction error variance:

$$\begin{aligned} V(\mathrm{PE_u}) &= E(\mathrm{PE_u^2}) \\ &\approx (n^{-1} + h_\mathrm{u})(\sigma_e^2 + \sigma_{\Delta y}^2 + \|\boldsymbol{\beta}\|^2\sigma_{\Delta X}^2) + \sigma_e^2 + \|\boldsymbol{\beta}\|^2\sigma_{\Delta X}^2 \end{aligned} \tag{3.56}$$

where $E(\mathrm{PE_u^2})$ is the expected value of the squared prediction error. According to Faber and Kowalski, this approximation is reasonable when almost all of the systematic variation in the measured spectra must be included in the model to obtain good predictions (a common case), or when very little of the systematic variation is important for prediction. The advantage of the zeroth-order expression is its interpretability in terms of the relative contributions of the various sources of error. Also implicit in Equation 3.56 are the assumptions that the errors in $\mathbf{y}$ and $\mathbf{X}$ are independently and identically distributed (*iid*) and that σ_e^2 is the same for new samples as for the calibration standards. Equivalent formulas but accounting for heteroscedastic, correlated noise in $\mathbf{y}$ and $\mathbf{X}$ can be derived from the treatment in Ref. [74].

Equation 3.56 includes three parameters that must be estimated somehow: $\sigma_{\Delta X}^2$ and $\sigma_{\Delta y}^2$ are the variances of the measurement errors in the spectroscopic and concentration variables (assuming that each sspectroscopicvariable has the same uncertainty), which may be determined by replications of the concentration and spectroscopic measurements; σ_e^2 is the variance of the residuals. A method for the estimation of this last parameter is given by Faber *et al* [96].

In the absence of bias, the mean squared prediction error (MSEC) is given by

$$\mathrm{MSEC} = \sigma_e^2 + \sigma_{\Delta y}^2 + \|\boldsymbol{\beta}\|^2\sigma_{\Delta X}^2 \tag{3.57}$$

MSEC can also be estimated from the calibration residuals:

$$\widehat{\mathrm{MSEC}} = \frac{1}{n - A - 1}\sum_{i=1}^{n}(\hat{y}_i - \tilde{y}_i)^2 \tag{3.58}$$

The factor $1/(n - A - 1)$ accounts for the degrees of freedom consumed by the estimation of model parameters. The expression is correct as given for PCR, but since PLS factors are constructed using both $\mathbf{X}$ and $\mathbf{y}$, a further correction may be necessary (for example, Van der Voet's pseudo-degrees of

freedom [75] could be used). Comparing Equations 3.57 and 3.58, an estimate for the residual variance is given by

$$\hat{\sigma}_e^2 = \widehat{\text{MSEC}} - \hat{\sigma}_{\Delta y}^2 - \|\hat{\boldsymbol{\beta}}\|^2 \hat{\sigma}_{\Delta X}^2 \tag{3.59}$$

It should be noted that if bias is present, it will inflate the estimate $\hat{\sigma}_e^2$ in Equation 3.59, which is supposed to be purely a variance measure. In this case, some bias will be present in the prediction error, so this inflation guards against underestimating the true prediction error.

Other authors [97] have recommended using $\widehat{\text{MSEC}}$ as a direct estimate of the residual variance. The above approach tries to account for the contribution that the errors in $\mathbf{X}$ and $\mathbf{y}$ make to $\widehat{\text{MSEC}}$. Obviously, if $\widehat{\text{MSEC}}$ is used directly, the prediction errors will be biased high.

Confidence intervals

Once the standard error of the new prediction has been estimated, by any of the above methods, it can be used to calculate a $(1-\alpha)\times 100\%$ confidence interval by assuming that the error in the prediction is normally distributed:

$$\hat{y}_u - t_{1-\alpha/2,\nu}\hat{\sigma}(\text{PE}_u) < y_u < \hat{y}_u + t_{1-\alpha/2,\nu}\hat{\sigma}(\text{PE}_u) \tag{3.60}$$

where ν is the number of degrees of freedom associated with $\hat{\sigma}(\text{PE}_u)$. Since $\hat{V}(\text{PE}_u)$ is a sum of several variance estimates (Equation 3.56), each with its own number of degrees of freedom, $\hat{V}(\text{PE}_u)$ is called a complex variance estimate. Its number of degrees of freedom may be estimated by Satterthwaite's rule [98] (see Appendix A).

3.4.3 Detection limits

The term "detection limit" has numerous definitions, with the most common probably being the "three times the baseline noise" rule of thumb. This rule states that if a measurement exceeds the noise level by a factor of three or more, then the sample is unlikely to be a blank. Statistically, if the noise is normally distributed and the "noise level" used in the estimate is its standard deviation, only ~1 % of blank measurements will exceed three times the noise level. Therefore, this definition of the detection limit provides protection against false positive detection decisions. However, there is another class of possible errors: a false negative decision occurs when a sample with a concentration higher than the stated detection limit returns an analytical result lower than the detection limit. IUPAC recommends [90] a definition of the detection limit [99] which also provides protection against false negative errors. The earlier definition of the detection limit is renamed the "critical level", L_C, and formally defined as being

the predicted analyte concentration above which a proportion α of blank measurements will fall. This concept is illustrated in Figure 3.6.

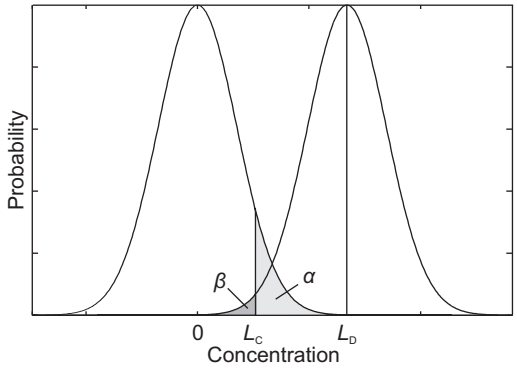

Figure 3.6: Graphical representation of the critical level, L_C, and the detection limit, L_D. The light grey area represents false positive detection decisions (with respect to L_C); the dark grey area represents false negative decisions.

The detection limit, L_D, is then defined as the true sample concentration that with probability β will lead to a predicted concentration $< L_C$. The value of L_C will depend on the distribution of errors for blank samples, while that of L_D will depend both on L_C and on the distribution of errors for the real sample measurement.

There are some complications in applying this methodology in multivariate calibration, as discussed by Boqué et al. [97] (see also Refs 100 and 101 for direct calibration models and Refs 90 and 102 for general reviews). Essentially, these problems arise because the uncertainty in the analyte concentration, y, is not simply a function of y. As shown above, the key quantity in determining the uncertainty is the leverage, h (Equation 3.37). While h depends on the concentration of the analyte, it also depends on the concentrations of the interferents as well as any instrumental artefacts. The most obvious consequence is that the detection limit can only be specified *per sample*, which is at odds with the usual concept of the detection limit as a property of an analytical *method*. (It will be shown in Chapter 7 how a method-specific detection limit can be obtained within certain assumptions.) A slightly more subtle problem is that, in multivariate calibration, the error distribution of the "blank" sample is unknown: merely defining a blank as having zero analyte concentration is insufficient to define its error distribution. Boqué et al. [97] recommend calculating a per-sample "blank" spectrum, for the α level chosen, by manipulation of the sample spectrum; details will not be given here.

A simpler approach is to take the prediction uncertainty of the blank is as being the same as the prediction uncertainty of the sample spectrum. In this case, the estimated detection limit will be biased upward by the contribution of the analyte spectrum to the leverage. However, since the detection limit is most important for samples with low analyte concentration, for which the bias will be small, this

approach is probably reasonable.

With the above assumption, the following equations can be used to calculate the detection limit:

$$\begin{aligned} L_C &= \hat{\sigma}(PE_u)t_{1-\alpha,\nu} \\ L_D &= L_C + \hat{\sigma}(PE_u)t_{1-\beta,\nu} \end{aligned} \tag{3.61}$$

where $\hat{\sigma}(PE_u)$ is an estimate of the prediction error (with ν degrees of freedom) and the t statistics are single-tailed.

3.4.4 Heteroscedastic errors in y

Ordinary least-squares regression gives the best ("maximum likelihood") parameter estimates if the error structures of the independent and dependent variables meet certain conditions [103]. Specifically, the errors in the dependent variables must be normally distributed, and, between observations, must be independently and identically distributed (*iid*). The errors in the independent variables must be negligible in comparison to those in the dependent variables.

If the error structures do not meet the requirements listed above, certain weighted regression methods can, in some cases, be used to obtain the maximum likelihood solution [103]. In the case where the errors in the independent variable are negligible but the dependent-variable errors are not *iid*, it is possible to transform the calibration set data, as described in Appendix A, so that they fulfill the assumption of *iid* errors in $\mathbf{y}$. For errors uncorrelated between observations this is achieved simply by dividing each concentration y_i and the corresponding row of $\mathbf{X}$, $\mathbf{x}_i$, by $\sigma_{\Delta y_i}$. After the transformation, the unweighted regression algorithm can be used as normal.

Another major consequence of heteroscedastic errors in $\mathbf{y}$ is that the apparent MSEP (or MSECV) is now a function of the concentrations of the standards used to determine it. In the common case of proportional errors (in which $\sigma_{\Delta y} = \gamma y$), if the test set consists mostly of low-concentration standards, the MSEP will be small, whereas if the test set consists mostly of high-concentration standards, the MSEP will be large. Assuming a perfect model, the expected squared apparent error for a single test-set object i is equal to the error variance of the reference value y_i:

$$E\left[(\hat{y}_i - y_i)^2\right] = \sigma^2_{\Delta y_i} = \gamma^2 y_i^2 \tag{3.62}$$

For an infinite test set with y having a probability density function $f(y)$ the apparent MSEP (again, for

a perfect model) is given by

$$\mathrm{MSEP}^0_{\mathrm{app}} = \gamma^2 \int_0^\infty y^2 f(y)\mathrm{d}y \tag{3.63}$$

and for a finite test set, the integral is replaced by a sum:

$$E(\mathrm{MSEP}^0_{\mathrm{app}}) = \frac{\gamma^2 \sum_{i=1}^N y_i^2}{N} \tag{3.64}$$

These equations can in principle be used to correct apparent RMSEPs, as in Equation 3.43, but no examples were found in the literature.

The performances of weighted and unweighted regression with data having proportional errors in the reference **y** vector are compared in Figure 3.7. For simplicity, univariate data with no errors in the independent variable were used. Calibration sets with n objects were generated, and noise with standard deviation $\sigma_{\Delta y} = \gamma y$ was added to the concentration vectors. PLS regression models were constructed, with and without object weighting, and used to predict concentrations for a large test set. The predicted concentrations were compared to the true test-set concentrations to obtain $\mathrm{RMSEP}_{\mathrm{true}}$, and with concentrations which have themselves been corrupted with noise (in the same manner as the calibration set) to obtain $\mathrm{RMSEP}_{\mathrm{app}}$.

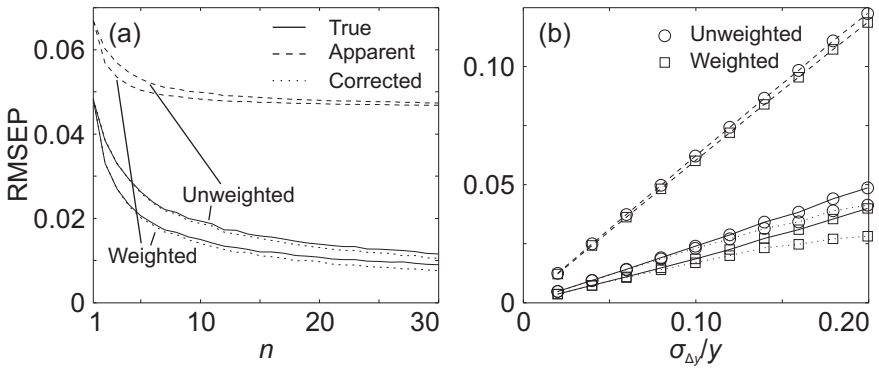

Figure 3.7: Effect of weighting in PLS regression with proportional errors in **y**. True, apparent and corrected RMSEPs are plotted with solid, dashed, and dotted lines, respectively. Each point is the average of 5000 Monte-Carlo replicates; $n_{\mathrm{test}} = 1000$. In (a), the number of calibration standards is varied and $\sigma_{\Delta y}/y = 0.08$. In (b), $\sigma_{\Delta y}/y$ is varied and $n = 10$; circles and squares denote unweighted and weighted regression, respectively.

The results are plotted in Figures 3.7a and b. In general, a small improvement can be seen with the use of weighted regression. With the exception of the case where there is only a single standard, the improvement in the true RMSEP due to weighted regression decreases slowly as the number of standards is increased. In Figure 3.7b, it can be seen that the correction based on Equations 3.43 and 3.64 is accurate when the errors are small, but overcorrects when the errors are large.

3.5 A way to mislead oneself with cross-validation

Often, it may be possible to make more than one measurement of a single standard, and the question arises as to how best to make use of the additional information present in the replicate measurements. The amount of extra information actually present is dependent on the type of replicate measurement: for example, if a UV absorption spectrum of a solution is measured twice in quick succession, the two spectra are likely to be essentially identical. The only new information that can be extracted from the second spectrum relates to the noise level. If the cuvette is removed and re-inserted, then information is obtained about the reproducibility of that process. If a significant period elapses between the two measurements, then information is gained about the stability of the instrument and the sample. If the measurement is made on a different day, by a different operator, on an independently prepared solution of the same nominal concentration (and so forth), then the second measurement constitutes a true statistical repeat of the first measurement, because all possible sources of variance are allowed [66].

For the data sets described in this thesis, several somewhat independent measurements are made from each standard. The sampling head is repositioned between measurements to investigate different regions of the sample, but measurements are made in quite rapid succession and all share a single background spectrum, so it is unclear to what extent the contributions from species other than the analyte (absorption by atmospheric gases and instrumental artefacts) will be correlated.

If each spectrum were independent of the others, it would be valid to perform a leave-one-out cross-validation in which a single spectrum is omitted each time ("leave-one-spectrum-out"). However, if there is correlation between the various components contributing to the spectra for each standard, such a cross-validation can produce an erroneously low estimate of the prediction error. In this situation, all the spectra from a single standard should be treated as a group and left out or retained together in cross-validation ("leave-one-standard-out").

The effect of correlation between the concentrations of interfering species in replicate measurements can be demonstrated easily. In this simulation, there are eight species in addition to the analyte. The spectrum of each was a vector of 100 random numbers. The concentration of the analyte was varied linearly from 0 to 1 unit in ten steps. For each value of the analyte concentration, eight replicate spectra were calculated. For each replicate, the concentration of each interferent was determined randomly, but in such a way that the extent of the correlation between the concentrations of a particular interferent in all the replicates for each sample was controlled.

This was achieved by constructing an appropriate correlation matrix (see Appendix A.1.1) for the concentrations of each interferent. The matrix resembled that in Equation 3.65 below, except that the

blocks along the diagonal were 8×8.

$$\mathbf{R}(\mathbf{C}_{\text{int}}) = \begin{bmatrix} 1 & R & R & 0 & 0 & 0 & \\ R & 1 & R & 0 & 0 & 0 & \\ R & R & 1 & 0 & 0 & 0 & \\ 0 & 0 & 0 & 1 & R & R & \\ 0 & 0 & 0 & R & 1 & R & \\ 0 & 0 & 0 & R & R & 1 & \\ & & & & & & \ddots \end{bmatrix} \tag{3.65}$$

The covariance matrix for the interferent concentration was obtained by multiplying the correlation matrix by the square of the relative concentration of the species. The relative concentrations of the interfering species ranged from ~4 times to ~0.2 times the average concentration of the analyte. Finally, the actual concentrations of the interferent were obtained by premultiplying a vector of normally distributed random numbers by the matrix square root (obtained by a Cholesky decomposition [61]) of the covariance matrix. This transforms the *iid* random numbers into ones having the specified correlation properties. This process was repeated for each interferent, and the calibration set spectra were constructed as in Equation 3.2 by multiplying the resulting concentration matrix by the random spectroscopic matrix.

A large test set ($n_{\text{test}} = 1000$) was constructed in a similar way (with evenly spaced analyte concentrations between zero and one), but with only a single spectrum per sample. For several values of R, the two types of cross-validation and a test-set validation were carried out, and the whole procedure was repeated 1000 times to give stable results.

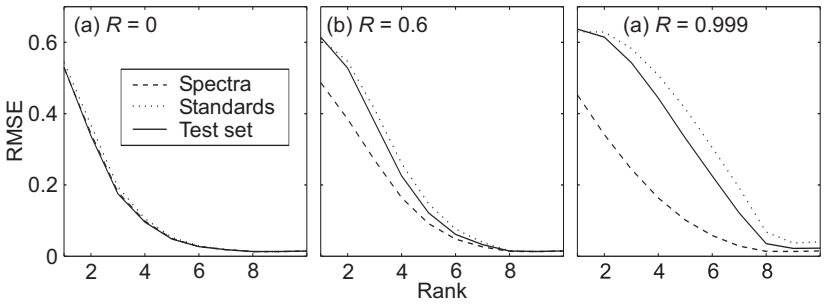

Figure 3.8: Comparison of validations with imperfect repeat points. The solid lines are for validation with a large test-set; the dashed and dotted lines represent leave-one-spectrum-out and leave-one-standard-out cross-validations, respectively. In each case there are ten standards each with eight spectra, and eight interfering species. (a) No "per-standard" correlation between interfering species ($R = 0$); (b) $R = 0.6$; (c) $R = 0.999$.

The results of the simulation (Figure 3.8) clearly illustrate that, while the two sorts of cross-validation are basically equivalent when there is no correlation of the interfering components, even a weak correlation, as in panel (b), leads to a substantial under-estimation of the error by the leave-one-spectrum-out cross validation. The leave-one-standard-out cross validation gives an overestimate of the error, but is clearly much closer to the true error even in the case when correlation is very strong (c). For this reason, and also to present conservative measures of the quality of a model, the leave-one-sample-out procedure has been preferred in this thesis. A more straightforward demonstration of this phenomenon with experimental data is given in Chapter 6.

Chapter 4

Experimental

This chapter provides background to and describes the experimental methods referred to elsewhere in this thesis.

The first section briefly discusses the principles of Fourier transform infrared (FTIR) spectroscopy in general, following the treatment of Griffiths and de Haseth [104]. The FTIR system used in this work consists of a Bruker Vector 22 spectrometer with a prototype Remspec SpotView fibre-optic grazing-angle accessory coupled to the external beam port. For spectrometer control and data acquisition the manufacturer's software, OPUS [105], has been used. The fibre-optic probe is described in detail in Section 4.2, where a very brief review of infrared fibre optics in general is also given.

The main target application of this IRRAS system is in pharmaceutical cleaning validation (see Chapter 1), where it has the potential to complement or replace the existing wet-chemistry methods. The first experimental step in calibration is to prepare calibration standards, which must strike a balance between being relevant to the target samples and being sufficiently well characterised for the calibration model to be evaluated. Methods for preparing suitable standards have been developed (in tandem with Michelle Hamilton [106]) and are discussed in Section 4.3.

4.1 FTIR spectrometry

4.1.1 Interferograms and spectra

Fourier transform spectrometers function in a radically different way from dispersive instruments. A dispersive spectrometer physically separates the wavelengths of light from a broadband source, usually by a diffraction grating, and uses a slit to select a narrow beam corresponding to a small range of wavelengths. The selected wavelength is varied by rotating the grating and a spectrum is measured directly in the frequency domain. The trade-off between resolution and signal intensity is obvious: a

narrow slit improves the resolution but at the expense of the intensity of light reaching the detector and therefore the signal to noise ratio (SNR).

At the heart of an FTIR spectrometer is an interferometer rather than a diffraction grating. A simple interferometer is illustrated in Figure 4.1. The three optical components are a fixed mirror, a moveable mirror and a beamsplitter. Light from the source is incident on the beamsplitter: a portion is transmitted (towards the moveable mirror) and the rest is reflected (towards the fixed mirror). The two beams are reflected by the mirrors and return to the beamsplitter. The beam returning from the moveable mirror will have travelled a different path length than the beam returning from the fixed mirror, the difference, δ, (called the retardation) being twice the mirror displacement. For a monochromatic source with wavelength λ, a retardation of $\delta = n\lambda$, where $n = 0, 1, 2, \ldots$, results in the two beams interfering constructively and all of the incident radiation passing to the detector. Contrastingly, if $\delta = (n + 1/2)\lambda$, the two beams will interfere destructively and all the radiation will return in the direction of the source.

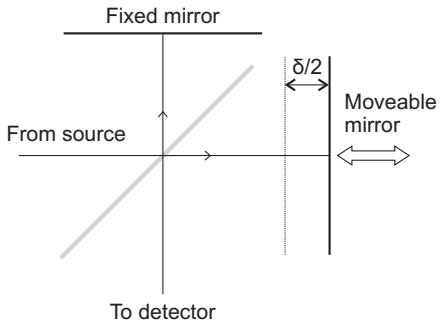

Figure 4.1: A Michelson interferometer. The thick grey line is the beamsplitter.

An interferogram is collected by measuring the intensity at the detector as a function of the retardation. If the radiation is monochromatic, the intensity at the detector is given by

$$I'(\delta) = \frac{B(\bar{\nu})}{2}(1 + \cos 2\pi\bar{\nu}\delta) \tag{4.1}$$

where $\bar{\nu} = 1/\lambda$ is the wavenumber, in the inverse units of δ. In addition to the intensity of the source, $B(\bar{\nu})$ is understood to include the effects of all factors contributing to the beam intensity, such as the non-ideality of the beamsplitter, the detector response, and the amplifier characteristics. The interferogram is the modulated component of I',

$$I(\delta) = 0.5B(\bar{\nu}) \cos 2\pi\bar{\nu}\delta \tag{4.2}$$

For source spectra consisting of several narrow lines, the interferogram is simply the sum of the interferograms due to the individual lines. For a broadband source, the interferogram is given by an

integral over all wavenumbers:

$$I(\delta) = \int_{-\infty}^{\infty} B(\bar{\nu}) \cos 2\pi\bar{\nu}\delta \, \mathrm{d}\bar{\nu} \tag{4.3}$$

This is one part of a cosine Fourier transform pair. The complementary part is

$$B(\bar{\nu}) = \int_{-\infty}^{\infty} I(\delta) \cos 2\pi\bar{\nu}\delta \, \mathrm{d}\delta \tag{4.4}$$

in which the spectrum $B(\bar{\nu})$ is calculated from the interferogram by taking the Fourier transform.

At the point of zero displacement (ZPD; $\delta = 0$), the interference is constructive for all wavelengths, so the intensity is very high. For this reason, the region around the ZPD is called the centreburst. FTIR spectrometers generally offer several options affecting the way the interferogram is measured. A single-sided interferogram has a small number of samples to one side of the ZPD and extends to $\delta_{\max}$ on the other, whereas a double-sided interferogram extends from $-\delta_{\max}$ to $\delta_{\max}$. A second double-sided interferogram can also be measured as the mirror is returned from $\delta_{\max}$ to $-\delta_{\max}$. This last mode of operation, referred to as "double-sided forward-backward" in the OPUS software, is the most efficient in terms of number of interferometer scans per unit time, and is the method used in this work.

4.1.2 Apodisation and resolution

Because the mirror can only be moved over a finite range, only a portion of the complete interferogram can be measured. This leads to distortions in the calculated spectrum, and also limits its resolution. Limiting the retardation is equivalent to multiplying the complete interferogram by a top-hat (or box-car) function, which is equal to one for $-\delta_{\max} \leq \delta \leq \delta_{\max}$ (assuming a double-sided interferogram) and to zero elsewhere. By the convolution theorem [107], the Fourier transform of the product of two functions is the convolution of their Fourier transforms. The calculated spectrum is therefore the convolution of the true spectrum and the Fourier transform of the top-hat function (which is a sinc function, $\sin(x)/x$ [107]). The effect of truncation of the interferogram is illustrated in Figure 4.2. Convolving the line in panel (a) with the sinc function due to the top-hat results in a sinc function centred on the frequency of the line, as shown in panel (c). The width of the sinc function is the inverse of the maximum retardation, so two lines can be resolved if they are separated by $\Delta\bar{\nu} \geq 1/\delta_{\max}$ [104].

Obviously, nothing can be done about the loss of resolution other than to increase the maximum retardation. However, the distortion of the lineshape can be reduced, at the cost of further broadening, by apodisation. In this process, the interferogram is multiplied by an apodisation function, $A(\delta)$, which goes to zero at $-\delta_{\max}$ and $\delta_{\max}$ and essentially serves to soften the edges of the top-hat function. A

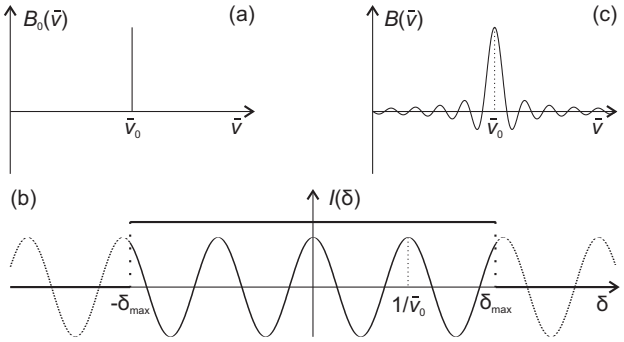

Figure 4.2: Schematic illustration of the effect of truncating the interferogram. (a) Spectrum of a monochromatic source. (b) Interferogram corresponding to (a). The dark horizontal lines represent the top-hat function (the nonzero portion is at $I = 1$) and the dashed part of the cosine is the portion of the interferogram lost by truncation. (c) Spectrum calculated from the interferogram in (b) after truncation. The width of the main peak at its base is $1/\delta_{\max}$.

simple example is the triangular function, defined as

$$A_\Lambda = \begin{cases} 1 - |\delta/\delta_{\max}| & -\delta_{\max} \leq \delta \leq \delta_{\max} \\ 0 & \text{otherwise} \end{cases} \tag{4.5}$$

With triangular apodisation, the largest side lobe is reduced from ~21 % of the peak maximum to ~4.5 %, but the peak width is ~1.5 times greater [104]. While this is generally a welcome trade, the triangular apodisation function is not optimal and a variety of alternatives are more commonly used [104]. In all the work described in this thesis, the Blackmann-Harris three-term apodisation function has been used. This function is given by [108]

$$A_{\mathrm{BH}} = 0.42 + 0.5\cos\left(\frac{\pi\delta}{\delta_{\max}}\right) + 0.08\cos\left(\frac{2\pi\delta}{\delta_{\max}}\right) \tag{4.6}$$

for $-\delta_{\max} \leq \delta \leq \delta_{\max}$ and 0 otherwise.

4.1.3 Phase errors and correction

In general, there will be wavenumber-dependent phase errors in the measured interferogram. These can arise from sampling error with respect to the ZPD and phase lag introduced by electronic filters [104]. Incorporating the possibility of a phase error $\theta_{\bar{\nu}}$ into Equation 4.3,

$$I(\delta) = \int_{-\infty}^{\infty} B(\bar{\nu})\cos\left(2\pi\bar{\nu}\delta - \theta_{\bar{\nu}}\right)\,\mathrm{d}\bar{\nu} \tag{4.7}$$

This has the effect of adding sine components to the cosine wave interferogram. The cosine Fourier transform of a truncated sine wave is a derivative-shaped band, so the presence of sine components in the interferogram causes distorted peaks in the spectrum. For this reason, phase correction is always applied. In this work, Mertz phase correction [104] was used.

4.1.4 Effect of optical divergence

Ideally, the beam entering the interferometer would be perfectly collimated, but this would require a perfect point source. An extended source results in a beam that contains a range of angles. A ray diverging from the ideal path will travel a greater distance inside the interferometer and consequently will incur a greater path difference (see Figure 4.3).

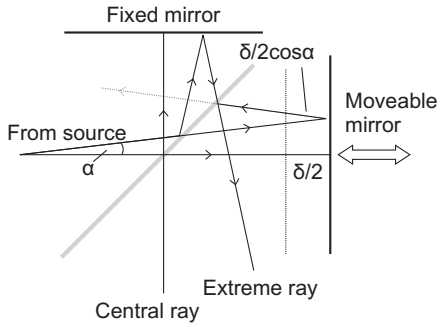

Figure 4.3: Michelson interferometer showing the effect of beam divergence. The path difference for the central ray is δ, while that for the extreme ray is $\delta/\cos\alpha$.

Beam divergence is problematic because a ray on a divergent path will produce an interferogram that is equivalent to a central ray with a longer wavelength. This can be seen by imagining a particular retardation δ of the central ray. The retardation δ_{ex} for the extreme ray is given (see Figure 4.3) by $\delta_{ex} = \delta/\cos\alpha$, where α is the angle of divergence. For the central ray,

$$\delta = n\lambda = n/\bar{\nu}$$

while for the extreme ray,

$$\delta_{ex} = n\lambda' = n/\bar{\nu}'$$

where λ' is the effective wavelength. Solving these equations gives

$$\bar{\nu}' = \bar{\nu}\cos\alpha$$

Since $\cos\alpha < 1$, this equation implies that the effective wavelength for the extreme ray is greater than

the actual wavelength. The two consequences of this are a reduction in resolution (since a divergent monochromatic source appears as a narrow band of wavelengths) and a shift toward lower wavenumber in the recorded spectrum [104]. The wavenumber shift varies linearly with respect to $\bar{\nu}$, so it is easily correctable.

The great precision of the wavenumber scale of an FTIR instrument derives from the use of a HeNe laser to control the sampling of the interferogram. The laser beam (with wavenumber about $15798\,\mathrm{cm}^{-1}$) passes through the interferometer and a separate detector measures its interferogram, which should be a cosine wave with very little attenuation even at large retardation. The peaks in the laser interferogram are used to trigger the sampling of the infrared interferogram. The abscissa spacing after the Fourier transform is proportional to the laser wavenumber; because the laser is unlikely to be perfectly aligned, its effective wavelength will be somewhat longer and the abscissa spacing will be slightly too small.

Because of these wavenumber shifts, FTIR instruments require wavenumber calibration against a suitable sample (water vapour and polystyrene film are common examples [109]). The calibration is effected simply by changing the wavenumber of the laser entered into the Fourier transform programme.

In addition to laser misalignment, a number of other optical adjustments can cause wavenumber shifts. The system is particularly vulnerable wherever the beam is focused. Changing the diameter of an aperture or using a sample with dimensions smaller than the size of the focused beam are common causes of wavenumber shifts [104]. It must be stressed that these shifts are small, as will be demonstrated later, in Section 4.2.6. The biggest danger is that a wavenumber shift will occur between the measurement of the background and the measurement of the sample spectrum. If this occurs, the background absorbance features such as water vapour will fail to cancel completely, leaving derivative-shaped bands. This problem is especially acute for water vapour because it is pervasive in normal laboratory environments and its IR absorption features are very sharp and strong. Section 4.2.6 in this chapter discusses this phenomenon as it pertains to the present system.

4.2 Fibre-optic reflectance probe

4.2.1 General principles of fibre optics

Optical fibres transmitting in the visible and near-infrared have long been used in imaging and communications applications. The current and potential applications of infrared fibre optics include remote thermometry, infrared imaging (via large bundles of fibres), laser power delivery for medical applications, and chemical sensing. The discussion below concentrates on chemical sensing applications

requiring transmission throughout the mid-infrared and fibre lengths of a few metres.

Total internal reflection

Transmission of light through optical fibres occurs because of total internal reflection within the fibre. Snell's law determines the angle of refraction when light passes from one material to another:

$$n_1 \sin\theta_1 = n_2 \sin\theta_2$$

When the light is incident from the first medium and $n_1 > n_2$, there will be a critical incidence angle, θ_c, for which $\theta_2 = 90°$ and therefore $\sin\theta_2 = 0$. At any angle greater than θ_c, no light is refracted into the second medium. If the first medium is an optical fibre, the light will continue to propagate along its length.

Attenuated total reflection

When total internal reflection occurs, there is no propagating wave outside the fibre core: however, the field amplitude does not fall to zero at the boundary. A field with exponentially decaying amplitude, called the evanescent wave, is present outside the core. This effect can be demonstrated easily by calculations of the sort described in Chapter 2. The mean square electric field amplitudes for a system with $n_1 = 1.5$ and $n_2 = 1$ are plotted in Figure 4.4. In this example, the critical angle is $\theta_c = \sin^{-1}(1/1.5) \approx 41.8°$.

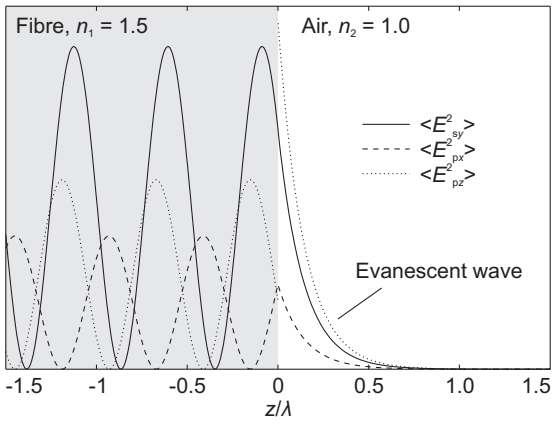

Figure 4.4: Mean square electric field amplitudes for total reflection at a boundary, showing the evanescent wave. The optical constants are $n_1 = 1.5$ and $n_2 = 1$ and the incidence angle is $\theta_1 = 60°$ (the critical angle is $\theta_c = 41.8°$). Note the discontinuity in the z component.

The evanescent wave amplitude falls off rapidly with distance from the boundary. The rate of the decay and the amplitude at the boundary both decrease as the incidence angle increases past the critical

angle. If any absorbing material is in close proximity to the fibre, energy will be lost by absorption of the evanescent wave. This phenomenon forms the basis of attenuated total reflection (ATR) spectroscopy [110, 111], but is clearly undesirable when the goal is to transmit radiation from one end of the fibre to the other. Absorption of the evanescent wave is prevented by coating the optical fibre in a transparent "cladding" layer with a lower refractive index than the core of the fibre. Radiation propagating through the fibre is totally reflected at the boundary between the core and the cladding, and the cladding is thick enough that the evanescent wave falls to zero before reaching its outer edge. The fibre is usually encased in an additional buffer layer which serves to protect the fibre from abrasion.

Losses in optical fibres

The processes resulting in transmission losses in fibre optics can be grouped into four classes [112]: intrinsic and extrinsic losses due to the fibre material itself, and Fresnel (reflection) and bending losses due to geometric factors and the optical constants of the fibre materials. Intrinsic losses include absorption due to electronic or lattice resonances as well as Rayleigh scattering. Extrinsic losses include scattering by structural defects in the fibre (such as grain boundaries in polycrystalline fibres, air bubbles or other inclusions) and absorption by impurities.

In addition to losses within the fibre, losses in the coupling of the fibre into the optical system must be considered. When light is incident on an end of the fibre (either from outside the fibre or from inside), a portion is reflected and lost. These losses can be calculated from Fresnel's equations (Equations 2.13–2.16). The losses will be large for fibres with high refractive index and for tightly focused beams where a significant proportion of the radiation is incident at large angles to the normal. Bending losses are due to the conditions for total internal reflection not being met where the fibre is curved. The extent of the bending loss depends on the ratio of the bending radius to the core diameter of the fibre, with a smaller ratio resulting in a larger loss [112].

4.2.2 Infrared optical fibre materials

Infrared optical fibres can be grouped by the type of material of which they are composed. The earliest optical fibres were fabricated from chalcogenide glasses. More recently, crystalline fibres and cylindrical hollow waveguides have been developed [113].

The loss spectrum of a given fibre will depend on the particular composition and manufacturing process, but some typical spectra for the most important classes of fibre are plotted in Figure 4.5.

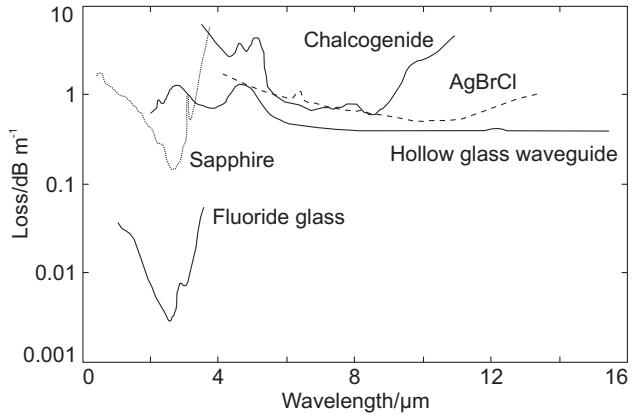

Figure 4.5: Loss spectra of infrared optical fibres (after Ref. 113).

Glass fibres

For visible and near-infrared wavelengths, silica fibres have low attenuation and excellent mechanical properties and are inexpensive. However, no oxygen-containing glass is free from absorption at wavelengths longer than about 6 μm [114]. Another important class of glass fibres are the heavy metal fluoride glasses (HMFG). Two of the most popular classes of HMFGs are fluorozirconates (such as ZBLAN: ZrF_4–BaF_2–LaF_3–AlF_3–NaF) and fluoroaluminates (such as AlF_3–ZrF_4–BaF_2–CaF_2–YF_3). These glasses have low attenuation for infrared wavelengths shorter than about 4 μm.

Glasses containing heavier elements are necessary for mid-infrared transmission, since the absorption edge occurs at longer wavelengths [114]. The most mature technology is for chalcogenide glasses [112]. This term encompasses many glass systems consisting of mixtures of one or more chalcogen (element in group 16 of the periodic table: O, S, Se, Te) and one or more of the elements As, Ge, P, Sb, Ga, Al, Si. Particularly important systems are As–S and As–Ge–Te–Se. Chalcogenide glasses are opaque to visible light and have high refractive indices in the infrared, which leads to high reflection losses, of the order of several tens of percent.

Crystalline fibres

Polycrystalline fibres of KRS-5 (TlBrI) or AgBrCl can be formed by hot extrusion. KRS-5 is problematic because of its toxicity. AgBrCl fibres transmit to wavelengths of around 18 μm, but suffer from several handling problems. The fibres are weak and, under strain, will deform plastically, creating grain boundaries and a high-loss region in the fibre. Additionally, exposure to visible or UV light will result in colloidal silver forming in the fibre, increasing the losses. Single-crystal fibres of sapphire can be grown in lengths of up to about a metre. These have very good mechanical properties, but do not

transmit at long wavelengths.

Hollow waveguides

Hollow waveguides with a circular cross-section and diameter less than 1 mm can be fabricated by starting with plastic, metal or glass tubing. The fibre consists of an air (or other gas) core surrounded by a dielectric layer, then a metallic layer, then (for plastic or glass waveguides) another dielectric layer. Glass waveguides are currently more popular because the smoothness of the tubing reduces scattering losses [113]. Hollow waveguides have major advantages for applications such as laser power delivery, but are handicapped by large bending losses proportional to the reciprocal of the bending radius [113].

4.2.3 Design of the grazing-angle probe

Most reflection accessories focus the infrared beam onto the sample surface. For cleaning validation, however, it is expected to be more useful to sample a larger area by using a collimated beam, as this should help to mitigate the effects of sample heterogeneity. The grazing-angle probe (a prototype of the SpotView instrument made by Remspec[1] [115]) is shown schematically in Figure 4.6. Not shown are the launch optics which couple the source end of the fibre bundle to the spectrometer's external beam port: these simply consist of an off-axis parabolic mirror, to focus the beam onto the end of the fibre bundle, and a mounting for the fibre bundle that allows translation in three dimensions (z, along the fibre axis; x, horizontally; y, vertically).

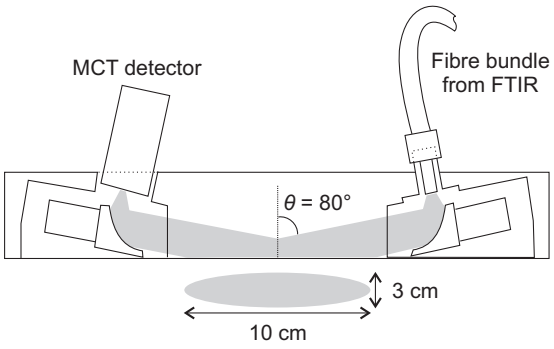

Figure 4.6: Diagram of the grazing-angle fibre-optic IRRAS probe.

Within the probe, two off-axis parabolic mirrors are used. The first collimates the beam from the fibre bundle and directs it toward the sample at an angle of 80° to the normal. The second focuses the beam reflected from the sample surface onto the detector. In the prototype configuration the probe is

[1] Remspec Corporation, Sturbridge, MA. Website: http://www.remspec.com

simply rested on the surface to be sampled. The production version is hand-held, to permit sampling of vertically oriented surfaces.

Optical fibres and packaging

The optical fibres are 500-micron diameter chalcogenide glass (As–Se–Te clad in As–Se–S with a polymer buffer) from Amorphous Materials, Inc. (Garland, TX). Nineteen fibres are gathered in the bundle and the ends are terminated with proprietary connectors, similar to the SMA type. The bundle is wrapped with flexible armouring. The length of the bundle is approximately 1 m.

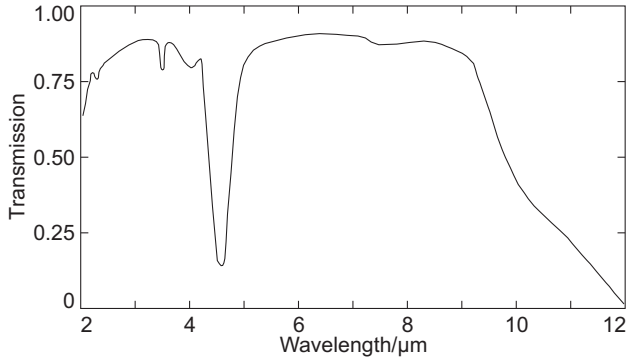

Figure 4.7: Transmission spectrum of a 1 m length of As–Se–Te glass fibre from Amorphous Materials, Inc. (Garland, Texas). The spectrum is obtained as the ratio of the spectrum of a 1.5 m fibre to that of a 0.5 m fibre, so does not include reflection losses. The spectrum is adapted from http://www.amorphousmaterials.com/images/IR-fib_C1.gif.

The transmission spectrum of the chalcogenide fibres used (from the manufacturer) is plotted in Figure 4.7. There is appreciable transmission from 2–11 μm, or about 5000–900 cm^{-1}.

Detector

An MCT-12-2.0 detector (Infrared Associates, Stuart, Florida) is mounted on the grazing-angle probe. The MCT detector is a photoconductive device [116]. When light is incident on the semiconductor detector element, electrons are excited into the conduction band, increasing the conductivity. A voltage is applied across the detector element, and the current is measured.

Since infrared light has low energy, the detector must be cooled to reduce the concentration of thermally excited charge carriers. Most laboratory systems (including the present system) use liquid nitrogen cooling, but detectors with thermoelectric cooling are available.

Infrared footprint

To determine the area illuminated by the probe, the intensity profile was measured in two dimensions by moving a $1 \times 1\,\mathrm{cm}^2$ mirror on a raster and recording the analog to digital converter (ADC) count. The smoothed image is plotted in Figure 4.8. The profile along each dimension is approximately described by a Gaussian curve, and the 2-dimensional profile has elliptical contours. All of the radiation is contained within an ellipse with axes of 100 and 30 mm; the ellipse containing 80 % of the intensity has axes of approximately 65 and 19 mm.

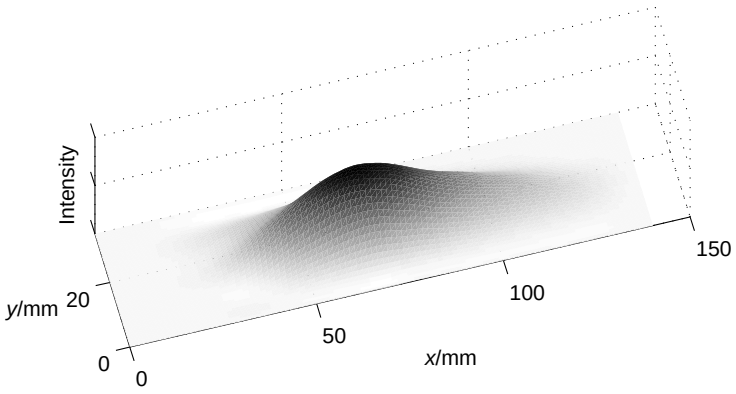

Figure 4.8: Smoothed intensity profile of the infrared footprint of the fixed-angle (80 °) IRRAS head. The edges of the mirrors are at $x \approx 0$ and $x \approx 140$ mm.

4.2.4 Single-beam spectra

A typical single-beam spectrum obtained with the FT-IRRS system, with a steel mirror in place of the sample, is plotted in Figure 4.9. The curve is a superposition of the blackbody emission of the source, the transmission of the fibre (see Figure 4.7), the transmission of the atmospheric gases in the beam path (note the intense absorption by H_2O and CO_2) and the response of the detector. The strong absorption at about $2200\,\mathrm{cm}^{-1}$ is due to an H–Se impurity in the fibre.

4.2.5 Noise characteristics

In infrared spectroscopy, shot noise is usually unimportant, since it is dominated by detector noise [104]. Since detector noise is independent of the signal level, the SNR increases linearly with increasing signal strength. For quantitative analyses, spectra are calculated in absorbance mode, that is $A = \log_{10}(B/S)$ where B and S are the background and sample single-beam spectra respectively. Assuming that the

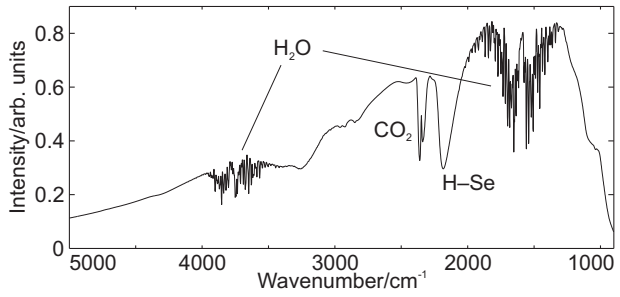

Figure 4.9: A typical single-beam spectrum obtained with the FT-IRRAS system, with a steel mirror in place of the sample (50 scans; ADC ≈ 28000). Major absorptions are indicated.

noise in B, N_B, is uncorrelated with the noise in S, N_S, the noise in A is given by

$$N_A = \frac{1}{\ln(10)} \sqrt{\left(\frac{N_B}{B}\right)^2 + \left(\frac{N_S}{S}\right)^2} \tag{4.8}$$

If A is small and B and S are measured in the same way (same number of scans, etc.), as is usually the case in IRRAS, $B \cong S$ and $N_B \cong N_S$, so $N_A \propto N_B/B$: the noise in the absorbance spectrum is inversely proportional to the signal level in the single-beam spectra.

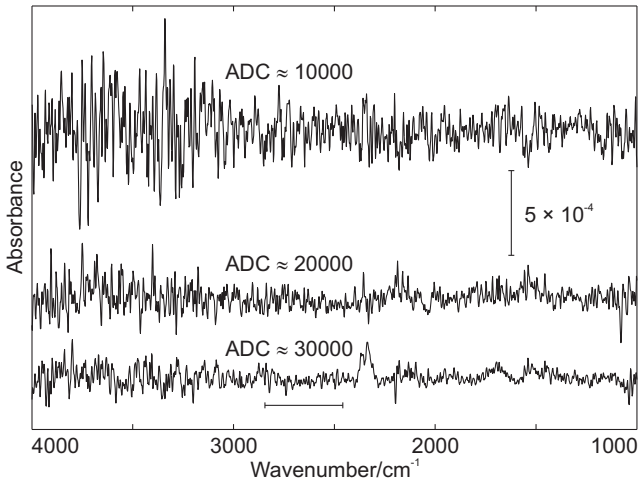

Figure 4.10: Blank spectra at various signal levels, measured as the analog-to-digital converter count (ADC). The sample single-beam spectrum is measured immediately after the background single-beam spectrum; each is calculated from the average of ten interferograms. The spectra are offset for clarity and the ordinate scale is indicated. The horizontal bar indicates the wavenumber range used for the noise calculations presented in Figure 4.11.

Figure 4.10 shows some "100 % line" (calculated from two consecutively measured blank single-beam spectra) measurements in absorbance mode for several values of the signal level. The signal level was adjusted by moving the source fibre bundle end away from the focus of the launch optics (in a direction parallel to the fibre axis). It is evident that the noise decreases with increasing signal level.

Also, since the signal is wavenumber-dependent, so is the noise. The ADC count is the digital signal from the analogue-to-digital convertor at the ZPD of the interferogram. It is taken here as a measure of the intensity of the single-beam spectrum, to which it is also proportional.

In Figure 4.11a, RMS noise values calculated over the wavenumber range 2840–2460 cm^{-1} (indicated by the horizontal bar in Figure 4.10) are plotted against reciprocal signal level. The noise is calculated by fitting a straight line through the indicated region of the spectrum and calculating the RMS of the deviations (since the RMS noise is similar to a standard deviation, the peak-to-peak noise will be ~5 times greater). The noise is inversely proportional to the signal level, as expected. It is demonstrated in Figure 4.11b that the noise is also inversely proportional to the square root of the number of scans.

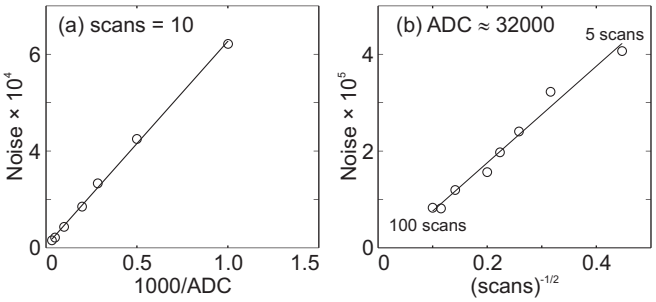

Figure 4.11: RMS noise (absorbance mode) over the range 2840–2460 cm^{-1} as a function of (a) signal level, and (b) number of interferometer scans.

4.2.6 Optical adjustments and wavelength shifts

While it might seem that accessories external to the spectrometer should be immune to the wavelength-shift problem discussed in Section 4.1.4, since the beam is already modulated when it leaves the interferometer, this is not the case. What matters is the average angle of radiation through the interferometer [104], but this only applies to radiation that actually reaches the detector. Thus, any accessory that does not preserve the angular distribution of the radiation leaving the spectrometer will affect the wavelength shift.

In the present system, there are foci at both fibre ends and at the detector. Each of these optical elements can be translated in three dimensions and rotated. Additionally, the angle and vertical separation between the probe and the sample can be adjusted. All these variables have the potential to affect the wavelength shift.

Figure 4.12 illustrates how optical adjustments between the background and sample single-beam measurements can introduce artefacts due to incomplete cancellation of water vapour. After measuring

the background spectrum (B), the probe end of the fibre was rotated 90°; then the sample spectrum (S) was measured. To prevent genuine water vapour absorption from obscuring the artefacts, the spectrum in Figure 4.12b was calculated as a modified difference spectrum rather than as absorbance:

$$D = S - bB \tag{4.9}$$

where b was found by minimising the integral of the absolute value of D (see Appendix B for a discussion of some related techniques). The result is that water vapour absorption features in D are suppressed, revealing the subtraction artefacts, which cannot be reduced by changing b. The magnitude of these artefacts is significantly greater than the noise level.

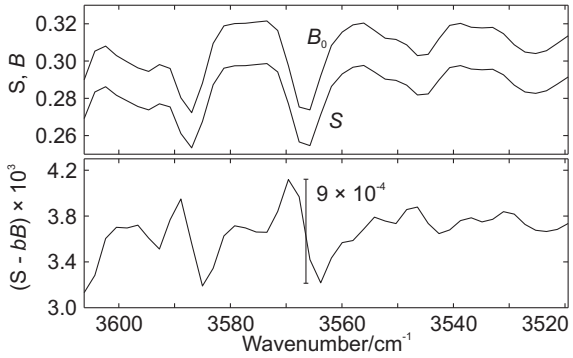

Figure 4.12: (a) Two single-beam blank spectra, and (b) their difference, as per Equation 4.9; $b = 0.92$. The end of the fibre was rotated ~90° between the two measurements.

The size of the wavenumber shift incurred by a number of optical adjustments was determined using a version of the OPUS software's wavenumber scale calibration routine modified to use the external beam port and detector (see Appendix C). The water vapour absorption band at

$$\bar{\nu}^0_{H_2O} = 1554.353\,\mathrm{cm}^{-1}$$

was used as a reference. The laser wavenumber was set to its nominal value of

$$\bar{\nu}^0_{HeNe} = 15798\,\mathrm{cm}^{-1}$$

A single-beam spectrum at $1\,\mathrm{cm}^{-1}$ resolution with no apodisation was measured and the precise peak location $\bar{\nu}_{H_2O}$ found using OPUS's peak-finding routine. This wavenumber was used to calculate the relative error ρ in the wavenumber scale:

$$\rho = \bar{\nu}^0_{H_2O}/\bar{\nu}_{H_2O} \tag{4.10}$$

The laser wavenumber required to correct the abscissa is given by

$$\bar{\nu}_{\mathrm{HeNe}} = \rho \bar{\nu}^{0}_{\mathrm{HeNe}} \tag{4.11}$$

As shown in Table 4.1, with all optical elements optimally positioned (the signal level is regulated by the z adjustment of the source end of the fibre; all other elements are positioned to give highest throughput), $\rho - 1 \approx 10^{-4}$. A similar value is obtained using the internal beam path and detector, so this error is probably mostly due to laser misalignment. The absolute error in the wavenumber scale is given by $(\rho - 1)\bar{\nu}$ which corresponds to $0.1\,\mathrm{cm}^{-1}$ at $1000\,\mathrm{cm}^{-1}$ and $0.4\,\mathrm{cm}^{-1}$ at $4000\,\mathrm{cm}^{-1}$. Additionally, some spread around this value of ρ is introduced by varying other optical parameters. These changes are on the order of 10^{-5}, contributing an additional error of 0.01–$0.04\,\mathrm{cm}^{-1}$.

Table 4.1: Wavenumber shifts introduced by various optical adjustments. The parameter ρ is the ratio between the measured wavenumber of a water vapour band and its true value. The magnitudes of the adjustments are approximate only.

Adjustment	**Magnitude**	**ADC count**	$(\rho - 1) \times 10^4$
None		29000	0.99
Sample height	5 mm	14070	1.30
Sample angle	+1.5°	20050	1.01
Sample angle	−1.5°	22590	1.18
Detector rotation	90°	26150	0.99
Probe fibre end rotation	90°	27550	1.19
Probe fibre end translation (z)	2 mm	27200	1.07
Source fibre end translation (x)	1 mm	24770	1.34
Source fibre end translation (y)	1 mm	24900	1.21
Source fibre end translation (z)	1 mm	26540	0.95

There are two ways for the wavelength shift to contribute error to chemometric models. The most obvious is the introduction of water vapour subtraction or ratio artefacts, as discussed above. Since the wavelength shift between background and sample will be slightly different each time, the shapes of the derivative spectra will differ, so these artefacts cannot be fully accounted for by increasing the rank of the chemometric model. Furthermore, they cannot be removed readily by subtraction methods. For this reason, these artefacts must be avoided. This requires that the background and the sample spectra must be obtained using optical configurations as close to identical as possible. This simply implies that in a production system, care be taken that all optical components can be firmly fixed in place, and that the probe is always held flat against the sample surface.

The second possible detrimental effect, which may occur on a sample-to-sample basis, is the subtle change in shape of the spectrum caused by stretching or compressing the abscissa slightly. This scaling

would occur if the optical configuration were changed between samples (each spectrum would have a slightly different ρ), but is not expected to be significant: bandwidths for solid-phase spectra are typically on the order of a few wavenumbers or tens of wavenumbers, at least 100 times greater than the expected wavenumber scale error.

Early in this work, the laser in the spectrometer failed and had to be replaced. The laser wavenumber in the software was not changed after this, and the oversight was not realised until a significant time later. To ensure that future spectra were compatible with those already measured, the laser wavenumber was not corrected. This means that all spectra are subject to a constant error corresponding to $\rho - 1 \approx 10^{-4}$.

4.3 Preparation of standards and IRRAS measurement

For a chemometric model to be relevant, the standards from which it is derived must be similar to real samples: that is, surfaces of pharmaceutical manufacturing equipment, possibly contaminated by active pharmaceutical ingredients or precursors, excipients, and cleaning compounds. The standards must be prepared on coupons (appropriately sized plates or sheets) made from the same material as the target surface. In particular, the composition and surface finish must be the same. Any compounds that may be present after cleaning must be included in the experimental design for the calibration, even if they are not being calibrated for. This may include cleaning agents and excipients as well as active pharmaceutical ingredients and precursors or degradation products.

The standards must also be well characterised. The loading (defined as mass per unit area, usually $\mu g\,cm^{-2}$) of every analyte must be known precisely. The standards should also be as homogeneous as possible: the reasons for this requirement and the consequences of heterogeneity are discussed below in Section 4.3.5.

4.3.1 Smear technique

The most obvious way to prepare a standard is to smear an aliquot of a solution of the analyte over a clean coupon of the substrate material, then allow the solution to dry. This method has the advantage that the mean loading can be quite precisely controlled. For example, a loading of $1\,\mu g\,cm^{-2}$ on a coupon with area $256\,cm^2$ could be obtained by spreading 1 mL of a $256\,mg\,L^{-1}$ solution of the analyte over the surface. The spreading tool must not scratch the surface, and care must be taken to ensure that no significant amount of material is lost by transfer to the spreading tool. A spatula cut from soft plastic works well for metal or glass substrates. The main drawback of this method is that it is

very difficult to prepare homogeneous standards. The problem is reduced if small coupons are used (with dimensions similar to the illuminated area). For more discussion of this method, see Michelle Hamilton's thesis [106].

4.3.2 Spray technique

The analyte can be dispersed more evenly over the substrate coupon by using an airbrush. The analyte is dissolved (typically at a concentration of around 1 mg mL^{-1}) in an appropriate solvent and the solution is sprayed over the substrate. As with the smear technique, there are a number of variables to consider. The volatility of the solvent is particularly important: for the aerosol to adhere to the substrate, the droplets must not have evaporated by the time they reach the substrate. This means that a more volatile solvent necessitates holding the airbrush closer. The smaller the spraying distance, the smaller the area in the path of the aerosol and the greater the effort required to ensure a homogeneous standard. A less-volatile solvent allows a greater spraying distance, but the coupon takes longer to dry.

The substrate coupon is placed in a custom-built cabinet which is enclosed on all sides except the front and has an extraction system. The coupon is rested vertically on a stand and oriented towards the airbrush-wielding experimenter. The airbrush used in this work was a double-action, internal-mix model from Paasche, operated with a small diaphragm compressor. The use of a ballast tank was investigated, but found to be unnecessary; the slight pulsing of the aerosol due to the pump does not seem to adversely affect the homogeneity of the standards produced.

Acetone and water were used as solvents for aerosol deposition. For acetone, the airbrush must be held no more than ~40 cm from the surface. Consequently, care must be taken to ensure even deposition of the analyte. Best results were obtained when the brush was moved in a raster fashion, with lateral strokes back and forth and each stroke offset vertically from the last. An area substantially larger than the coupon should be "painted" in this way. Because acetone evaporates quickly, it is possible to use a dilute solution of which several coats may be applied, to obtain an averaging effect.

When water is used as the spraying solvent, the airbrush can be held significantly further away, around 1–1.5 m. At this distance, the aerosol covers a significantly greater area than the coupon. Some motion of the brush is still required, however, since the density of the aerosol is greatest near its centre. Two patterns that work well are a slow, circular motion where the aerosol is directed near the edges of the coupon; or the raster-style movement as described for acetone.

Typically, at least 3–5 mL of solvent are required to produce an even coat. It is convenient to prepare ~20 mL of solution directly in the reservoir by weighing in a few mg of the analyte then dissolving it. Acetaminophen and aspirin, which are much more soluble in ethanol than in water, were

usually dissolved in $\lesssim$1 mL of ethanol, to which water was added to fill the reservoir.

When mixtures of more than one analyte are being prepared, it is important to vary the concentrations of the analytes independently. In this case, the best approach is to prepare stock solutions of the individual analytes and to prepare the spraying solution for each standard by mixing the stock solutions in a unique ratio.

The disadvantage of the spray method, compared to the smear method, is that the loading can only be loosely controlled by varying the concentration, spraying distance and spraying time. For the standard to be useful, the precise loading must be determined by another method.

4.3.3 Primary calibration

The most direct method to determine the loading would be to weigh the substrate coupon before and after spraying, with the loading being the ratio of the difference in the weights to the area of the coupon. In general, this is difficult because of the large dimensions of the coupon and the small difference in the weight. A 1 $\mu g\,cm^{-2}$ coating over 256 cm^2 weighs 256 μg. The best balance available in the Department can be read to ± 5 μg, but only for samples weighing <30 g; the precision is ten times worse for heavier samples. The area of a sheet of 2 mm-thick window glass (specific gravity $\sim$2.5 $g\,cm^{-3}$) weighing 30 g is $\sim$150 cm^{-2}. The smallest loading that can be measured with precision of 10 % is therefore $\sim$0.7 $\mu g\,cm^{-2}$. In practice, the precision is somewhat worse. Finally, weighing can only give the total loading, which is a disadvantage when there are multiple analytes. (If the standard is prepared in such a way that the analytes are present in known ratios, the individual loadings can be determined, however.)

A better approach is to rinse the analyte from the substrate and analyse the resulting solution. This depends on having a solvent in which the analyte dissolves readily and an analytical method to determine the concentration. The APIs used in this work are soluble in ethanol and have strong UV chromophores, so UV colorimetry is an ideal method. A calibration can be established by preparing a number of standard solutions of the analyte and measuring their spectra. CLS regression (see Section 3.2.1) provides a straightforward analysis while retaining the principal advantages of multivariate calibration (in particular, outlier detection via inspection of spectroscopic residuals). This procedure is illustrated below, using the primary calibration for the API mixtures work (see Chapter 7) as an example. More sophisticated extraction processes may be required in cases where the analyte is less soluble.

The two analytes are aspirin and acetaminophen. Their UV spectra (in ethanol solution) are plotted in Figure 4.13, along with the background absorbance for a 1 cm pathlength of ethanol. Spectra

were measured on a GBC 920 UV/vis spectrometer in single-beam mode with a resolution of 2 nm (corresponding to the largest slit-width setting) and with a scan speed of 180 nm min^{-1}.

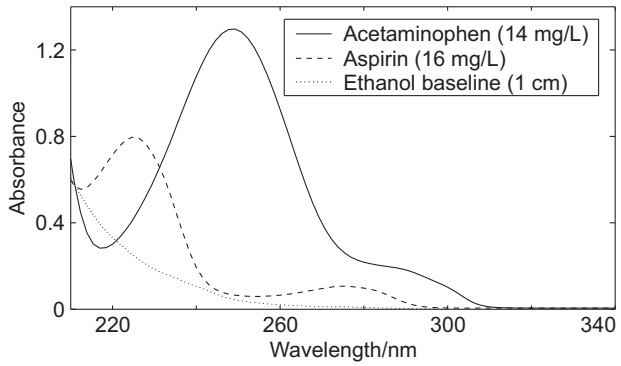

Figure 4.13: UV spectra of acetaminophen, aspirin and ethanol. The API spectra have been corrected by subtracting the absorbance due to the solvent.

The peaks are well enough separated that analysis of mixtures should be possible (an alternative would be to use HPLC). The peak absorbance for acetaminophen,

$$\epsilon_{249\,\mathrm{nm}} = 13700\,\mathrm{L\,mol^{-1}\,cm^{-1}}$$

occurs in a region where absorption by the solvent is minimal, while that for aspirin,

$$\epsilon_{225\,\mathrm{nm}} = 8900\,\mathrm{L\,mol^{-1}\,cm^{-1}}$$

overlaps with the onset of significant absorption by the solvent. Since ethanol has a fairly high coefficient of thermal expansion ($\rho = 0.7894\,\mathrm{g\,cm^{-3}}$ at 20 °C and $\rho = 0.7810\,\mathrm{g\,cm^{-3}}$ at 30 °C [117]), small changes in temperature between baseline and sample measurement could lead to measurable changes in the background absorption and potentially to errors in the measured concentration. If necessary, this situation can be rectified by including the solvent as an absorbing species in the CLS model, or by use of an inverse regression method, as described in Chapter 3.

Calibration was effected by measuring spectra of six solutions containing both compounds and using CLS regression to estimate the pure component spectra. Validation was by an external test set consisting of nine standards that were prepared independently of the calibration standards and whose true concentrations are shown as circles in Figure 4.14. The results of the validation are also summarised in Figure 4.14, by plotting the predicted concentrations as crosses.

With the exception of the point in the centre-bottom of the plot, agreement is excellent. Acetaminophen concentrations are predicted slightly high on average. The isolated significant error is

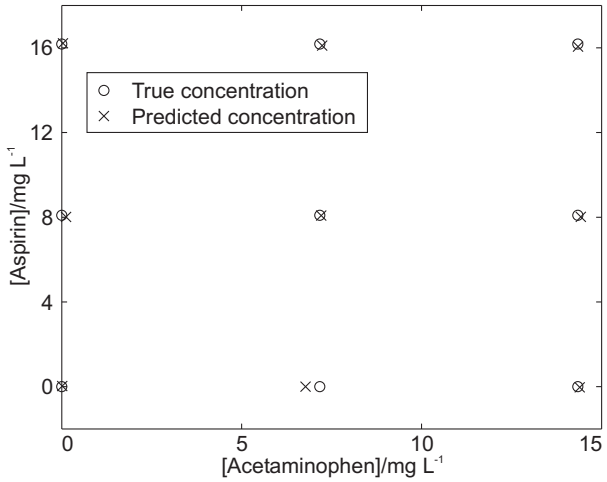

Figure 4.14: Test-set validation for acetaminophen and aspirin mixed solutions. Circles are "true" (gravimetrically determined) concencentrations; crosses are model predictions. Horizontal and vertical deviations represent errors in acetaminophen and aspirin concentrations, respectively.

most likely due to an error during standard preparation. The results are presented numerically in Table 4.2. The RMSEP is about 0.06 mg L^{-1} for both compounds. Assuming a rinse volume of 50 mL and a coupon area of 256 cm^2, this corresponds to a standard deviation of approximately 0.01 µg cm^{-2} in the determination of the loading. As will be seen in later chapters, this error is much smaller than that encountered in the validation of the IRRAS models.

Table 4.2: Validation results for UV + CLS calibration for mixtures of acetaminophen and aspirin in ethanol solution. C_{res} are the concentration residuals and A_{res} are the RMS absorbance residuals for each spectrum. The RMSEP value enclosed in parentheses is for acetaminophen with the outlier removed.

[Acet.]/mg L^{-1}			**[Aspirin]**/mg L^{-1}			
C_{true}	C_{pred}	C_{res}	C_{true}	C_{pred}	C_{res}	$A_{res} \times 10^3$
0	0.02	0.02	0	0.03	0.03	1.5
0	0.12	0.12	8.09	8.01	-0.08	2.8
0	0.04	0.04	16.18	16.21	0.03	3.7
7.17	6.78	-0.39	0	-0.001	-0.001	3.5
7.17	7.21	0.04	8.09	8.08	-0.01	3.1
7.17	7.24	0.07	16.18	16.11	-0.07	5.3
14.34	14.39	0.05		-0.03	-0.03	4.9
14.34	14.43	0.09	8.09	8.01	-0.08	3.8
14.34	14.35	0.01	16.18	16.06	-0.12	3.3
RMSEP: 0.14 (0.06)			**RMSEP**: 0.06			

For the rinse method to be accurate, it is essential that all of the analyte is rinsed from the surface. This could be problematic for some compounds, but for the APIs used in this work, 25 mL of ethanol was found to be adequate. Two methods were used to verify complete recovery. First, aliquots of

the APIs were deposited on glass and steel coupons and the solvent was then allowed to dry. Rinsing samples with loadings $\lesssim 3\,\mu g\,cm^{-2}$ with 50 mL of ethanol, apparent recovery percentages of between 98 % and 102 % were obtained.

A second check, which was employed throughout the work, consisted of collecting an additional small volume of rinsate after rinsing was complete. The concentration of API in this second volume should be near zero. This was almost always the case. Occasionally, a statistically significant concentration was found, but never enough to imply an error of more than about 1 % in the measured loading. As will be seen later, the error in the loading determination by the rinse-UV method is very small compared to the errors introduced by sample heterogeneity and the final errors associated with the IRRAS method as a whole.

4.3.4 IRRAS measurement

Once a standard has been prepared, IRRA spectra can be measured. In this work, large (in terms of the footprint size) coupons (about $16 \times 16\,cm^2$) have been used. The original reasons for this were twofold. First, the total amount of analyte is greater so the loading determination should be more precise. This is not a significant concern for compounds with strong UV absorptivity, but is important in a parallel study with the surfactant sodium dodecyl sulfate as the analyte, in which ^{1}H NMR is used as the reference method [106]. Second, it was thought that if several IRRAS measurements were made from each standard it might be possible to produce calibrations using fewer standards. This idea relied on the assumption that features unrelated to the analyte would vary randomly between measurements, which turned out to be unfounded (see Sections 3.5 and 6.3). However, the use of large coupons exacerbates the effects of sample heterogeneity, as discussed below.

Measurements are made by resting the probe on the sample surface and recording a single-beam spectrum. The IRRAS is then calculated as absorbance against a suitable background spectrum, usually measured beforehand from a clean coupon of the same substrate material.

4.3.5 Sample heterogeneity

A standard is heterogeneous if the loading in any given region of the coupon differs from the mean loading. This is a scale-dependent phenomenon: on the scale of the droplets deposited by the airbrush, the standard will necessarily be heterogeneous. On a larger size scale, the standard may be heterogeneous if the spots are not uniformly distributed, or are uniformly distributed but few enough in number that statistical variations in spot density from region to region are significant. Each IRRAS measurement

probes an area of approximately

$$\pi/4 \times 10\,\mathrm{cm} \times 3\,\mathrm{cm} \approx 24\,\mathrm{cm}^2$$

or about a tenth of the surface of the coupons used in this work. Since the centre is weighted much more strongly than the edges, this number is an overestimate: from Figure 4.8, 80 % of the total intensity falls in an area of

$$\pi/4 \times 6.5\,\mathrm{cm} \times 1.9\,\mathrm{cm} \approx 10\,\mathrm{cm}^2$$

The extent of the heterogeneity can be determined by measuring the loadings of different regions of the coupon using the UV method. Since the relative standard deviation (RSD) of the local loadings may be quite small, it is important to take the samples in such a way that the sampling method does not contribute appreciably to the variation in the analytical results. The most straightforward approach, and the one taken here, is to cut a coupon into several strips, arrange them into the shape of the original coupon, and spray them in the usual way. The strips can then be analysed individually by the rinse method with no sampling error. Another approach would be to inscribe lines on a coupon, dividing it into the desired number of regions. After spraying, each of these regions could be swabbed repeatedly with a cotton bud or similar implement moistened with solvent. This would require care in confining the swabbing to the desired region and would introduce the complication of absorbing compounds leaching from the swabbing tool. These problems would not be irresolvable, and the flexibility in selecting the sampling region (in particular, the ability to sample smaller regions) would be advantageous.

A polished stainless steel coupon ($15 \times 15\,\mathrm{cm}^2$) was cut into four strips of approximately equal area. The strips were weighed to establish their relative areas (under the assumption of constant thickness); the total spread in masses was about 4 %. They were sprayed with a solution of ~50 mg of acetaminophen in ~20 mL of water (with ~1 mL of ethanol used to dissolve the acetaminophen first). From experience, such a spraying solution is expected to give a mean loading on the order of 2 $\mu\mathrm{g\,cm}^{-2}$. Each strip was rinsed and the UV spectra measured.

Since it is the relative concentrations that are of interest, there is no need to determine the absolute loadings. The relative loadings were taken as the scores of a 1-factor PCA model (see Section 3.2.3) normalised to have unit mean and corrected for the different areas of the strips. The residuals were monitored to ensure that the 1-factor model was adequate. (Additionally, almost identical results were obtained by using the absorbance at 249 nm as a univariate measure of the relative loading.) The relative standard deviations (in percent) in 6 repeats of the experiment were 3.6, 1.5, 4.5, 9.6, 5.1, 6.2. It is obvious that there is considerable variability. The mean of these numbers is 5.1 %, or 4.2 % if the

outlying value is removed; this latter is taken as an estimate of the RSD in the loading.

It is not completely clear how best to extrapolate this value to estimate the RSD when an area smaller than a quarter of the coupon is sampled. The simplest statistical model for the spraying process is that of a multinomial experiment [118], in which each of N droplets has an equal probability of landing in each of n regions of equal area. The RSD is a function of both N and n; if N can be found such that $\mathrm{RSD}(N, 4) = 0.042$, then $\mathrm{RSD}(N, 10)$ can be calculated.

The RSD for the number of droplets landing in each of the equally sized areas (see Appendix A.1.2) is given by

$$\mathrm{RSD} = \frac{\sigma}{\mu} = \sqrt{\frac{n-1}{N}} \tag{4.12}$$

Solving Equation 4.12 for N,

$$N = \frac{(n-1)}{\mathrm{RSD}^2} \tag{4.13}$$

For $n = 4$, an RSD of 4.2 % corresponds to $N \approx 1700$; extrapolating to $n = 10$ gives RSD = 7.3 %.

The significance of this value is that there is appreciable variation in the true loading sampled in any given measurement from the mean loading of the coupon. The error in the loading therefore consists of a constant term, due to the error in the mean loading, and a term proportional to the loading, due to the heterogeneity. The consequences of this for chemometric modelling are discussed in Chapters 3 and 7.

Chapter 5

A variable-angle fibre-optic reflectance probe

5.1 Introduction

From Chapter 2, it is apparent that there are several parameters to be considered when designing an IRRAS experiment. The system to be studied is, at least for organic analytes without extremely strong absorption, characterised principally by the optical constants of the substrate and the thickness of the film. But, for a given system, the incidence angle and the state of polarisation are also important in determining the properties of the IRRAS.

The first section of this chapter defines quantitative measures of the properties that are desirable in an IRRAS spectrum and, using the thin-film model described in Chapter 2, isolates experimental conditions that are conducive to obtaining those properties for several substrate materials. The subsequent sections are related to measurements obtained with a variable-angle fibre-optic specular reflectance probe based on the grazing-angle probe described in Section 4.2. The second section describes the design of this probe, while the third details the experimental and data-processing procedures. The fourth and fifth sections discuss experimental results for two types of film material on glass substrates.

The theoretical work discussed in the first section was carried out after most of the experiments discussed elsewhere in this thesis had been completed, and, as such, serves to explain some of the shortcomings encountered and to suggest improvements; see Chapter 9.

5.2 Theoretical considerations

Several studies have been published that discuss optimum conditions for IRRAS of very thin organic films ($d < 100\,\text{nm}$) on a variety of substrates [30, 46], defining "optimum" in terms of the maximum reflection-absorbance. Blaudez et al. [119] studied very thin organic films on a glass substrate and considered the signal-to-noise ratio, as did Kattner and Hoffman [120]. Thick ($d > 1\,\mu\text{m}$) films on metallic substrates were discussed by Merklin and Griffiths [121], who defined the optimal conditions in terms of lack of (qualitative) band-shape distortion.

The present investigation aims to include both very thin (defined here as $d < 100\,\text{nm}$) and thin ($d < 1\,\mu\text{m}$) films and to consider several criteria for the quality of the spectra. Ideally, instrument parameters allowing linear calibration for films up to 1 µm will be found; or, failing that, the approximate limit of the linear response will be established.

5.2.1 Criteria for selecting instrument parameters

Signal-to-noise ratio

The most basic spectroscopic performance metric is the SNR of the measurement. As shown in Section 4.2.5, for the current spectrometric system, the noise level in single-beam measurements is independent of the optical throughput (a condition that is common in FTIR spectrometry [104]). From Equation 4.8, the relative noise in reflection-absorbance spectra is given by

$$N_{\text{RA}} \propto \sqrt{\left(\frac{1}{R_0}\right)^2 + \left(\frac{1}{R}\right)^2} \tag{5.1}$$

where the noise levels in R_0 (the background reflectance) and R (the sample reflectance) are taken to be equal. The noise for p- and s-polarised spectra is calculated using R_p or R_s, respectively, in Equation 5.1; for unpolarised spectra, $R = R_p + R_s$ is used. The halving of the maximum optical throughput that occurs when an ideal polariser is used is thus implicit in the calculations.

Band-shape distortion

Distorted bands (in the sense of differing in shape from the transmission-mode absorbance) are common in IRRAS [1]. It is desirable to obtain spectra that resemble transmission-mode absorption spectra: strongly distorted band-shapes can render spectra more difficult to interpret (in terms of peak positions, heights and areas) and compare to library spectra. (However, distorted bands pose no problem for chemometric methods provided that the shape is independent of film thickness.) Since plotting spectra

for a range of incidence angles, a range of film thicknesses, for several substrates and for three types of polarisation would result in an inordinate number of figures, the similarity between the absorption and reflection-absorption band shapes will be summarised by the correlation coefficient C between them.[1] This quantity ranges from -1 (perfect negative correlation) to 1 (perfect positive correlation). The variables are mean-centred in the calculation of C, which corresponds to fitting both a slope and an intercept in linear regression. Values of C near -1 or 1 indicate a low level of band distortion. The absorbance spectrum was calculated (see Section 2.1.4) as

$$A_{\text{trans}} \propto k\bar{\nu} \tag{5.2}$$

where k is the imaginary part of the complex refractive index. See Figure 5.6 later in this chapter for some calculations of distorted spectra and the corresponding correlation coefficients.

Linearity

The linearity of the reflection-absorbance with respect to film thickness is also important. Linearity means that the RA at all wavelengths is proportional to film thickness. The chemometric methods used in this work are all based on a linear model: while the factor-based methods have some tolerance for nonlinearity in the form of changes in spectrum shape, the best results should be obtained with a system that behaves linearly. Linearity is assessed by calculating p, the projection of the RA spectrum onto the transmission-mode absorbance spectrum of a 1 nm film, $A_{1\text{nm}}$, normalised by the vector length of the absorbance spectrum:

$$p = \frac{A_{\text{RA}} \cdot A_{1\text{nm}}}{\|A_{1\text{nm}}\|^2} \tag{5.3}$$

If p is found to be greater than d, the thickness of the RA film in nanometres, this means that the RA is more sensitive than the transmission; p is the equivalent transmission-mode thickness of the RA film. If the RA bandshape is distorted, p will be reduced by a factor equal to the reciprocal of the correlation coefficient between the RA and absorbance spectra, in which case it is hard to define a good measure of intensity.

5.2.2 Baseline corrections

A complication arises from the fact that the presence of even a non-absorbing film can cause a significant wavelength-dependent change in the reflectance. Consequently, reflection-absorption spectra

[1] Elsewhere in this thesis, the symbol R is used for the correlation coefficient. In this chapter, to prevent confusion, C is used, and R is always the reflectance.

feature baseline variations of various types as well as absorption-like bands. The term "baseline" here refers to effects due to the (almost) constant real part of the refractive index: effects due to the dispersions in n that accompany the absorption bands must be considered part of the RA spectrum. Examples of these baselines can be seen in Figures 2.10, 2.12 and 2.13 later in this chapter. The latter two figures illustrate that, near absorption bands of dielectric substrates, the baseline spectrum can be very much stronger than the RA bands due to absorption by the film. Consequently, to make comparisons between different film thicknesses or substrate media, the baseline should be removed first. The baseline, defined as above, can be calculated simply by repeating the reflection-absorbance calculation but using only the average of the real part of the film refractive index.

This correction may not be practical to apply to experimental data, since it requires knowledge of the complex refractive index of the substrate and of the real part of the refractive index of the film. In most cases, however, the baseline (over a small wavelength range) will be simple in form (constant or linear) and easy to remove by standard means; the correction described above is then just a convenient, automated way to do something that could be achieved by other methods. For thin films on metals, the baselines tend to be constant or linear. Some examples of curving baselines can be seen in Figure 5.1, where some calculated RA spectra of benzene films on glass are plotted along with their baselines calculated as described above. Similar baselines are observed experimentally; see Figure 7.1.

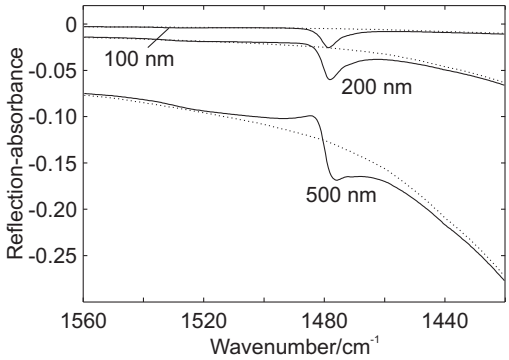

Figure 5.1: Calculated unpolarised RA spectra ($\theta_0 = 80°$; solid lines) and baselines (dotted lines) for films of benzene on glass.

5.2.3 Calculation procedure

For each of several representative metallic and dielectric substrates, the three criteria in Section 5.2.1 will be used to investigate the relative merits of different sets of experimental conditions.

As in Chapter 2, benzene is used as the model film material. The results should be generally applicable, however, since many organic compounds have $n \approx 1.5$ and $k \lesssim 0.2$ in the infrared. The wavenum-

ber range used here is 1560–1420 cm^{-1}, chosen to span a single strong band (ν_{13} CC + HCC in-plane at 1479 cm^{-1}; $n \approx 1.47$, $k_{\mathrm{peak}} = 0.153$) and a nearby weak band ($\nu_4 + \nu_{11}$ combination at 1528 cm^{-1}; $k_{\mathrm{peak}} = 0.007$). Refractive index data for the substrates were taken from the literature [40, 45, 122, 123] and interpolated linearly to match the abscissa of the benzene spectra. To investigate different regions of the substrate optical-constant spectra, the substrate spectra were translated along the abscissa so the desired region coincided with the band. This was done so that the absorbance of the band (which is proportional to wavenumber; see Equation 5.2) remained constant.

Six calculated graphs are presented for each substrate (or particular wavenumber range for a substrate):

1. The correlation coefficient C between the RA and the absorption spectrum for very thin films (10 and 100 nm) for each polarisation and as a function of incidence angle. This plot reveals whether the RA bands are positive or negative and the extent of their distortion.

2. As above, but for thicker films (200 and 500 nm).

3. The baseline-corrected RA for a 10 nm film, evaluated at the peak in the benzene k spectrum, for each polarisation and as a function of incidence angle.

4. The noise measure in Equation 5.1, also evaluated at the peak in the benzene k for each polarisation and as a function of incidence angle.

5. The SNR (baseline-corrected RA divided by noise and arbitrarily multiplied by 1000) for 10 and 200 nm films. This plot allows identification of instrument parameters providing particularly good or poor sensitivity. The SNR for the 200 nm film is divided by 200/10 for comparison with the 10 nm results (to correct for optical thickness).

6. A few values of θ_0 are chosen on the basis of the C and SNR plots. For these angles, the projection of the RA spectrum onto the absorbance spectrum of a 1 nm film is calculated as a function of film thickness. The thickness d_{99} at which the squared correlation coefficient, C^2, between the projection and the film thickness falls to 0.99 is determined: this can be taken as an estimate of the maximum allowable film thickness before nonlinearity plays a significant role.

The substrates considered are aluminium [45], a highly reflective metal; iron [123], a less reflective metal; silicon [122], a semiconductor with a high refractive index; and SiO_2 glass [40], a dielectric material featuring a strong absorption band. Selected results are discussed below and summarised in Table 5.1.

5.2.4 Metallic substrates

Aluminium

Figures 5.2a and b show that the RA band is positive ($C > 0$) and relatively undistorted ($C \approx 1$) for all incidence angles and even for a relatively thick 500 nm film.

The signal and noise plots (Figures 5.2c and d) are typical of metals. The *s*-polarised reflection-absorbance is several orders of magnitude less intense than the *p*-polarised RA (and is indistinguishable from zero in Figure 5.2c). The unpolarised RA is therefore roughly half as intense as the *p*-polarised RA, but is also only half as noisy. Consequently, the unpolarised SNR is almost identical to the *p*-polarised SNR, while the *s*-polarised SNR is very much smaller (Figure 5.2e). It is interesting to note the differences between the 10 nm and 200 nm SNR curves. Firstly, at incidence angles $\gtrsim 60°$ the 200 nm SNR is much less (after normalisation) than that for 10 nm, which indicates severe nonlinearity with respect to film thickness. Secondly, the peak in the SNR is broader and its maximum occurs at lower angle for the thicker film. For a thin film, the best SNR is obtained at ~88°; however, this angle is difficult to obtain experimentally,[2] so 85° is chosen for further investigation. Considering also the 200 nm film results, a smaller incidence angle should give better linearity with respect to film thickness, so 75° is also chosen. Following the recommendation of Merklin and Griffiths [121], the Brewster angle of the film,[3] 56°, is also considered.

The projection of the reflection-absorbance onto the transmission-mode absorbance of a 1 nm film is plotted, as a function of film thickness, in Figure 5.2f. Here, it can be seen that there are dramatic differences between the three angles. In all cases the *p*-polarised spectrum is much more intense than the *s*-polarised, except for thick films (for which complex interference effects occur; see Figure 2.11b). The deviation from linearity is much gentler when 75° or 56° is used, with d_{99} (indicated by the squares and asterisks for unpolarised and *p*-polarised light, respectively) being several times greater (see Table 5.1). The shape of the curve for 85° is problematic if the possibility of films thicker than ~200 nm cannot be excluded. The nonlinearity, coupled with the lack of significant band-shape distortion even for quite thick films, means that a film 75 nm thick is difficult to distinguish from one 700 nm thick (see Figure 5.3).

[2] For a system with a collimated beam, the length of the illuminated area on the surface is $w/\cos\theta_0$, where w is the width of the beam. For a 2.5 cm-wide beam, $\theta_0 = 88°$ corresponds to a spot 72 cm long.

[3] For transparent materials, the Brewster angle is the incidence angle at which the reflectivity of *p*-polarised light is zero [47]. It is given by

$$\theta_B = \tan^{-1}(n_2/n_1)$$

For absorbing materials, the *p*-polarised reflectivity never goes to zero: the angle at which it is a minimum is called the pseudo-Brewster angle, and is not given by the above expression but can readily be found numerically; see `BREWSTER` in Appendix C.

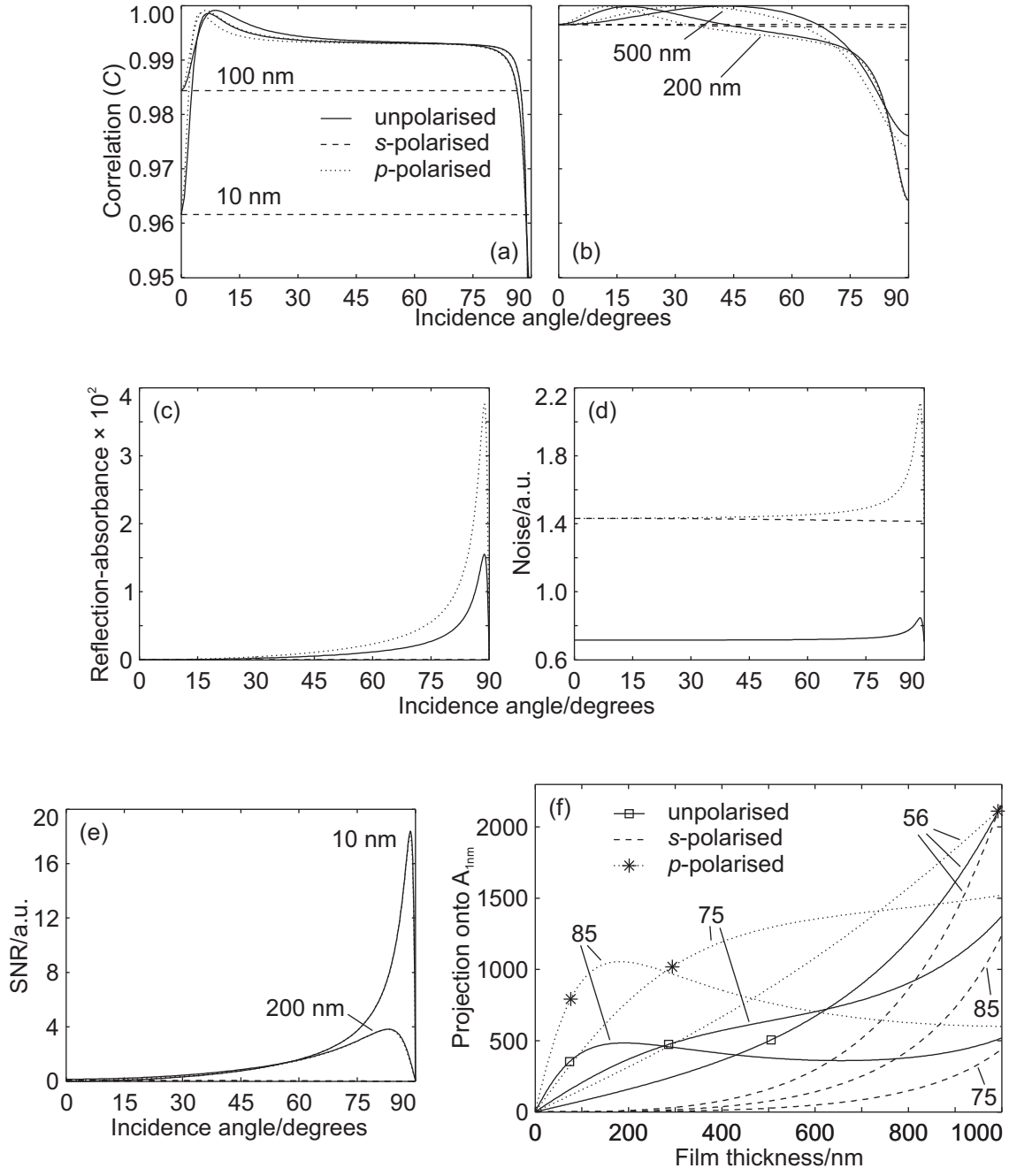

Figure 5.2: Results of calculations for films of benzene on aluminium. Panels a–e contain plots of various properties as a function of incidence angle and polarisation. (a) RA/absorbance correlation for thin films; (b) RA/absorbance correlation for thicker films; (c) Signal (peak RA) for a 10 nm film (the curve for *s*-polarisation runs along RA = 0); (d) Noise (as Equation 5.7); (e) SNR for 10 and 200 nm films. Panel f contains plots of the RA intensity at 56°, 75° and 85° as a function of film thickness and polarisation.

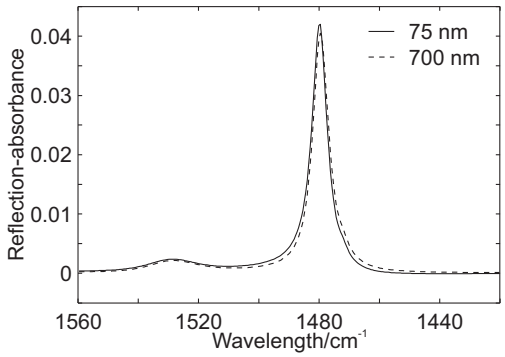

Figure 5.3: Calculated unpolarised, baseline-corrected spectra for films of benzene 75 and 700 nm thick on aluminium. The baselines (not shown) differ, but are essentially constant offsets, which often occur as instrumental artefacts and are not useful for distinguishing the two spectra.

By far the best linearity is obtained by using *p*-polarised light incident at the Brewster angle of the film; if the abscissa of Figure 5.2f were extended, it would be seen that the region of near-linearity extends to ~14 μm.

Near-grazing incidence provides the best sensitivity for very thin films, but poor linearity for thicker films.

Iron: 4000 cm^{-1}

Towards the low-wavenumber end of the mid-infrared, iron is very reflective and behaves similarly to aluminium. Around 4000 cm^{-1}, however, its optical constants are significantly smaller: $n \approx 4$; $k \approx 8$) [123]. Consequently, it is less reflective, and gives somewhat different results as a substrate for IRRAS. Figures 5.4a and b show that, even for thin films, there is significant distortion at small ($\lesssim 30°$) and very large ($\gtrsim 75°$) incidence angles. Interestingly, the distortion lessens as the film thickness increases.

The SNR plot (Figure 5.4e) has several differences from the plot for aluminium. Most obviously, the peak SNR is almost an order of magnitude less. For the thinner film, the peak is at $\theta_0 \approx 80°$. Increasing the film thickness to 200 nm has a less dramatic effect than for aluminium: the peak in the weighted SNR decreases slightly and shifts to slightly smaller incidence angle (~75°), and the difference between the *p*-polarised and unpolarised curves increases slightly as the *s*-polarised contribution becomes more significant. Figure 5.4f shows that, as for aluminium, the smaller incidence angle extends the linear range substantially, even though the difference between the plotted incidence angles is only 5° in this case. The linearity extends to ~800 nm when using unpolarised light incident at 75°. Again, the linearity extends to ~13 μm when *p*-polarised light at 56° incidence is used. However, this angle requires a polariser to be useful: both the sensitivity and the linear range are better at 75° incidence if

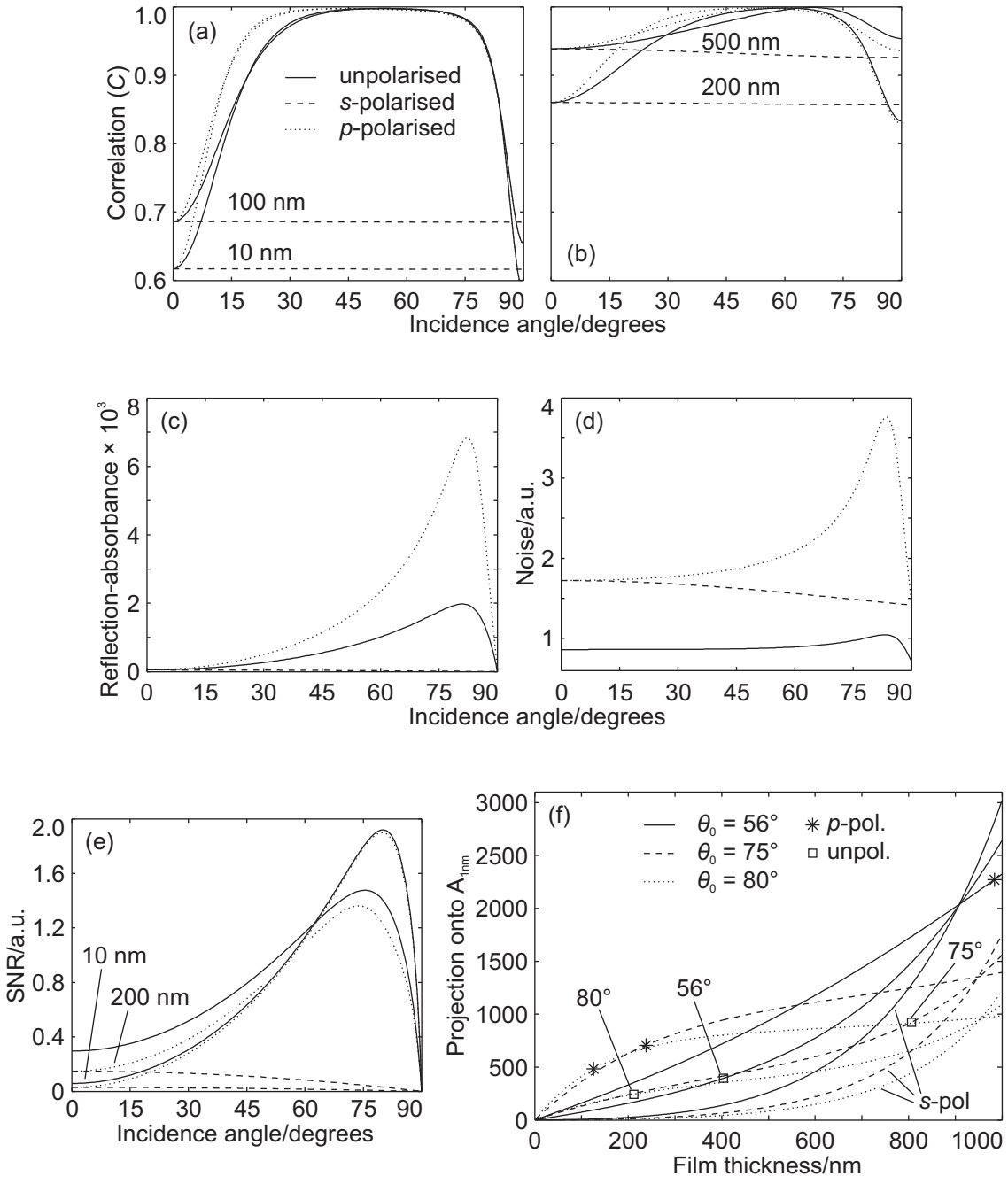

Figure 5.4: Results of calculations for films of benzene on iron (using iron optical constants from ~4000 cm^{-1}. Panels a–e contain plots of various properties as a function of incidence angle and polarisation. (a) RA/absorbance correlation for thin films; (b) RA/absorbance correlation for thicker films; (c) Signal (peak RA) for a 10 nm film; (d) Noise (as Equation 5.7); (e) SNR for 10 and 200 nm films. Panel (f) contains plots of the RA intensity at 56°, 75° and 80° as a function of film thickness and polarisation.

unpolarised light is to be used.

5.2.5 Non-metallic substrates

Silicon

The optical constants of silicon do not vary significantly over the mid-infrared: from 4000 to 1000 cm^{-1} n decreases from 3.44 to 3.42 while k remains $<10^{-3}$ [122].

The correlation coefficients between the absorbance spectrum and the baseline-corrected reflection-absorbance spectra are plotted in Figure 5.5a–b. For films less than 100 nm thick, the s-polarised spectra are negative and essentially undistorted ($C \approx -1$). For p-polarisation at incidence angles $\lesssim 30°$, undistorted negative bands are also seen. However, as the incidence angle is increased the band inverts, passing through a series of intermediate sigmoidal shapes (see Figure 5.6). By ~45°, the band resembles a positive absorbance band. A second inversion occurs at ~75°, near the Brewster angle for the substrate (74°).

For the 100 nm film, these inversions take place more gradually. The unpolarised curves are similar to the s-polarised curves (because of the greater reflectivity for s-polarised light) but with a range of incidence angles, centred on ~60°, at which some distortion occurs. For thicker films, significant distortion is present for almost all combinations of polarisation and incidence angle. The general trends are the same as for the thin films, with s-polarisation giving distorted, negative peaks at all angles and p-polarisation exhibiting negative peaks at extreme incidence angles and positive peaks for intermediate ones.

The RA intensity and noise are plotted in Figure 5.5c–d. The RA is negative at normal incidence (so the SNR is also negative). As the incidence angle is increased, the s-polarised RA decreases monotonically to zero at grazing incidence. The p-polarised RA initially decreases in magnitude with increasing incidence angle, then passes through zero and increases asymptotically as θ_0 nears the Brewster angle. For θ_0 greater than the Brewster angle, the p-polarised reflection-absorbance is negative. The greatest RA intensity is found with p-polarisation near the substrate Brewster angle. However, at these angles the reflectivity of the substrate is very low, and consequently the absorbance noise is much greater. In contrast, the noise for s-polarisation decreases monotonically with increasing incidence angle. The absolute largest SNR (Figure 5.5e) is for p-polarisation at 86°, but a more practical choice would be s-polarisation at $\theta_0 \lesssim 60°$. If unpolarised radiation is to be used, the incidence angle should be kept as close to normal as possible, both to maximise the SNR and to minimise the distortion.

As can be seen in Figure 5.5f, at 15° incidence the linearity between RA and film thickness extends to ~400 nm and there is very little difference between the two polarisations. For 63° incidence, however,

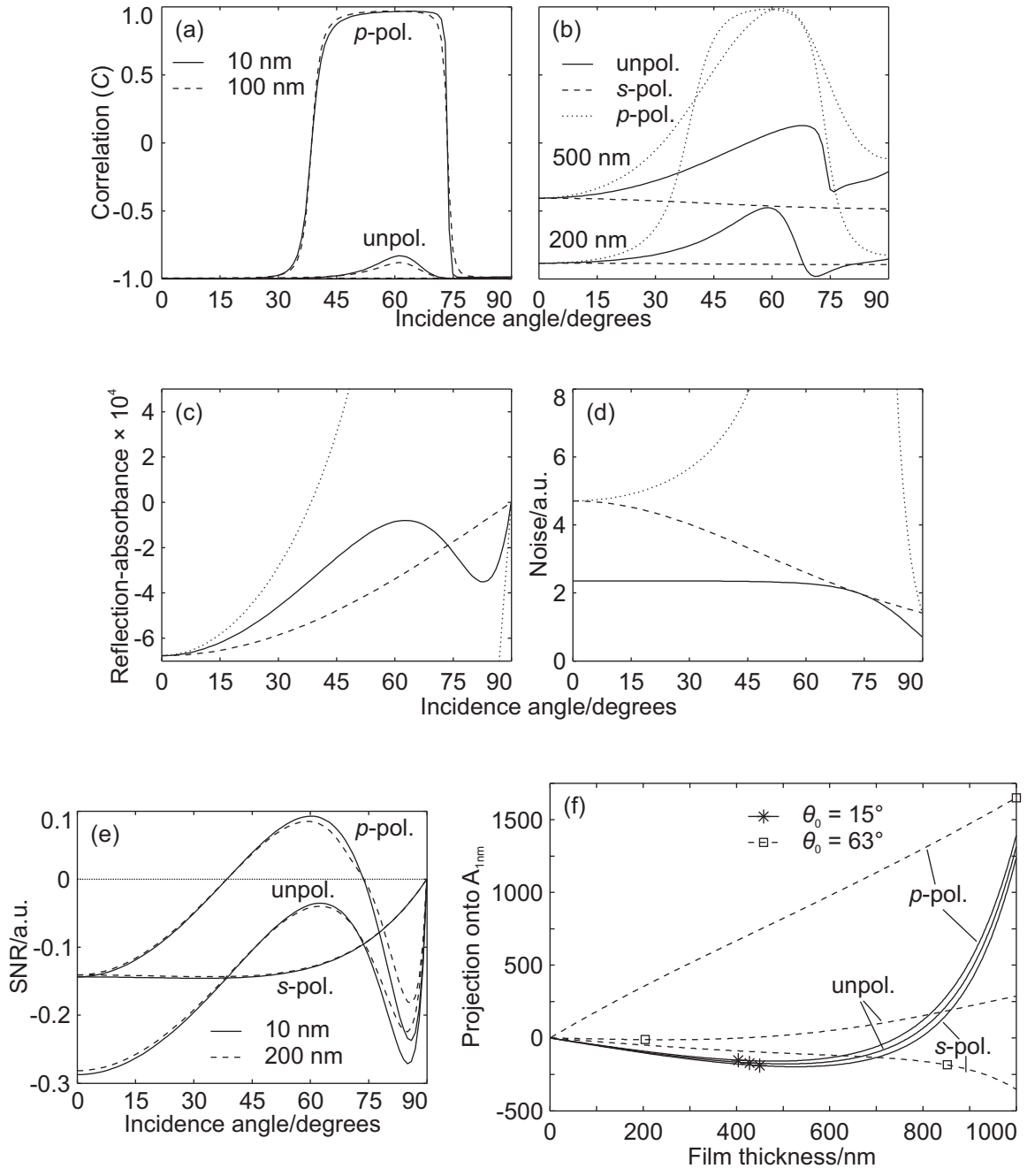

Figure 5.5: Results of calculations for films of benzene on silicon. Panels a–e contain plots of various properties as a function of incidence angle and polarisation. (a) RA/absorbance correlation for thin films (the *s*-polarised curves run along $C = -1$ almost exactly); (b) RA/absorbance correlation for thicker films (note the different line-type coding); (c) Signal (peak RA) for a 10 nm film; (d) Noise (as Equation 5.7); (e) SNR for 10 and 200 nm films. Panel (f) contains plots of the RA intensity at 15° and 63° as a function of film thickness and polarisation.

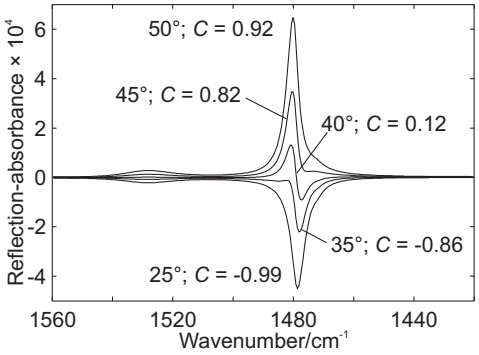

Figure 5.6: Calculated RA spectra (using *p*-polarisation) of a 10 nm film of benzene on silicon. The incidence angles are indicated. *C* is the correlation coefficient between the RA spectrum and the absorbance spectrum of benzene calculated from its optical constants.

the RA increases linearly to 850 nm for *s*-polarisation and to greater than 1000 nm for *p*-polarisation. This may be related to 63° being near to the Brewster angle for reflection between phases with $n_1 = 1.47$ (benzene) and $n_2 = 3.43$ (silicon), which is 66°. It is not immediately apparent why this should be the case, however, and it may simply be a coincidence.

SiO_2 glass

The refractive index of silica glass varies rapidly in the infrared, due to the strong absorption by the Si–O stretching mode at ~1100 cm^{-1} (see Figure 2.3). For this reason, it will be treated as several different substrates in the present investigation. The dominant feature of IRRAS spectra with a glass substrate is a band, which can be positive or negative, or dispersive in shape, centred on the frequency at which the refractive index of glass crosses unity (see, for example, Figures 2.12 and 7.1). Our experimental work has shown that reflection-absorption measurements very near the glass RA feature lack reproducibility, presumably because small local variations in the refractive index of the glass or of the incidence angle cause changes in the baseline that dominate the IRRAS. Consequently, the present investigation is limited to three regions relative to the Si–O reflection-absorbance feature: far to the blue (~3000 cm^{-1}), where glass is transparent; just to the blue (~1500 cm^{-1}), where it is still only weakly absorbing and has a refractive index near unity; and just to the red (~1200 cm^{-1}), on the short-wavelength shoulder of the absorbance band.

SiO_2 glass: 3000 cm^{-1}

At 3000 cm^{-1}, glass, like silicon, is essentially transparent, but glass has a much smaller refractive index of $n_s = 1.4$ [40]. For thin films of benzene on glass, negative, undistorted bands are predicted when using *s*-polarised or unpolarised light at any incidence angle (Figures 5.7a and c). If the light is

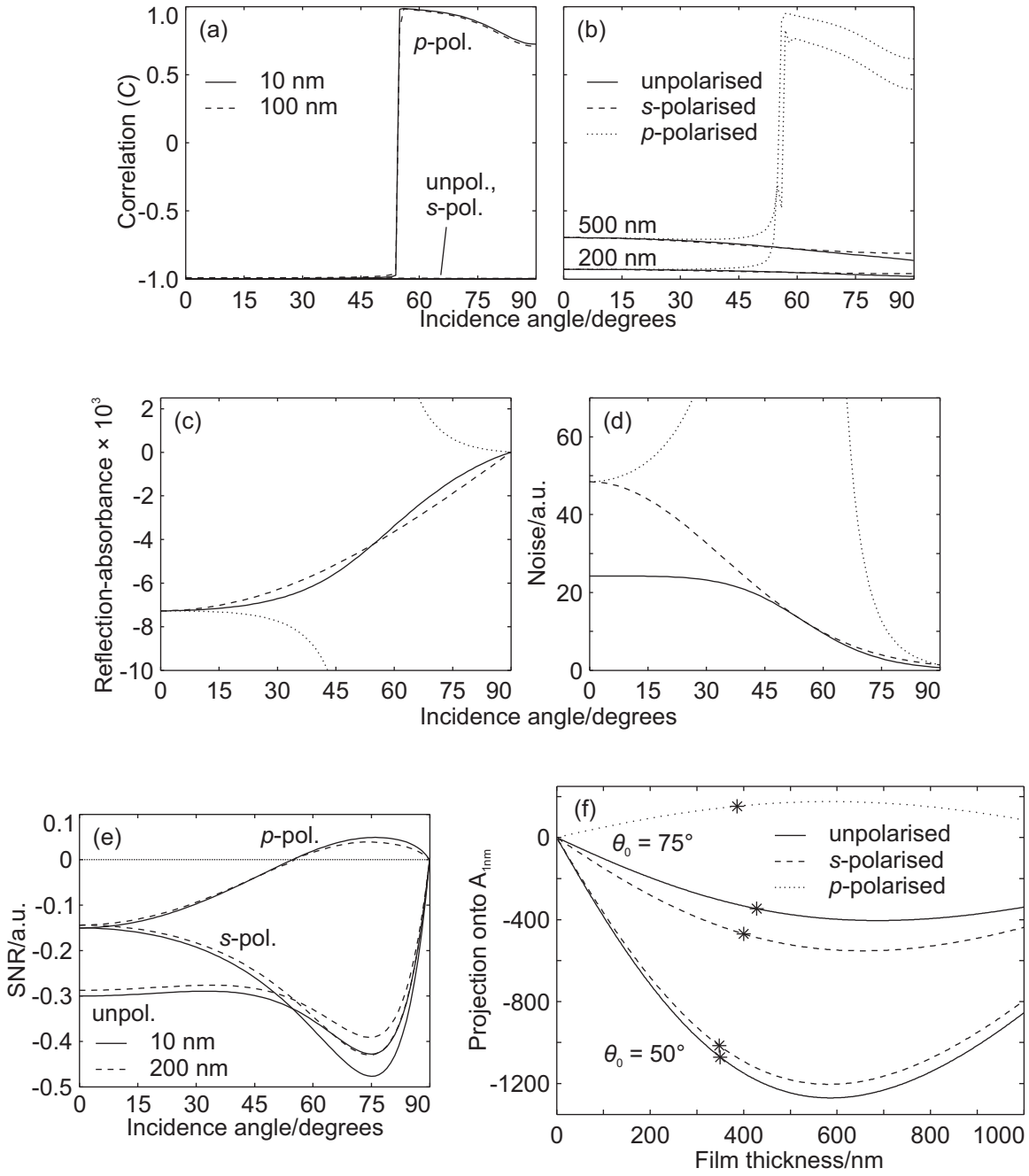

Figure 5.7: Results of calculations for films of benzene on glass (using glass optical constants from ~3000 cm^{-1}). Panels a–e contain plots of various properties as a function of incidence angle and polarisation. (a) RA/absorbance correlation for thin films; (b) RA/absorbance correlation for thicker films; (c) Signal (peak RA) for a 10 nm film; (d) Noise (as Equation 5.7); (e) SNR for 10 and 200 nm films. Panel (f) contains plots of the RA intensity at 50° and 75° as a function of film thickness and polarisation. The *p*-polarised curve for 50° (not shown) is about three times the magnitude of the *s*-polarised one, but the corresponding SNR is very poor.

p-polarised, the bands are negative and undistorted for incidence angles less than the Brewster angle ($\theta_B \sim 55°$) for the substrate. The transition to positive bands at θ_B is much more rapid than for silicon, and the distortion increases as the incidence angle is increased past θ_B. For thicker films (Figure 5.7), the results are similar, except that the extent of the distortion is greater.

Figures 5.7c and d show clearly the divergent behaviour of the *p*-polarised RA and noise about θ_B. The basic trends are the same as for silicon, except that the unpolarised results more closely resemble the *s*-polarised ones. The best SNR (Figure 5.7e) is obtained with *s*-polarised or unpolarised light incident at 75°. At this incidence angle, $d_{99} \approx 400$ nm for *s*-polarised or unpolarised light (Figure 5.7f). For 50° incidence, the RA is about a factor of 2.5 greater, but the SNR is slightly worse; the linearity is similar.

SiO_2 glass: 1500 cm^{-1}

In this region of the spectrum, n_s is decreasing rapidly with increasing wavelength while k_s is increasing slowly. At 1479 cm^{-1}, $n_s \approx 1.2$ and $k_s \approx 8 \times 10^{-3}$ [40], so the substrate is very weakly reflective. Figures 5.8a and b show that the general trends for band distortion are similar to those at 3000 cm^{-1}, the major differences being that in *p*-polarisation the RA band becomes negative again at high incidence angles ($\gtrsim 80°$) and that the distortion encountered on increasing the film thickness is much greater. The SNR (Figure 5.8e) follows the same basic pattern as the previous case, except that the unpolarised results even more strongly follow the *s*-polarised, because the reflectivity for *p*-polarisation is so small. The optimum SNR is now obtained at around 81°, and the advantage over smaller incidence angles is more marked. The threshold for linearity with respect to film thickness is also much smaller, with $d_{99} \approx 175$ nm. The normalised *s*-polarised SNR for the 200 nm film (Figure 5.8e) is significantly smaller (less negative) at all incidence angles than that for the 10 nm film, indicating that there is no incidence angle that will give improved linearity, unless *p*-polarised light, with its much smaller SNR, is used.

SiO_2 glass: 1200 cm^{-1}

This region of the spectrum coincides with the high-wavenumber side of the SiO_2 absorption band and the descending lobe of the corresponding dispersive feature in the refractive index spectrum (see Figure 2.3). At 1179 cm^{-1}, $n_s \approx 0.47$ and is roughly constant over the selected range (1260–1120 cm^{-1}), while $k_s \approx 0.91$ and increases from ~0.2–2 over the selected range [40].

It is apparent from Figures 5.9a and b that some band-shape distortion is introduced when *s*-polarised light is used at any incidence angle. The only exception is for the 500 nm film at near-normal

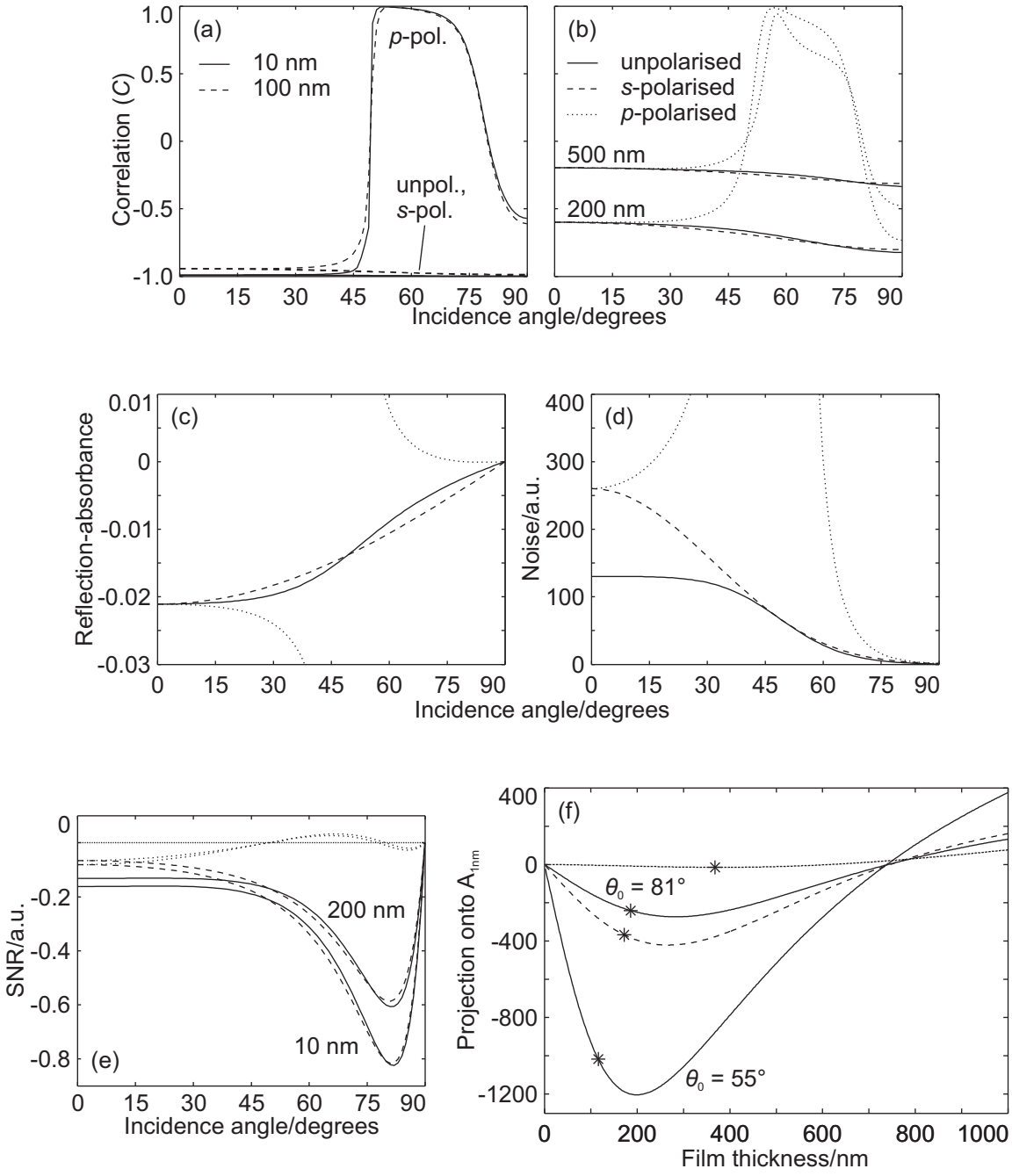

Figure 5.8: Results of calculations for films of benzene on glass (using glass optical constants from ~1500 cm^{-1}). Panels a–e contain plots of various properties as a function of incidence angle and polarisation. (a) RA/absorbance correlation for thin films; (b) RA/absorbance correlation for thicker films; (c) Signal (peak RA) for a 10 nm film; (d) Noise (as Equation 5.7); (e) SNR for 10 and 200 nm films. Panel f contains plots of the RA intensity at 55° and 81° as a function of film thickness and polarisation.

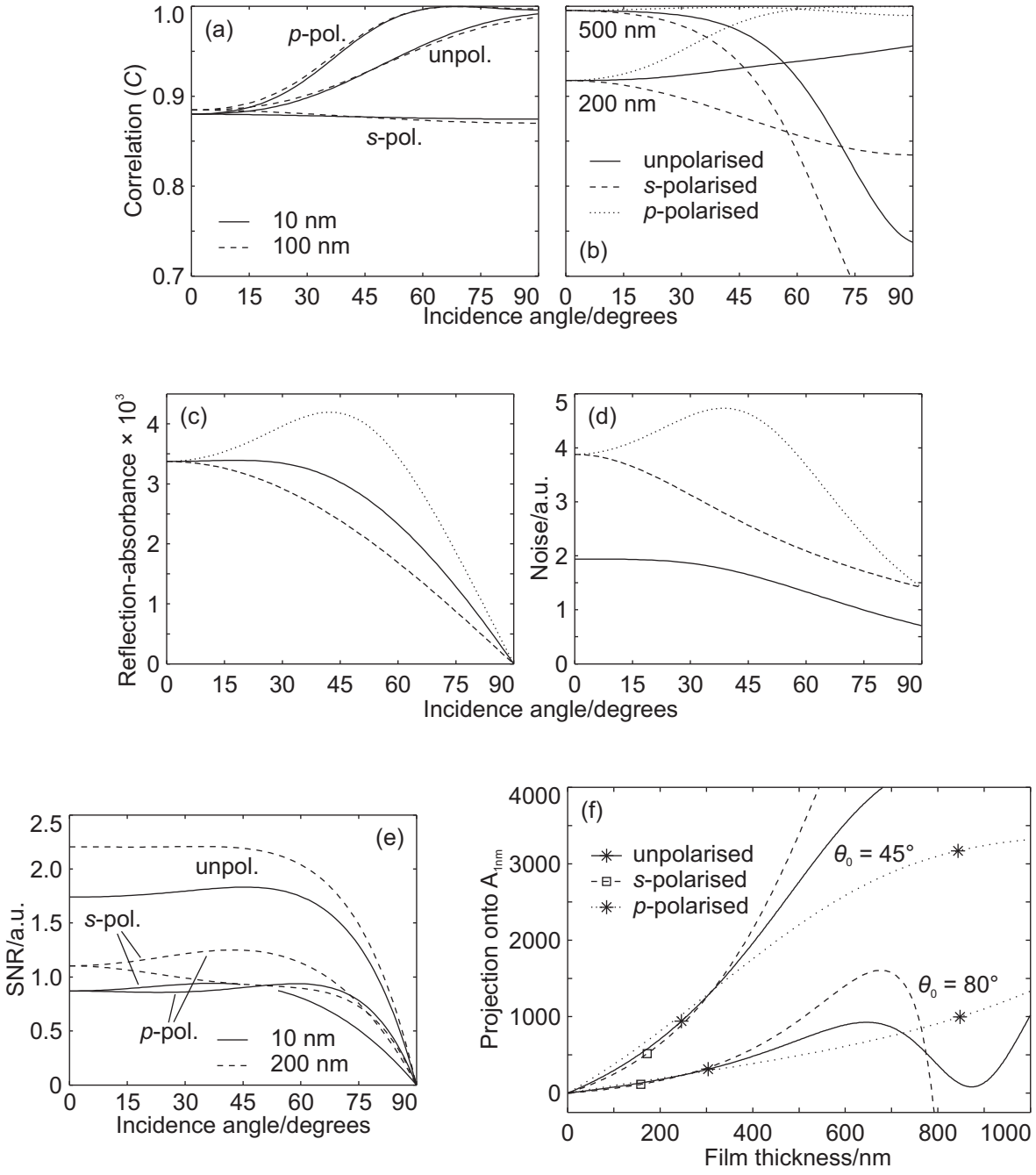

Figure 5.9: Results of calculations for films of benzene on glass (using glass optical constants from $\sim$1200 cm^{-1}). Panels a–e contain plots of various properties as a function of incidence angle and polarisation. (a) RA/absorbance correlation for thin films; (b) RA/absorbance correlation for thicker films; (c) Signal (peak RA) for a 10 nm film; (d) Noise (as Equation 5.7); (e) SNR for 10 and 200 nm films. Panel f contains plots of the RA intensity at 45° and 80° as a function of film thickness and polarisation.

incidence. At high incidence angles, p-polarised light gives relatively undistorted spectra. In all cases the bands are positive, in contrast with the other regions of the spectrum for this substrate. This figure also illustrates the fact that the unpolarised spectrum contains significant contributions from both the s-polarised and the p-polarised spectra.

Because k_s is large, the reflectivity minimum for p-polarised light does not approach zero ($R_p \approx 0.3$ at the pseudo-Brewster angle[4] of 39°). Consequently, the divergent behaviour of the RA signal and noise around the Brewster angle is strongly damped (Figures 5.9c and d) compared to the transparent regions (Figures 5.7 and 5.8 panels c and d).

The resulting SNR plot (Figure 5.9e) has a number of interesting features. The differences between s- and p-polarisation are not as marked as in the other cases, and the dependence on incidence angle is relatively weak between normal incidence and ~60°; as the angle is increased, the SNR falls towards zero at grazing incidence. The unpolarised SNR is about a factor of two greater than in either of the other wavenumber ranges examined. Particularly interesting is the result that the SNR is greater for the 200 nm film than for the 10 nm film. This is also apparent from Figure 5.9f. At both of the chosen incidence angles, linearity in the s-polarised RA is limited to ~175 nm; beyond this, the RA actually increases faster than linearly (before, in the 80° case, going through a maximum, plummeting and becoming negative at ~700 nm). Linearity extends slightly further with unpolarised light, and to ~900 nm when p-polarised light is used.

5.2.6 Discussion and conclusions

Metallic substrates

The best sensitivity for metallic substrates is generally obtained at near-grazing incidence with p-polarised or unpolarised light. However, much better linearity with respect to film thickness is obtained if the incidence angle is reduced somewhat. Particularly good linearity is obtained for p-polarised light incident at the Brewster angle of the film (which is ~55° for many organic materials). This phenomenon can be understood physically in terms of the multiple-reflection model (Figure 2.4). In the limit of small k for the film, p-polarised light incident at θ_B is not reflected from the front face of the film: it is entirely refracted into the film at an angle θ_1, which can be calculated approximately by Snell's law. The light passes through the film (with a small amount being absorbed) then is almost completely reflected from the metal substrate before passing through the film again. It can be shown[5] that θ_1 is the Brewster

[4] See Footnote 3 on page 98 for the definition of the pseudo-Brewster angle.

[5] If $n_0 = 1$ and $\theta_0 = \theta_B^{0,1} = \tan^{-1} n_1$, then $\theta_1 = \sin^{-1}(\sin(\theta_0)/n_1)$ and $\theta_B^{1,0} = \tan^{-1}(1/n_1)$. Using the trigonometric identities [124] $\tan(\sin^{-1} x) = x/\sqrt{1-x^2}$ and $\sin(\tan^{-1} x) = x/\sqrt{1+x^2}$, it can be shown that $\theta_1 = \theta_B^{1,0}$.

Table 5.1: Results of RA calculations for the ν_{13} band of films of benzene. θ_0 is the incidence angle in degrees. The SNR is in arbitrary units; d_{99} is defined in the text and is given in nanometres. The missing values for *s*-polarisation and metallic substrates are all very near zero.

Substrate	$\bar{\nu}$/cm^{-1}	n	k	θ_0	Unpol.		*s*-pol.		*p*-pol.	
					SNR	d_{99}	SNR	d_{99}	SNR	d_{99}
Al	1479	13	64	56	13	506			13	13516
				75	37	286			37	294
				85	100	74			100	76
Fe	1479	4.5	20	56	12	470			12	13408
				75	32	282			32	270
				85	61	68			60	62
	3979	4.2	8.1	56	9.7	404			9.5	13096
				75	18	806			18	238
				80	19	212			19	126
Si	1479	3.42	$\sim 10^{-5}$	15	-2.6	428	-1.4	450	-1.2	404
				63	-0.3	204	-1.2	852	0.9	1736
SiO_2	1179	0.47	0.91	45	18	246	9	172	9	844
				80	9.8	304	3.6	158	6.2	848
	1479	1.16	$\sim 10^{-2}$	55	-2.5	128	-2.6	124	0.1	374
				81	-8.1	186	-8.0	172	-0.1	368
	2979	1.4	0	50	-3.1	350	-2.9	348	-0.2	232
				75	-4.2	428	-4.7	400	0.5	386

angle for light encountering the film-air boundary from inside the film, so all the light passes through the boundary and to the detector. Thus, to the extent that k is very small, this configuration acts like a transmission experiment: there is no reflection component to introduce nonlinearity. The effective path length is somewhat longer than the geometric path length ($2d/\cos\theta_1$), because of the enhanced electric field intensity near the metal surface (Figure 2.6).

In Figure 5.10, plots similar to that in Figure 5.2f are shown for thicker films and for two weaker bands of benzene. For each band, the projection varies sinusoidally about the linear fit. These deviations are due to the nonzero values of k. Even in the case of the ν_{13} band of benzene, which is quite strong for an organic compound, the deviation from linearity is not extreme. The period of the oscillations is ~0.42 times the wavelength of the band maximum. For the ν_{13} band, the RA falls below the line at around 13 μm. The same occurs at much greater thicknesses for the other bands, since the importance of the small reflection components increases with the proportion of light absorbed by the film.

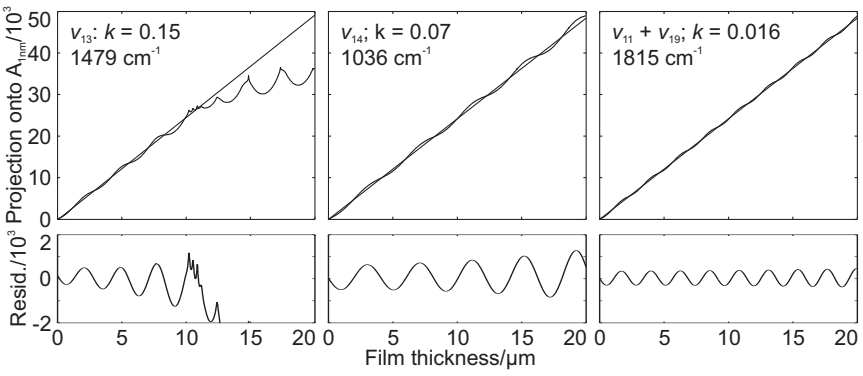

Figure 5.10: Plots demonstrating the linearity of Brewster-angle metallic RA with respect to film thickness. Calculations are as described in the text and plotted in, for example, Figure 5.2f. Three bands of benzene are chosen, decreasing in intensity from left to right. The plots in the upper row are of projections of the p-polarised RA spectrum onto the 1 nm-film transmission-mode absorbance spectrum; those in the lower row are of the residuals from the linear fits.

The efficacy of p-polarised Brewster-angle measurements can also be predicted from the formula for the reflectance of the air-film-substrate system. Equation 2.23 gives the Fresnel reflection coefficient as

$$r_{\text{tot}} = \frac{r_1 + r_2 e^{2i\delta_1}}{1 + r_1 r_2 e^{2i\delta_1}}$$

where r_1 and r_2 are the Fresnel reflection coefficients for the air-film and film-substrate interfaces (near zero and unity, respectively) and

$$\delta = 2\pi(n_1 + ik_1)\cos(\theta_1)d/\lambda$$

If k_1 is small and θ_0 is the Brewster angle for the film, then $\theta_1 \approx \sin^{-1}(\sin(\theta_0)/n_1)$, and $r_1 \ll 1$ for p-polarisation. Consequently, Equation 2.23 can be approximated by

$$r_{\text{tot}} \approx r_2 e^{2i\delta_1} \tag{5.4}$$

The exponent $2i\delta_1$ has an imaginary component proportional to n_1 and a negative, real component proportional to k_1. The reflectance is given by $R = |r_{\text{tot}}|^2$, to which only the real part of the exponent contributes.

The reflection-absorbance is thus given approximately by

$$\begin{aligned} \text{RA} &= \log_{10} \frac{R_0}{R} \\ &\approx \log_{10} \frac{R_0}{|r_2|^2 e^{-8\pi k_1 \cos\theta_1 d/\lambda}} \end{aligned} \tag{5.5}$$

Equation 5.5 is a linear function of d with a small intercept (due to the difference between R_0 and $|r_2|^2$) and a slope proportional to k_1. The periodic deviations from this function, seen in Figure 5.10, are due to the neglected term ($r_1 r_2 e^{2i\delta_1}$) in the denominator of Equation 2.23. As the film thickness is increased, the neglected term in the numerator (r_1) gains in importance: for completely absorbing films, $r_{\text{tot}} = r_1$ and is independent of film thickness.

For Brewster-angle RA measurements to provide significantly better linearity with respect to film thickness than other incidence angles, it is important that r_2 be large. This condition is met well for organic films on metals, and somewhat for organic films on dielectric or semiconductor materials with high refractive indices. For a glass substrate, however, the linear approximation is not at all valid.

Dielectric substrates

For dielectric substrates with high refractive indices (and consequently high reflectivity), such as silicon, both polarisations contribute significantly to the optical throughput and to the reflection-absorbance when no polariser is used. At small incidence angles and for thin films, negative, relatively undistorted RA is observed for both polarisations; and unpolarised measurements give the best SNR. At greater ($\gtrsim$30°) incidence angles, the p-polarised spectrum exhibits severe distortion, except in a window of angles around 60°. Use of this range of angles with p-polarised light gives the best linearity with respect to film thickness (since it is near to the film Brewster angle); the SNR is also reasonable.

Substrates (such as glass) that have smaller refractive indices tend to give the best SNR with s-polarised or unpolarised light, because the optical throughput is low for p-polarisation. The optimum

RA conditions for glass depend strongly on where the analyte absorbance bands fall. In the C–H stretch region (Figure 5.7), an incidence angle of around 75°, with either *s*-polarised or unpolarised light, is preferred for both sensitivity and linearity. In the portion of the fingerprint region where $n_s > 1$ (Figure 5.8), 81° is best, but has only a slight advantage over 75°. Further to the red, near the middle of the Si–O absorption band (Figure 5.9), unpolarised light offers the best SNR because both polarisations give positive RA of similar intensity. An incidence angle around 45° appears to offer the best linearity. In this region, the Brewster-angle method discussed above would be applicable for weak bands, because the reflectivity of the substrate is quite high. The high incidence angle preferable for the other spectroscopic regions performs poorly here in terms of SNR, but fairly well in terms of band-shape distortion and linearity.

Dynamic range issues

When comparing the SNR for the dielectric surfaces to that for the metallic ones, there is another factor that must be considered. As discussed by Kattner and Hoffmann [120], a sensitive MCT detector can fill the dynamic range of the ADC with only a small portion of the full source intensity. This means that the higher optical throughput obtained with a metallic substrate is not as helpful as the calculations would indicate. Even with the fibre-optic system used in this work, which significantly reduces the throughput, it is sometimes necessary to further reduce the amount of light reaching the detector to prevent saturation of the ADC. If a sensitive detector is used, the magnitude of the RA becomes more important (that is, the contrast between the background and the sample single-beam spectra) for determining the SNR when the substrate is highly reflective. The high reflectivity of metallic substrates at all incidence angles, coupled with the relatively low RA for incidence angles far from grazing, may result in much poorer sensitivity for measurements at the Brewster angle for the film.

5.3 Design of the variable-angle probe

To test the applicability of calculations like those described in the previous section to real systems, a fibre-optic variable-angle reflectance probe was built, using the grazing-angle probe (Section 4.2) as the starting point for the design. The probe was fabricated by Danny Leonard in the Mechanical Workshop of the Department of Chemistry. The parabolic mirrors are mounted inside blocks of aluminium, to which the fibre bundle and MCT detector are attached in such a way that translation in three dimensions is possible (see Figure 5.11). Each of the aluminium blocks is mounted on rails (arms), which are in turn attached independently to a framework. Also attached to the framework is a sampling tray.

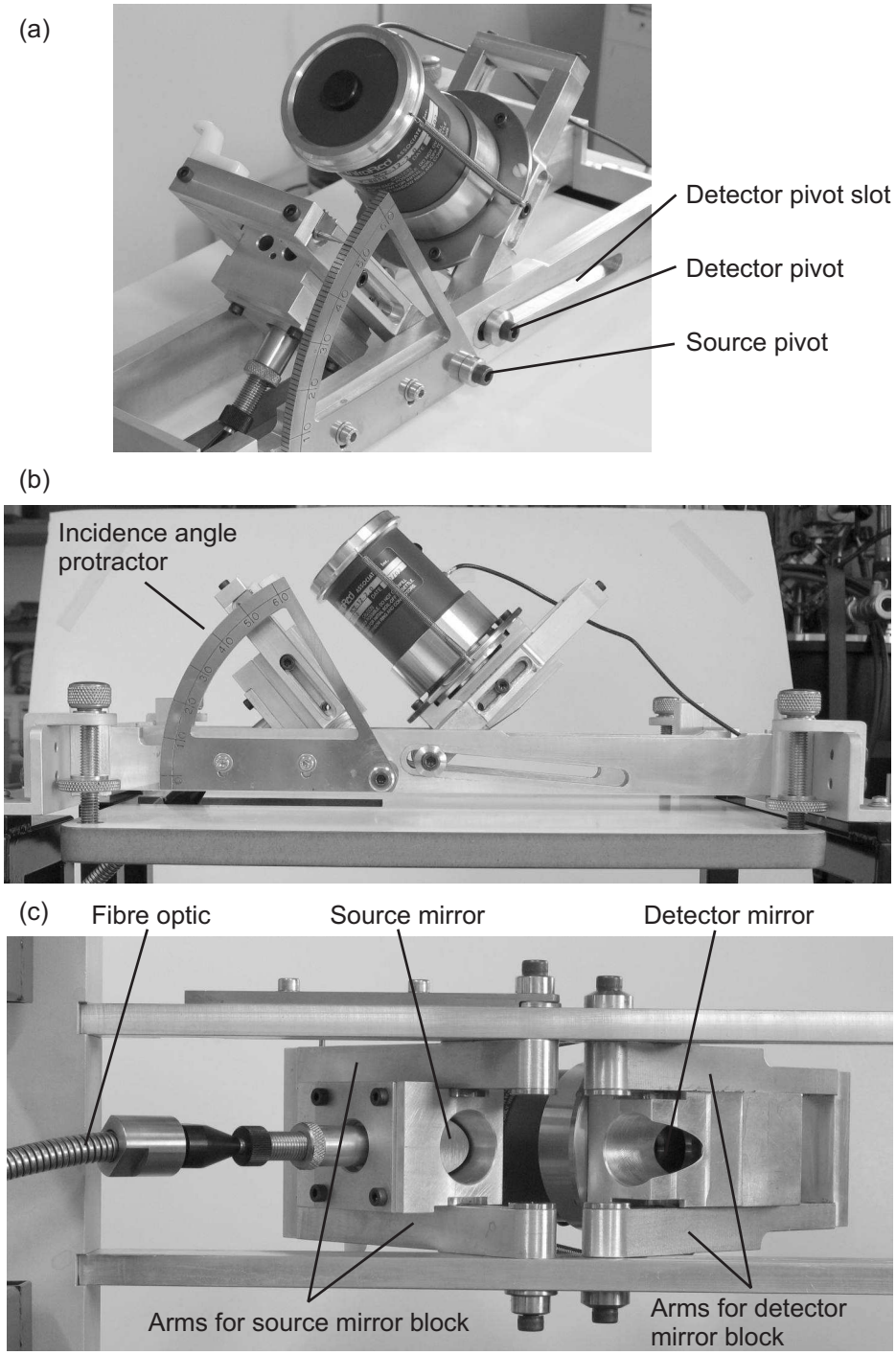

Figure 5.11: Photographs of the variable-angle probe. (a) Three-quarter view. (b) Side view. (c) View from underneath.

If the source and detector arms share the same pivot, that pivot must be level with the sample; otherwise, the area illuminated by the source will not coincide with the area seen by the detector. To avoid placing restrictions on the size of the sample, an alternative strategy was used. The arms are mounted independently, and the pivots are raised from the sample. As the source arm is moved towards grazing incidence, the illuminated area increases in size and moves towards the detector side of the apparatus (the right-hand side in Figure 5.11). To accommodate this shift, the detector-arm pivot may be moved along a slot cut into the frame (visible in Figure 5.11a–b).

The incidence angle can be varied over the range 30–80°. The source and detector blocks can be moved along their respective arms: for most of the range they can be left close to the sample, but angles smaller than about 40° require that they be moved outwards a few centimetres. The disadvantages of this design are that the path length changes as the incidence angle is varied (see Section 5.4.4 below), and that an extra adjustment must be made when changing the angle.

When the source arm is positioned at the desired incidence angle, the detector-arm horizontal position and angle are adjusted to maximise the ADC count (with a gold or aluminium mirror in the sample position). The approximate position can be found by setting the angle to match the source (with a protractor) and sliding the arm along the slot until the maximum is found.

5.4 Experimental procedure

The aim is to determine the dependence of the SNR on the incidence angle. The steps in the investigation are (for each sample) as follows:

1. Prepare a standard via the spray method (Section 4.3.2).
2. Measure spectra at a range of incidence angles.
3. Determine the loading of the standard via UV colorimetry (Section 4.3.3).
4. Calculate signal and noise measures.
5. Correct the SNR for instrument throughput differences and normalise.

Details for steps that have not been described fully elsewhere are given below.

5.4.1 Measuring spectra with the variable-angle probe

So that the SNR measurement can be corrected according to the optical throughput, the latter quantity must be measured during the experiment. This is achieved by measuring the ADC count with a gold or

aluminium mirror as the sample.

The maximum throughput is found at the smallest angle that can be used before the source block must be moved outwards to accommodate the detector; this is about 35–40°. First, the source arm is set to 40° and the angle and translation of the detector arm are adjusted so that the ADC count with the mirror in place is at a maximum. If the detector/ADC is saturated, the position of the spectrometer end of the fibre bundle is adjusted (along the fibre axis, keeping the lateral position optimal) so that the ADC count is about 30000. This ensures that saturation will not occur at any incidence angle.

For each angle, the ADC count with the metal mirror in place is recorded. Then, the mirror is replaced with a clean coupon of the substrate material and the background spectrum is recorded. Finally, spectra are measured from several places on the sample coupon.

5.4.2 Signal measures

A straightforward signal measure is obtained by calculating a band integral after an appropriate baseline correction has been made. The baseline correction has made by picking a point on either side of a band or group of bands, drawing a line between them and subtracting this line. The band is then integrated by the trapezium method, in which

$$\int_a^b f(x)\,\mathrm{d}x \approx \frac{\delta x}{2}\left(f(x_1) + f(x_n) + 2\sum_{i=2}^{n-1} f(x_i)\right) \tag{5.6}$$

where the x_i are evenly spaced by δx, $x_1 = a$ and $x_n = b$. A more sophisticated method for determining the band integral is to use nonlinear regression to fit band-shape and baseline functions to the data and then to integrate the fitted function analytically. This alternative approach would be potentially more precise if the forms of these functions were known, since the noise in the data would be averaged out in the estimation of a small number of parameters; with direct integration, the noise is incorporated directly into the result. However, in IRRAS, the band shapes differ from the standard functions offered by fitting software, so it would be necessary to write custom software. Given that the noise is not excessive (see Figures 5.13,and 5.19), the small potential improvement is outweighed by the vast increase in complexity.

Water vapour absorption bands are a significant problem in the calculation of band integrals, so derivative-minimisation subtraction of reference spectra (see Appendix B) has been used as a standard processing step. Several water vapour spectra were measured and, for each sample spectrum, the water vapour spectrum giving the most improvement (as measured by the reduction in the integrated absolute derivative) was used.

5.4.3 Noise measures

It has been shown in Section 4.2.5 that, for this spectrometric system, the noise is inversely proportional to the single-beam signal level. The simplest way to measure the noise is therefore to take the reciprocal of the single-beam intensity at the wavenumber of interest. This is the only straightforward way of measuring the noise without making additional "100 % line" measurements at each incidence angle, and has the added advantage of much greater precision. Since the aim is to determine a relative SNR, absolute noise measurements are not necessary, and the relative noise is calculated (as in Equation 5.1) as

$$N \propto \sqrt{\frac{1}{R_0^2} + \frac{1}{R^2}} \tag{5.7}$$

where R_0 and R are the mean values of the background and sample single-beam spectra over the wavenumber range spanned by the band. (The data in Figures 4.11a and b could be used to derive a relationship between the ADC count and number of interferometer scans and the noise, allowing the absolute SNR to be determined.)

5.4.4 Optical throughput correction

The variation of the path length with incidence angle is plotted in Figure 5.12. Panel (a) is a plot of atmospheric CO_2 absorbance (calculated from single-beam spectra by fitting a baseline) against incidence angle. With the exception of 30°, the path length increases roughly exponentially with increasing incidence angle. This trend is in good agreement with the geometrically calculated path length plotted in panel (b) (with the 30° point excepted, $C^2 = 0.999$ for these data). Depending on the divergence of the beam, a greater path length may reduce the optical throughput. The ADC count with a metal mirror is plotted in Figure 5.12c, from which it can be seen that the throughput is negatively correlated with the path length (excluding the 30° values, $C^2 = 0.99$).

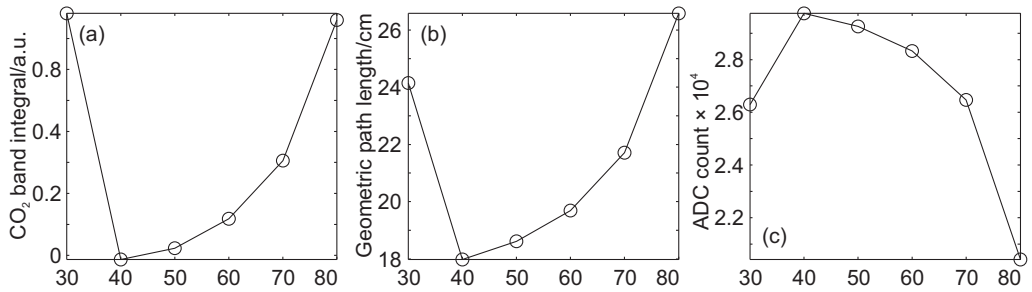

Figure 5.12: Dependence of the path and optical throughput length on incidence angle. (a) CO_2 band integral in a single-beam absorbance spectrum. (b) Path length calculated geometrically. (c) ADC count (aluminium mirror).

Since the dependence of the path length on the incidence angle is an artefact of the design of the instrument, it should be taken into account when comparing signal levels at different incidence angles. The correction is made by dividing the SNR at each incidence angle by the ADC count obtained from a metal mirror at the same angle.

5.4.5 Normalisation of the SNR

Finally, the SNR is normalised to allow comparison between different bands and different samples. Since there are several measurements for each angle, there are several ways to carry out the normalisation. The method adopted here is to calculate the mean SNR for each angle and to divide all the SNR values by the mean SNR having the largest absolute value. This normalisation ensures that the SNR measurements at the best incidence angle have unit mean.

5.5 Acetaminophen on glass

5.5.1 Preparation of the standards

Three standards were prepared by the spray method, as described in Section 4.3.2, on $16 \times 16\,\mathrm{cm}^2$ glass coupons. After the IRRAS measurements, the loadings were determined by UV colorimetry (Section 4.3.3) to be 0.83, 1.29 and 2.11 $\mu\mathrm{g\,cm}^{-2}$. The substrate material was soda-lime float glass (window glass), which has somewhat different optical properties in the infrared than pure silica glass [125]. This point is elaborated in Section 5.5.3 below.

5.5.2 Representative spectra

Some typical (unpolarised) spectra are plotted in Figure 5.13. There is no obvious change in the shape of the spectrum as the incidence angle is varied. This is as expected from Figure 5.8a, which shows that the unpolarised IRRAS with a glass substrate in this wavenumber range remains almost undistorted at all incidence angles. The trend of decreasing reflection-absorbance with increasing incidence angle (also seen in Figure 5.8c) is evident. It is also apparent that the noise is much less at the higher incidence angles, in agreement with Figure 5.8d. The reason for this is clear from the single-beam spectra (plotted in Figure 5.14): the signal level is much greater at higher incidence angles. The two features dominating the single-beam spectra are absorption by water vapour between about 2000 and 1300 cm^{-1} and the reflectance band of the glass at about 1250 cm^{-1}. The spectroscopic features of the acetaminophen layer are indiscernible in the single-beam spectra. It should also be noted that the signal effectively reaches zero at about 1280 cm^{-1} for incidence angles less than 70°.

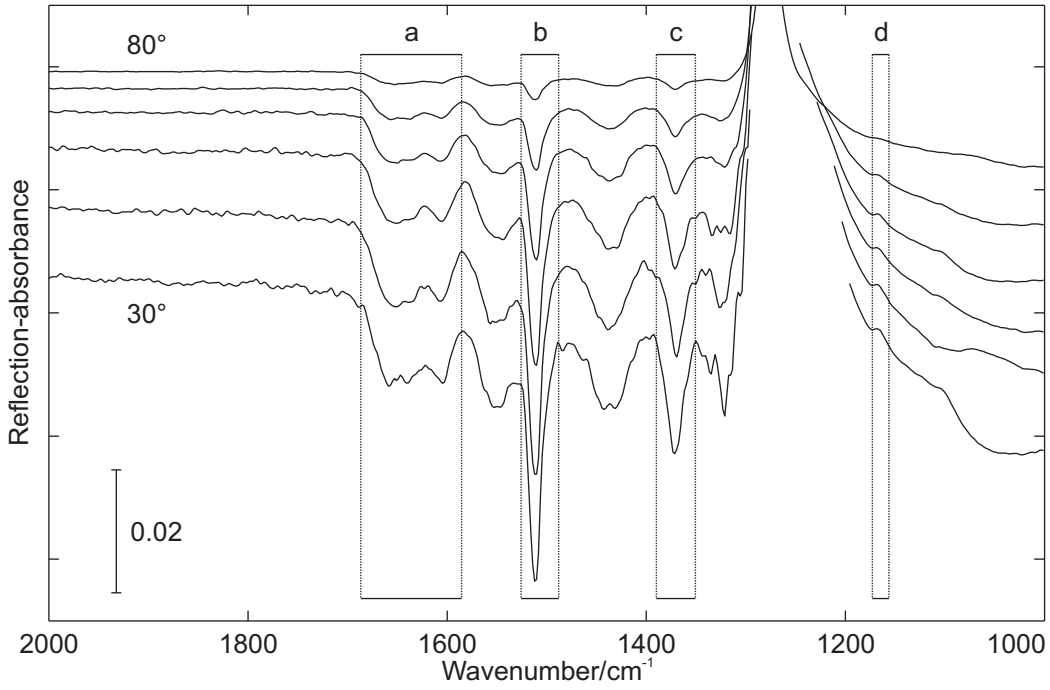

Figure 5.13: Some typical IRRAS of acetaminophen on glass (0.83 $\mu g\,cm^{-2}$). The incidence angle was varied in 10° increments from 30° to 80°. The spectra are offset for clarity; the scale is indicated. Regions for integration are indicated by the dotted vertical lines.

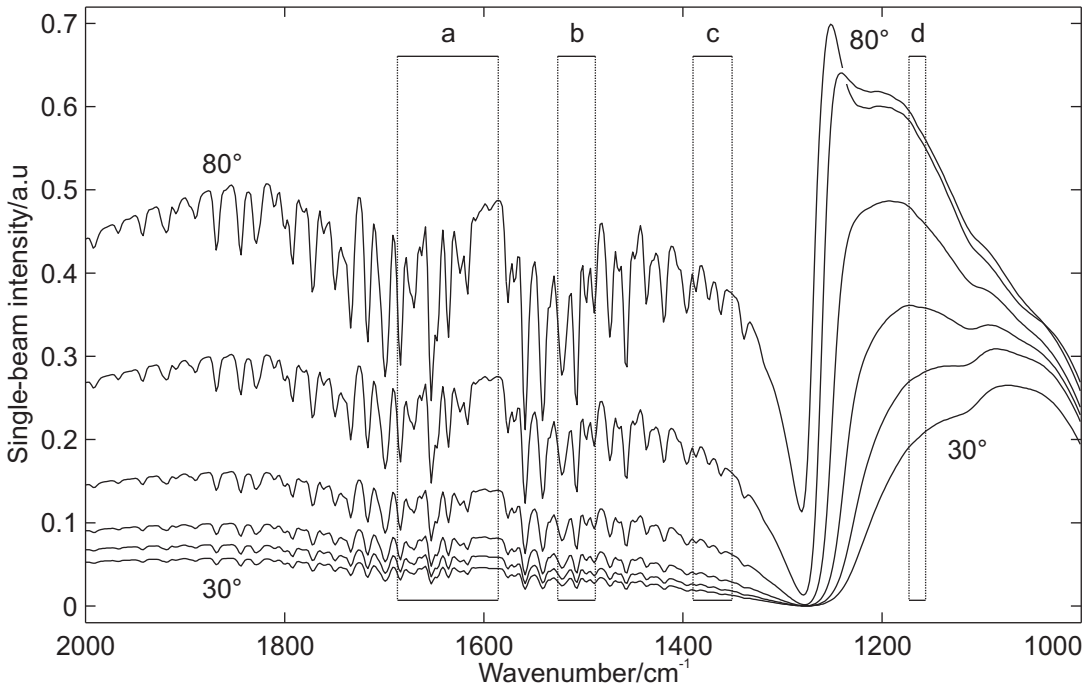

Figure 5.14: Some typical single-beam spectra of acetaminophen on glass (0.83 $\mu g\,cm^{-2}$). The incidence angle was varied in 10° increments from 30° to 80°. Regions for integration and are indicated by the dotted vertical lines.

5.5.3 Signal and noise calculations

Four regions (labelled as a–d in Figure 5.13) are chosen for integration. The band integrals and noise (Equation 5.7) for one sample are plotted in Figure 5.15, showing the expected trends of decreasing magnitude of both the signal measure and the noise measure as the incidence angle is increased. The corresponding plots for the other two samples (not shown) are essentially the same.

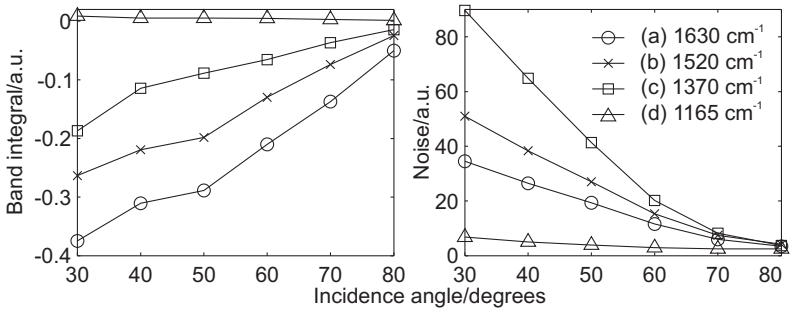

Figure 5.15: (a) Signal (RA band integral) and (b) noise measurements for 0.83 μg cm^{-2} acetaminophen on glass.

The SNR curves for all three samples and four band integrals are plotted in Figure 5.16, where they are compared with calculated SNR curves for thin benzene films on a pure silica glass substrate (dashed lines). The calculated curves were taken from Figures 5.7e, 5.8e, and 5.9e and normalised, in the same way as the experimental SNR curves, to have maxima of unity. Both the film and the substrate materials differ between the calculations and the experiments, so the results are not expected to agree quantitatively.[6]

Bands a–c all show a trend of increasing SNR with increasing incidence angle, in agreement with the calculations. This trend is similar in magnitude for bands a and b, and more pronounced for band c. The reason for the latter observation is related to the shape of the substrate refractive index spectrum, n_s. These bands are located just to the blue of the Si–O absorbance band (see Figure 2.3); in this region, n_s is decreasing with decreasing wavenumber. As n_s decreases, so does the reflectance, and this effect is more pronounced at small incidence angles than at large ones. The noise depends inversely on the reflectance, so the nett result is that decreasing n_s leads to an SNR curve more strongly favouring large incidence angles.

The choice of which calculated curve to plot for each band was made based on the fit to the experimental data, rather than on matching the wavenumber. For bands a and b, the best fit was obtained for

[6]It should be noted, though, that the agreement would be somewhat better than it is if the experimental data were normalised differently. The present approach is very sensitive to errors because the normalisation factor is determined from measurements at a single angle. Since the comparisons were intended to be qualitative, this is not regarded as a significant problem.

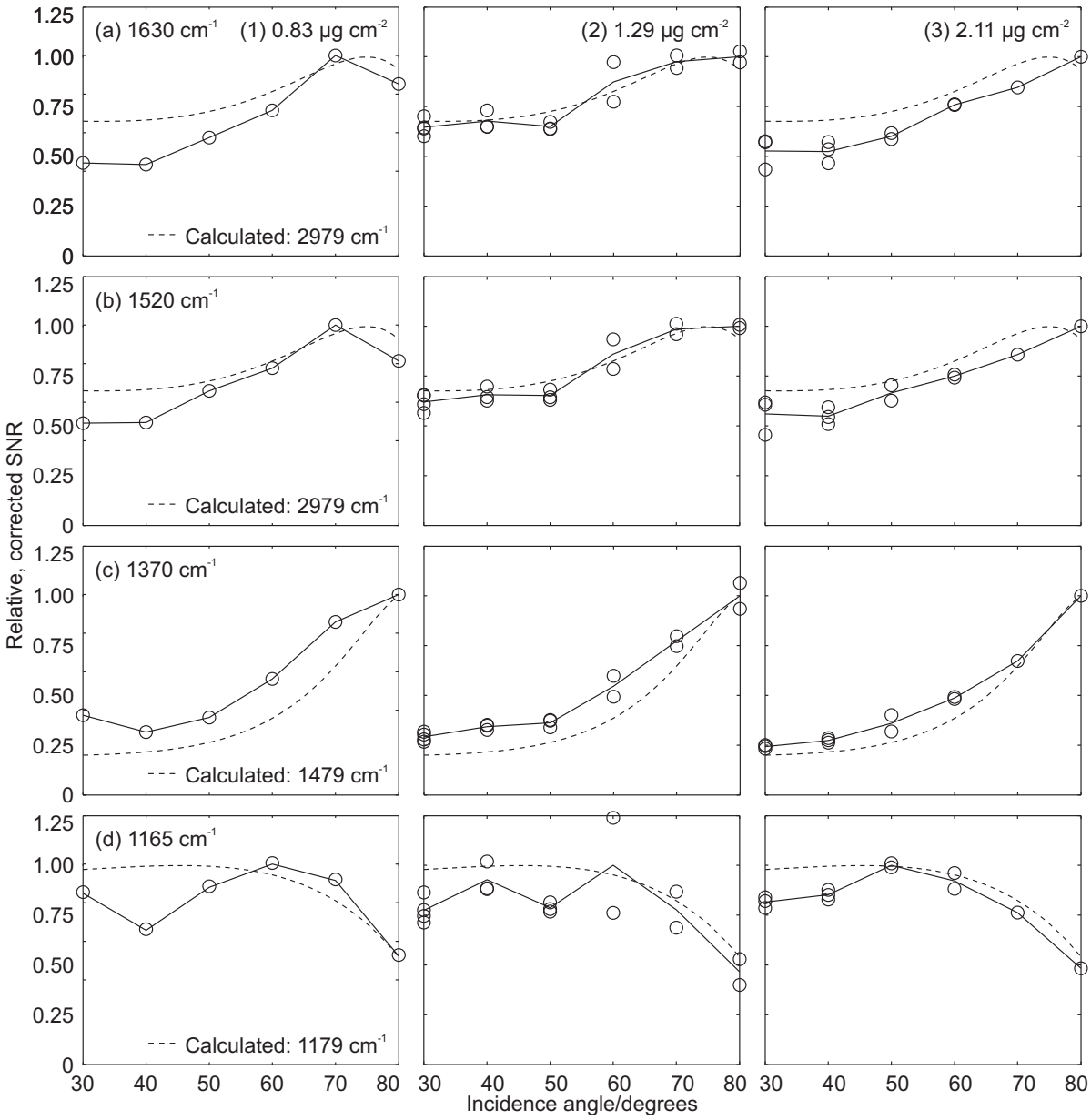

Figure 5.16: Relative, corrected SNR for acetaminophen on glass. Each column represents a sample (loadings are indicated in the top row) and each row represents a band (label and wavenumber range are indicated in the left column). The dashed lines are calculated values for benzene on pure silica glass, excerpted from Figures 5.7e (for bands a and b), 5.8e (for band c) or 5.9e (band d). Circles represent individual measurements; the solid lines join the mean values for the measurements at each angle.

the calculations pertaining to silica glass at ~3000 cm^{-1}, roughly twice the frequency of the bands themselves. However, the refractive index of glass changes very little from 3000 cm^{-1} to ~1700 cm^{-1} [40], at which point it starts descending into the downward lobe of the dispersion accompanying the Si–O absorption. Because the Si–O band is shifted to longer wavelength in glasses containing significant amounts of sodium (such as the window glass used here) [125], the refractive index of window glass over the wavelength range encompassing bands a–c is likely to be higher than that of pure silica glass. Accordingly, using the value of n_s at 3000 cm^{-1} for bands a and b is reasonable.

The band integral measurements for band d are much less reliable than the others, because the band is so much weaker. Nevertheless, they show same general trend as the calculations: the SNR remains roughly constant for $\theta_0 \lesssim 60°$ and then falls off as the incidence angle is increased further.

5.6 Electrostatically self-assembled multilayer on glass

Two shortcomings of the results just discussed are that there no easily integrable bands at high wavenumber, and that, where more than one spectrum has been measured for a given sample and incidence angle, there is considerable scatter in the results. The first problem can be rectified simply by choosing a different analyte, but the second is at least partially due to the heterogeneity of samples prepared by airbrushing, which is somewhat more difficult to improve. From a brief survey of the literature, the simplest method for preparing homogeneous films on glass appears to be the technique of electrostatic self-assembly (ESA). In this section, a brief background to ESA is given, then results of a similar study to that discussed above are presented.

5.6.1 Background

It was first shown by Iler in 1966 [126] that coatings consisting of alternating layers of positively and negatively charged species could be produced by exposing a charged substrate alternately to each species. The technique was rediscovered around 1991 by Decher [127] and has been greatly extended since then by a number of groups [128], with coatings having been produced from an enormous variety of polyelectrolytes and colloids on a wide range of substrates. The technique is referred to as the "layer-by-layer" method of electrostatic self-assembly.

The basic process is illustrated in Figure 5.17. The (here) negatively charged substrate is immersed in the polycation solution. Polymer chains bind, through electrostatic interactions with the surface, in sufficient number to reverse the charge of the surface. This occurs fairly quickly, on the order of seconds to minutes [129]. The substrate is withdrawn and rinsed thoroughly to dislodge any material

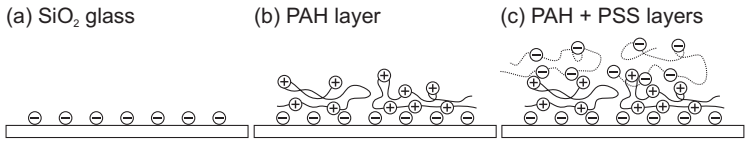

Figure 5.17: Schematic depiction of the ESA process. (a) The glass substrate is negatively charged in a basic solution. (b) A layer of polycations adsorbs and the surface charge is reversed. (c) A layer of polyanions adsorbs, reversing the surface charge again.

that is not firmly bound to the substrate, and is then immersed in the polyanion solution. Polyanion chains attach to the surface and the surface charge is reversed again. The procedure is repeated to build up as many layers as desired.

If solutions of a polycation and a polyanion are mixed, a precipitate forms immediately even if they are relatively weak electrolytes, since there are many charge centres on each molecule. The formation of multilayer films by ESA can be thought of as the controlled, templated formation of such a precipitate.

In the present work, the only real requirements are that the film be homogeneous and that it have absorption bands distributed throughout the mid-infrared; since these criteria are satisfied by a wide range of systems, simplicity and cost are the deciding factors. Fortunately, glass is an ideal substrate. It is extremely smooth and can easily be induced to have a negative charge. The surface of glass consists of silanol (Si–O–H) groups, which are weakly acidic, so if glass is immersed in a basic solution, some of the silanol groups are deprotonated, leaving a negatively charged surface [130].

A very well studied ESA system has poly(allylamine hydrochloride) (PAH) as the cation and poly(sodium 4-styrenesulfonate) (PSS) as the anion (Figure 5.18) [129]. After the first few layers have been deposited, films grown from these polyelectrolytes increase linearly in thickness with each bilayer up to quite thick films [131]. The mass deposited in each PAH layer can be controlled by adjusting the pH of the PAH solution. PAH is a weak base, so the pH determines the proportion of charged groups on each polymer chain: a higher pH means that there are fewer charged groups, which means that more polymer chains can attach to the surface, giving a thicker film. The same trick does not work so well for PSS, because it is a much stronger electrolyte. Instead, the thickness can be increased by increasing the ionic strength of the PSS solution [132]. A solution with greater ionic strength screens charge more effectively, reducing the repulsion between PSS chains and allowing more chains to attach to the surface. If the layers are made too thick, there is a risk of heterogeneity [128].

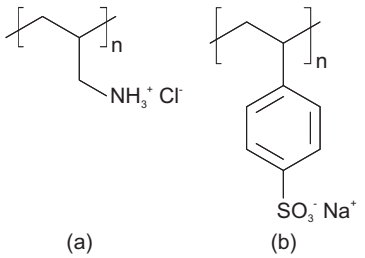

Figure 5.18: A polyelectrolyte pair suitable for forming ESA films. (a) poly(allylamine hydrochloride), PAH; (b) poly(sodium 4-styrenesulfonate), PSS.

5.6.2 Preparation of the standard

First, the glass substrate ($60 \times 150\,\mathrm{mm}^2$, 2 mm-thick window glass) was cleaned. Piranha solution (a 1:3 mixture of hydrogen peroxide and concentrated sulfuric acid) is often recommended for this purpose [129], but seemed excessive for the present application. The substrate was cleaned with hot water and detergent, then with acetone and ethanol. It was then immersed in ~50 % sulfuric acid solution for half an hour, followed by extensive rinsing with milli-Q water. The acid displaces any adsorbed cations, regenerating the silanol groups [130]. It was noted that after being treated with acid, the surface was much more hydrophilic than it had been prior to the treatment, with water tending to spread out rather than form droplets.

The solutions were prepared in tall, narrow beakers to facilitate the dipping procedure. The polycation solution was made by dissolving ~500 mg of PAH (MW 70,000; Aldrich) in ~550 mL of milli-Q water and adjusting the pH to 8.0 with $1.0\,\mathrm{mol\,L^{-1}}$ sodium hydroxide. The polyanion solution was made by dissolving ~2 g of PSS (MW 70,000; Aldrich) and ~30 g of NaCl (analytical grade; Aldrich) in ~550 mL of water. The pH of the PSS solution was adjusted to 2.0 with $1.0\,\mathrm{mol\,L^{-1}}$ hydrochloric acid.

The cleaned substrate was immersed in the PAH solution for 2 minutes, then rinsed thoroughly with water and dipped in the PSS solution for the same length of time. This was repeated ten times to give a total of 20 polyelectrolyte layers. After deposition was complete, the coupon was rinsed thoroughly and blown dry with nitrogen. The film was not visible to the naked eye. No efforts were made to determine the thickness, but Kolarik et al. [132] used similar conditions and obtained 10-bilayer films with loadings of approximately $5\,\mu\mathrm{g\,cm^{-2}}$. If the loading here is similar and assuming the film has density $>1\,\mathrm{g\,cm^{-3}}$, the thickness of the film is $\lesssim 50$ nm, much less than the limit of linearity, $d_{99} \gtrsim 120$ nm (Table 5.1).

5.6.3 Representative spectra

Spectra were measured at incidence angles of 35–80° at 5° intervals. Three spectra were measured at each angle. The PAH/PSS bilayers have reasonably strong, broad absorption bands between about 3700 and 2500 cm^{-1}, due to N–H and C–H stretching modes and the symmetric stretching mode of water present in the film (see the top plot in Figure 5.19). Since the optical constants of glass vary only slowly in this region, these overlapping bands have been collected into a single integral. There are many bands in the range 1700–900 cm^{-1}; these have been integrated separately, as shown in the bottom plot of Figure 5.19.

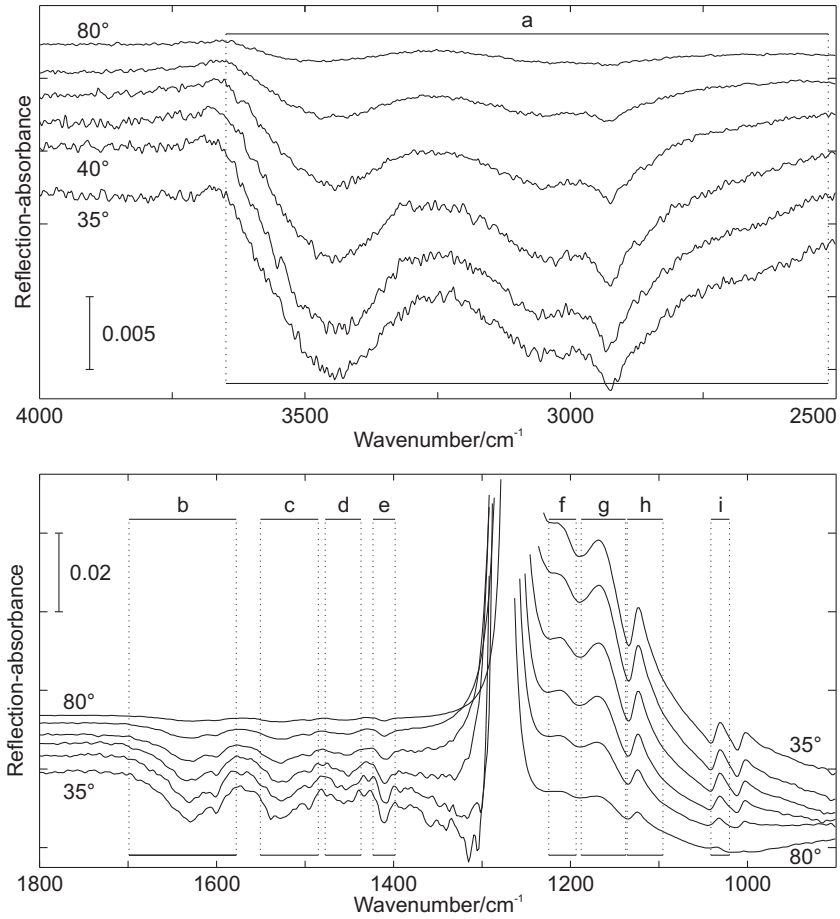

Figure 5.19: Unpolarised IRRA spectra of a 10-bilayer film of PAH/PSS on glass (top: high-wavenumber region; bottom: low-wavenumber region). The incidence angles are, from top to bottom, 80, 70, 60, 50, 40 and 35° (note that the order is reversed to the red of the discontinuity in the low-wavenumber region). Spectra have been offset for clarity; the ordinate scales are indicated. The dashed vertical lines denote the regions of integration.

5.6.4 Signal and noise calculations

The normalised SNR curves for bands a–i are plotted in Figure 5.20, where they are compared with the results of calculations. In these calculations, the optical constants of the substrate have been varied (by manual trial and error) to optimise the fit to the experimental data. The refractive index of the film, n_1, is not known, but has been determined for similar PAH/PSS films to be 1.539 at 623.8 nm [129]. In the infrared, n_1 will be slightly lower (due to dispersion), and its value at each absorption band will depend on the strength and proximity of other nearby bands. A constant value of $n_1 = 1.5$ was used for the calculations. Because of the normalisation, the calculated SNR curves are almost independent of the film thickness, d, and absorption index, k_1, provided that $d \lesssim 200$ nm and $k_1 \lesssim 0.5$. Both of these conditions are reasonable, and the values used were $k_1 = 0.1$ and $d = 50$ nm.

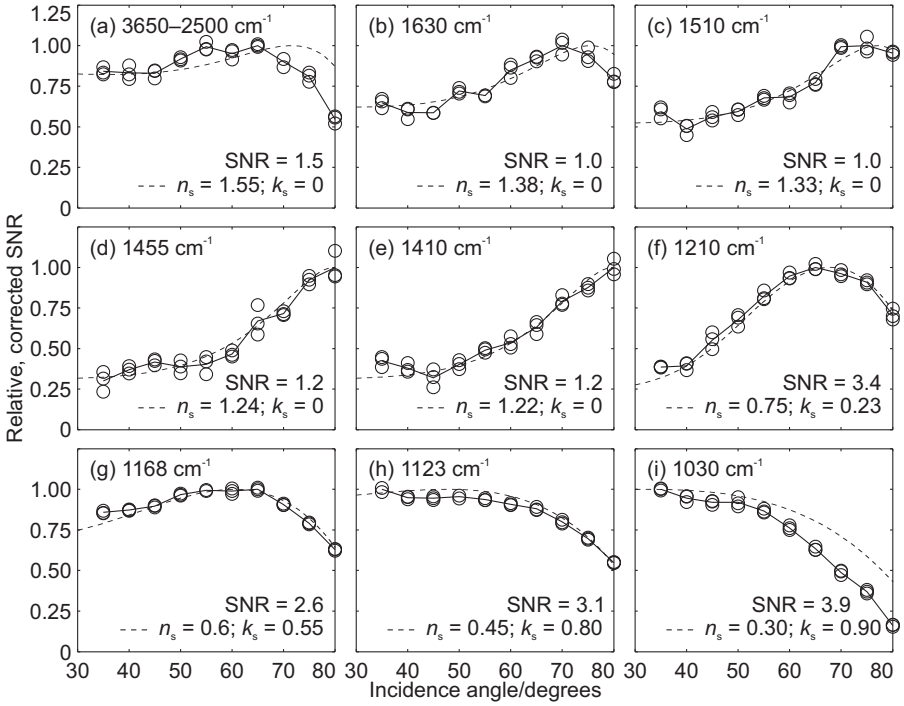

Figure 5.20: Relative, corrected SNR for a 10-bilayer PAH/PSS film on glass. The band labels and peak wavenumbers are indicated. Circles represent individual measurements; the solid lines join the mean values for each angle. The dashed lines are from optical calculations; the substrate optical constants n_s and k_s were determined for each band by starting with approximate values and adjusting them to optimise the fit to the experimental data. The film optical constants were $n_1 = 1.5$ and $k_1 = 0.1$; the thickness was $d = 50$ nm. The dashed lines have been normalised to unit maximum; the parameter "SNR" indicates the relative SNR for comparison between panels.

The estimated values for the substrate optical constants (see Figure 5.21) are consistent with what is known about the refractive index of window glass. In the visible, soda lime glasses have slightly higher refractive indices than pure silica glass (e.g. ~1.52 for crown glass and ~1.46 for silica glass [123]); this should also be true for the high-wavenumber region of the infrared. From the minimum in the

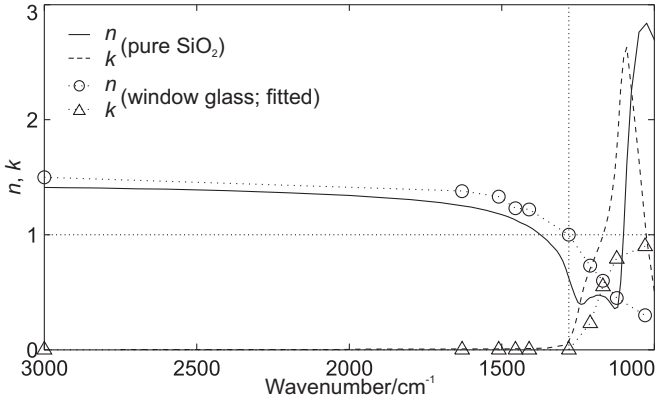

Figure 5.21: Optical constants for pure SiO_2 glass (from Ref. 40; n: solid line, k: dashed line) and values obtained for soda-lime float glass by optimising the fit to the SNR curves in Figure 5.20 (dotted lines; n: circles, k: triangles). The unity-crossing of n estimated by the reflectance minimum in Figure 5.14 is also shown.

reflectance spectrum (Figure 5.14), the unity-crossing of n_s occurs at $\sim$1280 cm^{-1}: to the blue of this value, the estimated values of n_s are greater than one, while to the red they are less than one and decreasing with wavenumber. The k_s values are also reasonable: zero to the blue of the reflectance minimum and increasing with decreasing wavenumber. However, it is not certain whether this is a reliable method for determining the optical constants of the substrate.

In general, the agreement between the calculated and experimental values is qualitatively quite good except in panels a and i, for the highest- and lowest-wavenumber bands, respectively. These results reinforce the conclusion reached in Section 5.5: as the frequency is decreased from the high-wavenumber edge of the mid-IR towards the Si–O band, the substrate becomes less reflective and incidence angles closer to grazing are preferred. However, once n_s dips below unity, k_s starts to increase, as does the reflectivity of the substrate, and the best SNR is obtained at near-normal incidence. Bands f and g represent intermediate cases where incidence angles around 60° are preferred.

Clearly, the optimal incidence angle depends strongly on the region of the spectrum in which the absorbance bands of the analyte are found. For analytes having bands throughout the mid-IR, it may be simplest to focus attention on the bands to the blue of $\sim$1400 cm^{-1}. While, in general, the SNR for bands having equal absorbance (in the transmission-mode sense) is much greater for the bands at frequencies $\lesssim$1200 cm^{-1}, coinciding with the Si–O absorbance of the substrate (see the parameter "SNR" in the panels of Figure 5.20), these measurements can be subject to complicated and intense baseline features that are quite variable. If bands to the blue of 1400 cm^{-1} are regarded as more important, an incidence angle of 75 ° can be recommended as providing near-optimal SNR for all frequencies $\gtrsim$1400 cm^{-1}.

5.7 Conclusions

These experimental studies support the conclusions drawn from the calculations presented in Section 5.2 and suggest that calculations pertaining to model systems are useful for determining the optimal conditions for IRRAS. For both acetaminophen and an electrostatically self-assembled polyelectrolyte multilayer on glass, qualitative agreement between calculated and experimental results was demonstrated.

If a single probe with fixed incidence angle is to be used for both metallic and glass substrates, the incidence angle should be 70–80°. This entire range is adequate for glass (for frequencies to the blue of the point where the refractive index crosses unity), but, for metallic substrates, the two extremes of the range have quite different properties: at the low end, sensitivity is somewhat lower but linearity with respect to film thickness is better; at the high end, the sensitivity is improved at the expense of the linearity. A polariser does not appear to be necessary, provided that the films are thin, but may provide an SNR improvement (of up to a factor of two) for metallic substrates if the detector is sufficiently sensitive.

Several avenues remain unexplored. Measurements with *p*-polarised light incident at the Brewster angle for the film (~55–60°) should be investigated for cleaning validation of metal surfaces. It seems likely this would reduce the nonlinearity encountered in some of our work (see Chapter 8); but it remains to be seen whether the sensitivity would be acceptable.

The importance of the SNR should not be overemphasised, however. There is evidence that spectroscopic noise is not the dominant factor in the uncertainty associated with the chemometric models used in this work (see Section 6.6). If the incidence angle of the probe were decreased slightly, to 70°, the improved linearity with respect to film thickness on metallic substrates would probably be more beneficial than the loss of sensitivity would be detrimental.

Chapter 6

Acetaminophen residues on glass

6.1 Introduction

Fibre-optic grazing-angle IRRAS for cleaning validation has been investigated previously by workers at the University of Puerto Rico in collaboration with Remspec [133]. These authors used the method to quantify loadings of an unspecified API on aluminium substrates, finding an RMS prediction error of 0.08 $\mu g\,cm^{-2}$ over the loading range 4–16 $\mu g\,cm^{-2}$. They contrasted this result with much poorer results for an HPLC method (RMSEP 2.0 $\mu g\,cm^{-2}$), but no details of the analyte or the HPLC method were given. A limitation of their study was that the gravimetric method they used to determine the loadings could not be used quantitatively for loadings below 4 $\mu g\,cm^{-2}$, at which loading most compounds are visible by eye anyway [15]. Very recently, a study by workers at Novartis [134] extended these early results to stainless steel substrates and (again, unspecified) APIs with excipients present. They found RMS errors of prediction of 0.04–0.07 $\mu g\,cm^{-2}$ for loadings in the range 0–1.5 $\mu g\,cm^{-2}$, a range much more relevant for cleaning validation.

Another important surface for cleaning validation is silica glass. As discussed in Chapters 2 and 5, the optical properties of glass differ greatly from those of metals; consequently, IRRAS with a glass substrate exhibits a number of interesting features that are not seen with metallic substrates. These features are described in the above-mentioned chapters, and some glass-IRRAS and transmission spectra are compared in Figure 7.1 in the next chapter. Briefly, there are two challenges involved in using glass as an IRRAS substrate. The first is sensitivity: glass IRRAS is generally much less intense than metallic IRRAS, and the reflectivity is lower, so the noise is greater. The second challenge relates to the effect of the strong Si–O absorption bands between ~1300–900 cm^{-1}. As shown in Chapters 2, 5 and 7, these bands cause very strong baseline features that can dominate the IRRAS. The results presented here illustrate that both of these challenges are quite easily overcome.

This chapter describes the application of IRRAS to the determination of loadings of a single API, acetaminophen, on a glass substrate. A subset of the work discussed in this chapter was published in Ref. 135. Acetaminophen was chosen as a model API because it is inexpensive, safe and easily handled; due to its solubility, however, it is unlikely to pose any particular cleaning validation difficulty.

6.2 Experimental

Acetaminophen (4-acetamidophenol, 98 %) was obtained from Sigma-Aldrich and used without further purification. The solvents used were milli-Q water, ethanol (solvent grade) and acetone (solvent grade). Two glass coupons ($15 \times 10\,\mathrm{cm}^2$) were cut from 3 mm-thick window (soda-lime float) glass and were roughened on one side by sand-blasting to prevent reflection from the back face.

Fifty-three standards were prepared by the spray method (Section 4.3.2) with acetone as the solvent. For the first 17 standards, three spectra were measured with 100 interferometer scans each; for the rest, five were measured with 50 scans each.

After IRRAS measurement, the coupon was rinsed with 25–100 mL of ethanol and the loading was determined by UV colorimetry using a univariate method based on the absorbance at 249 nm. In later work this method was superseded by the superior CLS method described in Section 4.3.3.

6.3 Model optimisation by cross-validation

The data were divided into calibration and test sets by the following method.

1. List the standards in order of loading.
2. Assign the first and last standards to the calibration set.
3. Starting with the second standard, assign every third standard to the test set.
4. Assign the remaining standards to the calibration set.

This procedure resulted in calibration and test sets with 36 and 17 standards, respectively. In each set, the distribution of loadings was approximately equal.

On the basis of inspection of the spectra, the wavenumber range 1880–1340 cm^{-1} was chosen (see Figure 7.1 and the associated discussion in Chapter 7). This range includes most of the bands due to acetaminophen in the fingerprint region, but excludes the prominent feature due to the Si–O stretch. The effect of including this feature will be discussed later.

A few typical spectra are plotted in the top-left panel (labelled NP) of Figure 6.1. While the bands due to acetaminophen are clearly visible, there are also two obvious interfering features. The most prevalent of these is a baseline offset; there are also strong absorption bands due to water vapour in some of the spectra.

Essentially, there are two approaches to dealing with this variation in the spectra that is uncorrelated with the analyte loading. The first is simply to allow it to be modelled implicitly by increasing the rank of the PLS model. The second is to remove as much of it as possible prior to the regression analysis: the constant offset can be removed either by fitting a constant to the flat region at the high-wavenumber end of the truncated spectrum or by taking the first derivative; the water vapour bands can be minimised by subtraction of a reference water vapour absorbance spectrum (see Appendix B). It is not obvious whether any of these pretreatments are beneficial, so several will be used in parallel, and the results compared to those obtained with no pretreatment.

Combinations of the various options lead to six different pretreatments:

- No pretreatment (NP)
- Offset elimination (OE)
- Water vapour subtraction (WV)
- First derivative (FD; 15-point quadratic Savitzky-Golay [85, 86, 136] filter)
- Water vapour subtraction and offset elimination (WV + OE)
- Water vapour subtraction and first derivative (WV + FD)

The effects of these pretreatments are illustrated in Figure 6.1, where some representative spectra are plotted before and after each treatment. Clearly, the pretreatments are effective at removing the interferences, but whether or not this translates into improvement of the model is yet to be seen.

In Section 3.5 it was shown that "leave-one-spectrum-out" cross validations can underestimate the prediction error when the several spectra for a sample are not true replicate measurements, but are correlated in some way. This phenomenon is demonstrated here in Figure 6.2, in which the results of four cross validations are plotted. The two lines near the bottom of the plot are for regular cross validations, and it can be seen that the per-spectrum cross validation gives a slightly smaller estimate of the error, and continues to decrease as the rank is increased. More revealing are the two lines near the top of the plot, which depict results for cross validations prior to which the vector, $\mathbf{y}$, of loadings was permuted randomly (on a per-sample basis). This process makes it impossible to produce

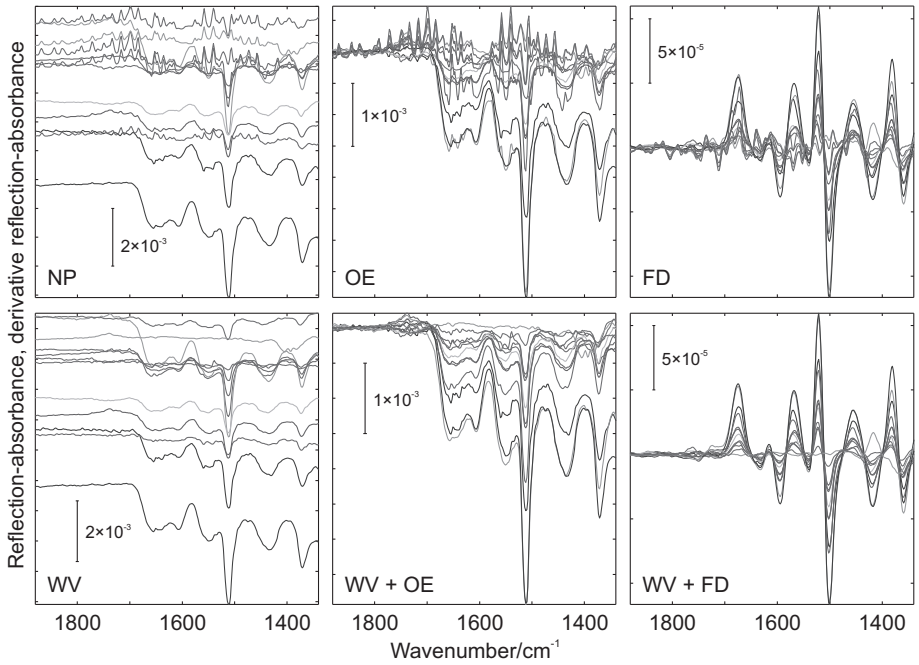

Figure 6.1: Effect of the various pre-treatments (see the text for descriptions) on a small subset of the (averaged per-sample) spectra over the wavenumber range used in the PLS modelling. In each panel the ordinate scale is indicated. NP = no pre-treatment; OE = offset elimination; WV = water vapour subtraction; FD = first derivative.

a genuine predictive model, because the relationship between the loadings and the spectra is destroyed. The dramatic decrease in the RMSECV with increasing rank for the per-spectrum cross validation is due to correlations between the amounts of the various interferences present in the spectra from each standard; essentially, the model is recognising similarities between the spectra from each sample rather than systematic variations between samples. The effect is weaker here than in the synthetic example depicted in Figure 3.8, but is sufficient reason to use only per-sample cross-validation.

For each pretreatment, a leave-one-sample-out cross validation (see Sections 3.3.1 and 3.5) was conducted. The results are summarised in Figure 6.3 and Table 6.1. For the unprocessed spectra, the optimal rank (determined by Martens' method [82] with $s = 0.02$; see Section 3.3.2) is 5. Independent of whether it is combined with another pre-treatment, subtracting out the water vapour spectrum does not reduce the optimal rank but does improve the RMSECV slightly. Using derivative spectra has the biggest impact on the rank; the optimal model has a single factor compared with a rank of 5 for the untreated data. In comparison, removing the baseline offset by subtraction reduces the optimum rank by only one compared to the untreated spectra. None of the pretreatments has a dramatic effect on the RMSECV at optimal rank. It is interesting that a single-factor model is successful for the derivative spectra even without reduction of the water vapour bands. Possible explanations for this are that the overlap between the derivative water vapour spectrum and the derivative acetaminophen spectra is not

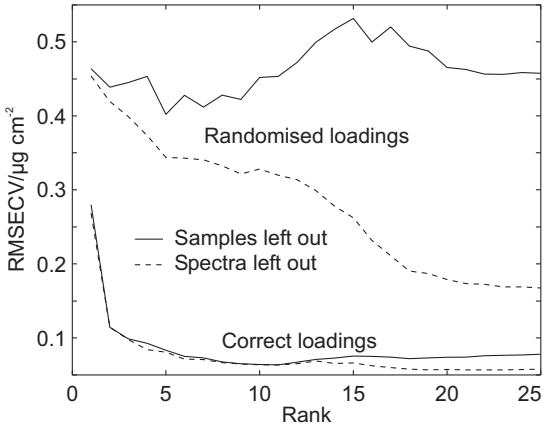

Figure 6.2: Comparison of per-sample and per-spectrum cross-validations. The bottom two lines use the correct loadings; the top two are for cross validations in which the vector of loadings, **y**, has been randomised. Solid lines: leave-one-sample-out cross validations; dashed lines: leave-one-spectrum out cross-validations.

extreme, or that the magnitude of the water vapour bands is much less than the magnitude of the acetaminophen bands.

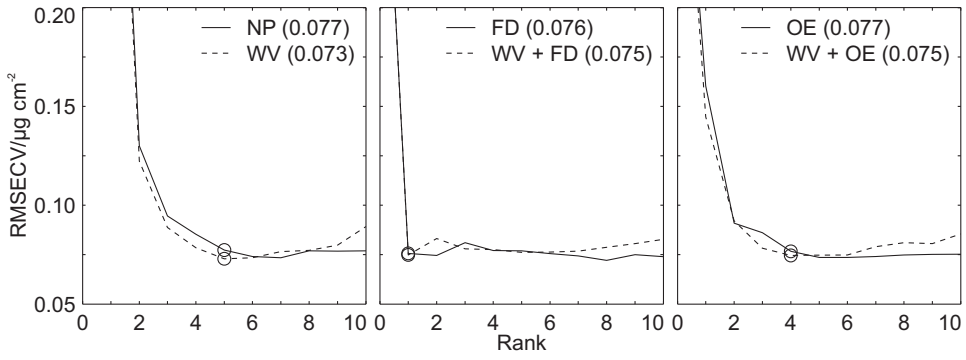

Figure 6.3: Cross-validation results. The numbers in parentheses are RMSECV values in µg cm^{-2}. The circles indicate the optimum rank determined according to Martens' method [82] with $s = 0.02$. NP = no pre-treatment; OE = offset elimination; WV = water vapour subtraction; FD = first derivative.

The next step is to check for outliers in the calibration set. Three plots are useful for this. The first (Figure 6.4) is a plot of the predicted loadings against the true loadings. The second (Figure 6.5) has the (calibration) leverage plotted against the studentised cross-validation residual. The average leverage is A/n, where A is the rank of the model; points falling above the horizontal line are especially influential, having leverage greater than $4A/n$. Points falling outside the vertical lines correspond to unusually large residuals (absolute studentised residual >3). Suspect points are labelled with their sample indices. The third plot (Figure 6.6) is of mean square spectroscopic residuals, calculated as the mean square of the difference between the fitted spectrum and the measured spectrum. For each model, the standard deviation of the RMS residuals was calculated: spectra with RMS residuals more than four

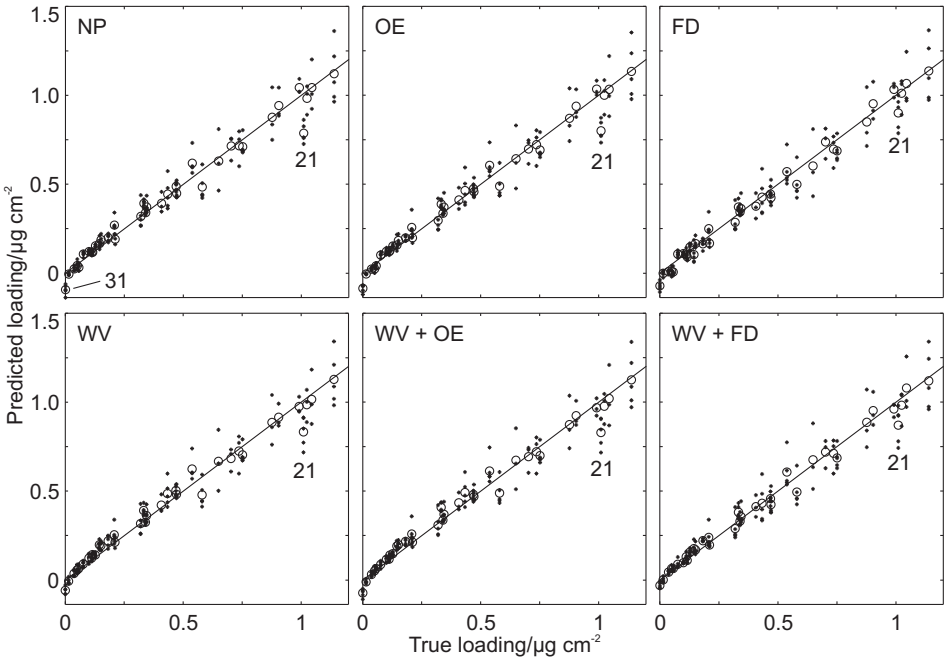

Figure 6.4: Predicted (in cross-validation) loadings versus true loadings for the six models. Asterisks: individual spectra; circles: per-sample averages. NP = no pre-treatment; OE = offset elimination; WV = water vapour subtraction; FD = first derivative.

times the standard deviation are labelled with their sample indices. This is an arbitrary limit: formal hypothesis testing in this case is difficult because of the lack of a reliable degrees-of-freedom estimator (see Section 3.3.4 for elaboration of this point).

Outliers can be removed on either a per-spectrum basis or a per-sample one. Which is appropriate depends on the cause of the outlier nature, which may itself be a per-spectrum or a per-sample phenomenon. For simplicity, the per-sample basis will be used here.

The worst standard in terms of its cross-validation loading residual is sample 21, which has a loading of 1.0 $\mu g\,cm^{-2}$. Its spectra also have high leverage and high spectroscopic residuals; this sample is clearly an outlier. Inspection of the spectra (not shown) reveals that the derivative-shaped substrate feature is unusually strong, causing a significantly curved baseline over the region used for the modelling.

Sample 31, a blank, has large spectroscopic residuals (Figure 6.6) and, from Figure 6.4, has larger loading residuals than other low-loading standards. Because the residuals increase with increasing loading, though, when its residuals are studentised with respect to all the other residuals, they are not unusually large (in the NP model, for example, the studentised residuals for the spectra from sample 31 range from 0.8–1.8 and the leverages from 0.01–0.05). However, this sample is also an outlier that should be removed.

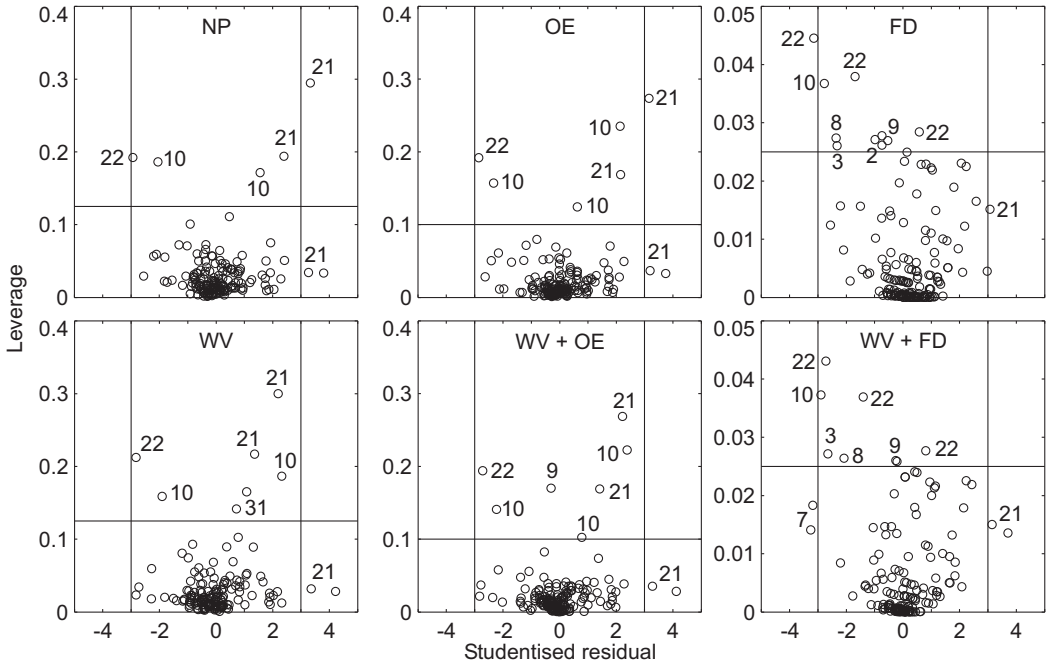

Figure 6.5: Leverage plotted against studentised residual. Points falling outside the vertical lines have absolute studentised residuals greater than three; points above the horizontal line have leverage greater than four times the calibration-set average. Each point represents a single spectrum; the labels are sample indices. NP = no pre-treatment; OE = offset elimination; WV = water vapour subtraction; FD = first derivative.

Other standards that may be considered outliers are samples 10 (1.0 µg cm^{-2}) and 22 (1.1 µg cm^{-2}). Sample 22 has high leverage (Figure 6.5) and, in the first-derivative models, large spectroscopic residuals (Figure 6.6). The loading residuals for its individual spectra are quite large, but the mean predicted loading is very close to the measured loading, so this standard is retained. The spectroscopic residuals for sample 10 are small in the first-derivative models, but large elsewhere. It is not obvious from inspection of the spectra why this is the case. Comparing the fitted and measured spectra (Figure 6.7) reveals some small differences in relative band heights, but it is uncertain what would cause this. Again, the mean predicted loading differs only slightly from the measured loading, so this standard is retained.

When the cross-validation is repeated with the two outlying samples (21 and 31) removed, the results change somewhat (compare the first and second columns of Table 6.1). For almost all models, the RMSECV decreases very slightly (by ~0.01 µg cm^{-2}) and R^2 increases slightly. These changes are due mostly to the samples with large residuals being removed from the calculation of the statistics; the actual improvement to the model is smaller again.

For most models, the optimal ranks change when the outliers are removed. In particular, the water vapour subtraction pretreatment now reduces the optimal rank by one in each case. Outliers can have unpredictable effects on the behaviour of RMSECV with respect to rank, which often cause an under-

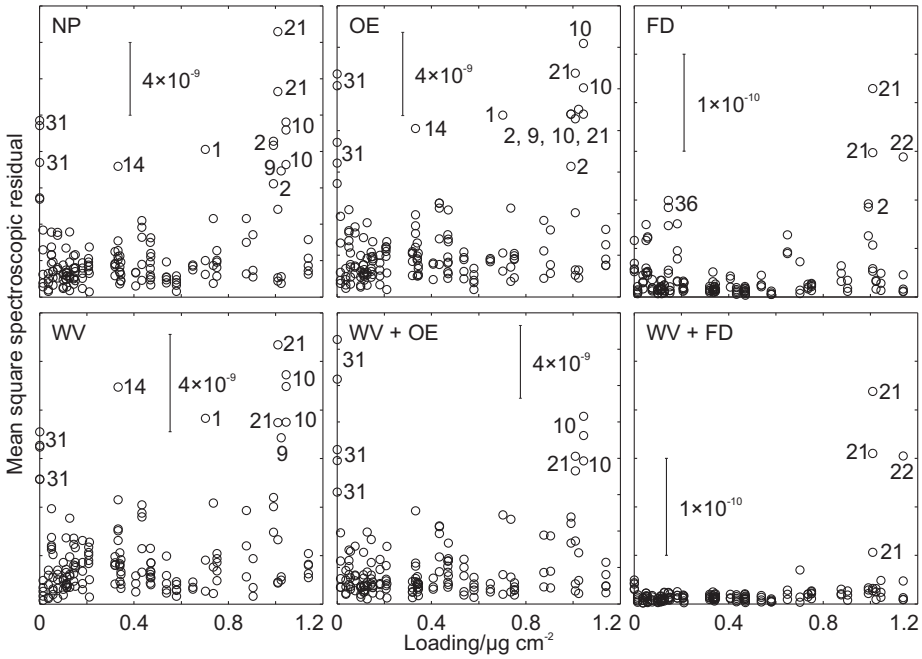

Figure 6.6: Mean-square cross-validation spectroscopic residuals plotted against loading. Points with RMS residuals greater than four standard deviations are labelled with their sample indices. NP = no pre-treatment; OE = offset elimination; WV = water vapour subtraction; FD = first derivative.

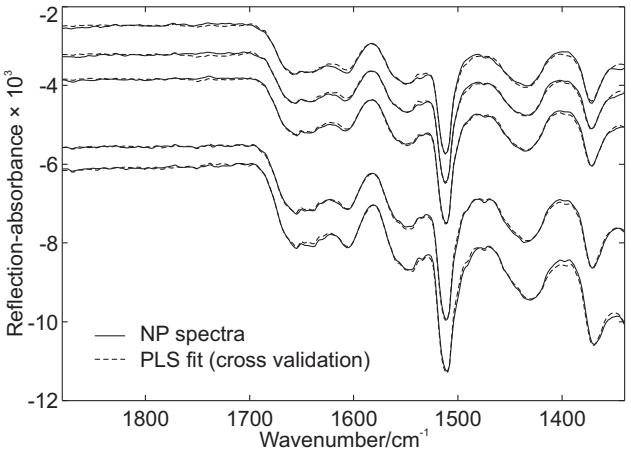

Figure 6.7: The five spectra (with no pre-treatment) for sample 10 (solid lines), and spectra fitted by the PLS model in cross-validation (dashed lines).

fitted model to be chosen. This effect is seen in the first-derivative models, where removing the outliers increases A_{opt} from one to two or three, depending on whether water vapour correction is also used. Plots analogous to those in Figures 6.4–6.6 do not reveal any new outliers.

Table 6.1: Cross-validation results for the six models, before and after outlier removal. A_{opt} is the optimal rank, R^2 is the determination coefficient, and RMSECV is in $\mu g\,cm^{-2}$. NP = no pre-treatment; OE = offset elimination; WV = water vapour subtraction; FD = first derivative. The data for "All standards" and "Outliers removed" were obtained using the wavenumber range 1880–1340 cm^{-1}; the "Larger wavenumber range" data were obtained by extending the lower limit to 1000 cm^{-1} to include the strong Si–O feature of the substrate (see Figure 7.1 in the next chapter).

	All standards			**Outliers removed**			**Larger wavenumber range**		
Treatment	A_{opt}	RMSECV	R^2	A_{opt}	RMSECV	R^2	A_{opt}	RMSECV	R^2
NP	5	0.077	0.95	5	0.064	0.96	5	0.074	0.95
OE	4	0.077	0.95	4	0.065	0.96	4	0.076	0.94
FD	1	0.076	0.95	3	0.065	0.96	6	0.063	0.96
WV	5	0.073	0.95	4	0.066	0.96	5	0.074	0.95
WV + OE	4	0.075	0.95	3	0.067	0.96	4	0.077	0.94
WV + FD	1	0.075	0.95	2	0.066	0.96	5	0.064	0.96

The data in Table 6.1 constitute very little basis for choosing one of these models over the others, either before or after the removal of the outliers. The RMSECV values can be compared by F-tests, as described in Appendix A. The critical value of the F distribution with n (here, $n = 150$, the number of calibration spectra) and n degrees of freedom and at the 95 % confidence level is 1.31. The largest ratio of two MSECV values (squares of the RMSECV values in the sixth column of Table 6.1) is $(0.067/0.064)^2 = 1.1$, so none of the differences between the RMSECV values is significant at the $\alpha = 0.05$ level.

In this thesis, the preferred approach has been to use minimal pre-processing: if the modelling method can account for the uncorrelated variation by itself, why add extra steps to the process?

One potential advantage of aggressive pre-processing relates to detection limits and the leverage of low-loading samples. As discussed in Section 3.4, the uncertainty in the predicted loading of a new sample increases with its leverage. The leverage has a quadratic dependence on the loading, but it also depends on the other factors contributing to the spectrum, such as water vapour or a baseline offset. Pre-treatment of the spectrum should reduce its leverage, reducing the uncertainty in the predicted loading. Furthermore, models with more aggressive pre-treatment should require fewer factors, further reducing the leverage. Figure 6.8 contains plots of the leverage as a function of loading for two models: the one with five factors and no pretreatment, and the other with two factors and both water vapour subtraction and first-derivative pre-treatments (WV + FD). In the latter model, the leverages are much

smaller than in the former (note the difference in the ordinate scales of the two plots). Additionally, the quadratic relationship between loading and leverage (compare these plots with Figure 3.4 in Chapter 3) is much more evident for the WV + FD model, because the interferents that also contribute to the leverage (chiefly a baseline offset and water vapour absorbance) have been removed. In the present work, however, this advantage of lower leverages seems to be academic, because the leverages are small enough to be negligible in either case. No significant improvement in the RMSECV with pre-treatment of the spectra is seen, and it is likely that the reference-method errors arising from sample heterogeneity preclude the detection of the more subtle effect of the difference in leverage.

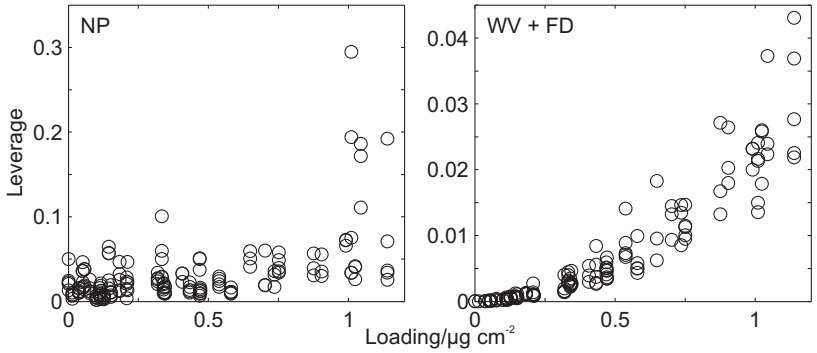

Figure 6.8: Leverage plotted against measured loading for spectra with no pre-processing (NP) and for spectra with water vapour subtraction and first-derivative pre-processing (WV + FD).

The modelling process was repeated with the wavelength range enlarged to 1880–1000 cm^{-1}, to include the prominent feature due to the substrate. The same standards were identified as outliers. The RMS errors (Tables 6.1 and 6.2) are similar to those obtained with the smaller wavelength range: in general, slightly lower in the cross-validation and slightly higher in the test-set validation.

6.4 Test-set validation

Strictly, the test set should be used only once, after the model parameters have been determined. However, since this is an exercise in model optimisation and validation methodology, it is interesting to use the test set with all the models, including those obtained before outlier removal. Pretreatments were applied in exactly the same manner as for the calibration set.

The results are summarised in Table 6.2; predicted loadings are plotted in Figure 6.9. The RMSEP values are similar for all the pretreatments, and slightly higher than the corresponding RMSECV values. It is seen that removing the two outliers from the calibration set improves the test-set predictions, as evidenced by the slight general reduction in the RMSEP.

It is evident from Figure 6.9 that some loadings are predicted very poorly: in particular, those

Table 6.2: Test-set validation results for the six models, before and after outlier removal. A_{opt} is the optimal rank determined by cross-validation, R^2 is the determination coefficient, and RMSECV is in µg cm^{-2}. NP = no pre-treatment; OE = offset elimination; WV = water vapour subtraction; FD = first derivative. The data for "All standards" and "Outliers removed" were obtained using the wavenumber range 1880–1340 cm^{-1}; the "Larger wavenumber range" data were obtained by extending the lower limit to 1000 cm^{-1} to include the strong Si–O feature of the substrate (see Figure 7.1 in the next chapter).

	All standards			**Outliers removed**			**Larger range**		
Treatment	A_{opt}	RMSEP	R^2	A_{opt}	RMSEP	R^2	A_{opt}	RMSEP	R^2
NP	5	0.079	0.93	5	0.076	0.93	5	0.083	0.92
OE	4	0.077	0.93	4	0.076	0.93	4	0.086	0.92
WV	5	0.075	0.94	4	0.076	0.94	5	0.076	0.93
FD	1	0.086	0.92	3	0.077	0.93	6	0.077	0.93
WV + OE	4	0.076	0.93	3	0.074	0.94	4	0.080	0.93
WV + FD	1	0.080	0.93	2	0.073	0.94	5	0.077	0.93

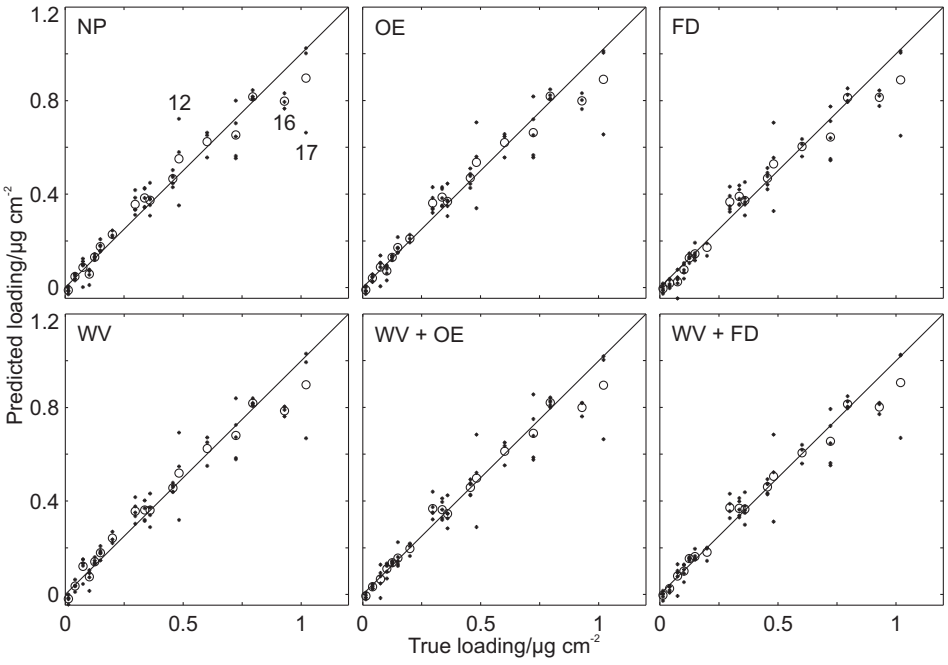

Figure 6.9: Predicted versus true test-set loadings for the six models (after calibration-set outlier removal). Asterisks: individual spectra; circles: per-sample averages. NP = no pre-treatment; OE = offset elimination; WV = water vapour subtraction; FD = first derivative.

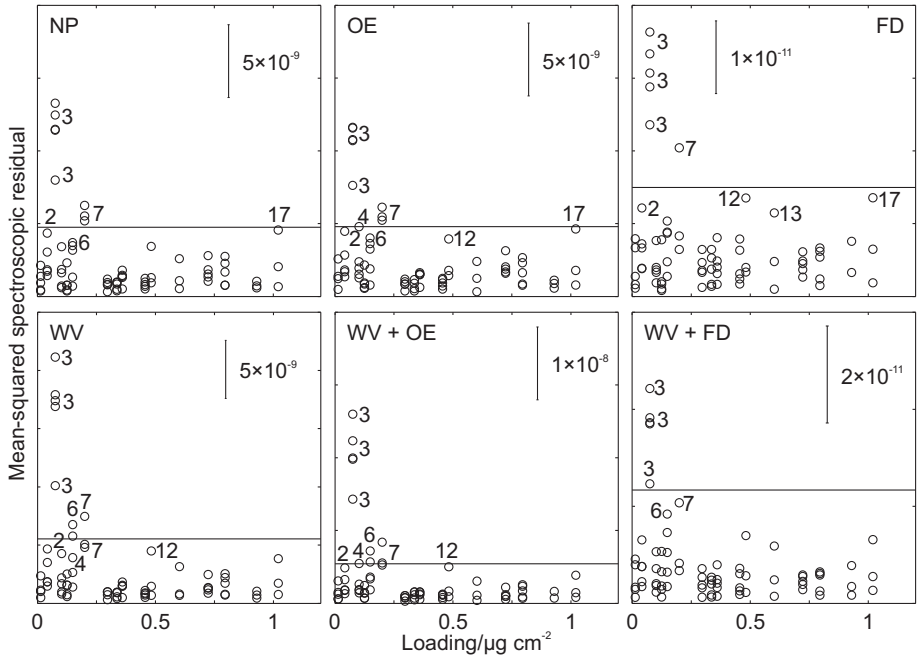

Figure 6.10: Mean-square test-set spectroscopic residuals plotted against loading. Points with RMS residuals greater than four (calibration-set) standard deviations are labelled with their sample indices. NP = no pre-treatment; OE = offset elimination; WV = water vapour subtraction; FD = first derivative.

for samples 12, 16 and 17 (labelled in the top-left panel). An important question arises: if the actual loadings were unknown, could these samples be identified as outliers by their spectroscopic residuals? If so, they could be flagged as such and their predicted loadings treated as unreliable. The spectroscopic residuals are plotted in Figure 6.10. If the same arbitrary limit as for the calibration set is used, the only sample persistently identified as an outlier is sample 3, for which the prediction is actually quite good in most models. From inspection of the spectra (the top spectrum in Figure 6.11), however, it is clear that some additional absorbing species or artefact was present, giving rise (for example) to the features near 1480 and 1860 cm^{-1}.

The third spectrum from sample 17 has a large spectroscopic residual in most models, yet it is the first spectrum from this sample that has the large concentration residual: the loading predicted from the third spectrum is close to the measured loading. The three spectra from this sample are plotted in Figure 6.11. For clarity, the spectra with water-vapour-subtraction pre-treatment are shown. The three spectra look very much alike, except that the first one is much less intense. The large spectroscopic residual for the third spectrum is probably due to some small irregularities, the two most visible of which are at ~1660 and ~1560 cm^{-1} (circled in Figure 6.11). These features may be due to incomplete subtraction of water vapour, perhaps because of a small wavenumber shift (see Section 4.2.6). The most likely explanation for the difference in intensity between the first spectrum and the other two is

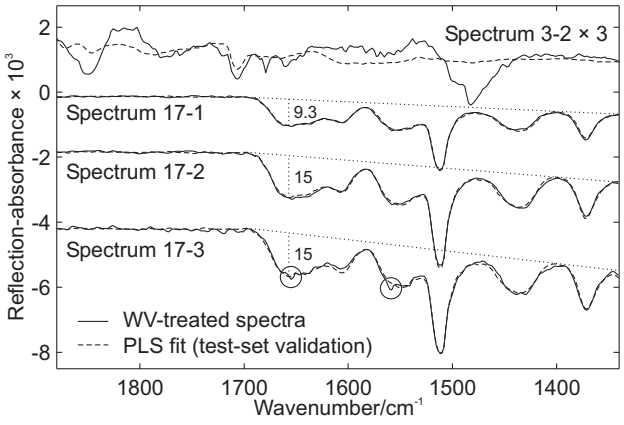

Figure 6.11: Spectra (with water-vapour-subtraction pre-treatment) for test-set samples 3 and 17 (solid lines), and the corresponding spectra fitted by the PLS model (dashed lines). The second spectrum for sample 3 is shown, multiplied by a factor of three; the other two spectra for the same sample are similar. The dotted lines represent baselines and "peak" heights. The negative peak heights relative to the baselines are listed numerically in milli-RA units.

extreme heterogeneity of the standard: the first spectrum is less intense because there is less analyte present on the corresponding region of the coupon. The ratios of the peak heights in Figure 6.11 are almost exactly the same as the ratios of the predicted loadings. Sample 12 (spectra not shown) appears to be similarly heterogeneous. The three spectra for sample 16 (not shown) are all of similar intensity, and are all fitted very well by the model. The errors may still be due to heterogeneity, if the analyte were concentrated near the edges of the coupon.

If these errors are due to the heterogeneity, then the model may be performing perfectly adequately, and the predicted concentrations may be closer to the true, local loadings than is the mean loading (that determined colorimetrically). However, while it seems likely, there is no way to be certain that these errors are generally due to the heterogeneity of the standards. In future work, accurate characterisation of the standards should be a priority. If large coupons are to be used, the homogeneity of the standards should be thoroughly investigated, and optimised by adjustment of the preparation method. If extent of the heterogeneity is known, more accurate estimates of the MSEP can be obtained.

6.5 Confidence intervals for predicted loadings

As discussed in Section 3.4.2, there are many methods for deriving confidence intervals for concentrations predicted by multivariate methods. Here, four approaches will be applied to the test set predictions and compared: the jack-knife [92], the object and residual bootstraps [93], and the errors-in-variables (EIV) formula introduced by Faber and Kowalski [74, 96]. Details of all of these these methods are given in Section 3.4.2.

Assessing the adequacy of the confidence intervals is difficult when the reference data are known to have errors (and when there are few samples). To simplify the analysis and to reduce the effect of the heterogeneity of the standards, the spectra for each standard in both the calibration and the test sets were averaged, giving a single spectrum per standard. This process had only a small effect on the model, and the values for A_{opt} found by cross-validation were unchanged. To reduce the amount of data while preserving the important features, only the NP, WV + OE and WV + FD pretreatments were considered.

For the bootstrap methods, 2000 replicates were used. The variance estimates for the EIV formula were obtained as follows. The spectroscopic residuals were estimated from Figure 4.11 to be on the order of $\sigma_{\Delta X} \approx 5 \times 10^{-5}$. The noise level in the first-derivative spectra differs because they are scaled differently and have been smoothed as well as differentiated; the noise was estimated by applying the same Savitzky-Golay filter to "spectra" consisting solely of noise with standard deviation $\sigma_{\Delta X}$; a standard deviation of $1.6 \times 10^{-6}\,\text{cm}^2$ was found.[1] As it happens, the spectroscopic noise is of little importance in this case; replacing these estimates with zero has little effect on the confidence intervals (not shown).

Deriving an appropriate variance estimate for the errors in the reference-method loading values is more difficult, because the errors are heteroscedastic. Rather than attempting to modify the formula to account for this, an approximation was used. The mean calibration-set loading is $0.4\,\mu\text{g}\,\text{cm}^{-2}$. Assuming the reference-method errors have a relative standard deviation of 7 % (see Section 4.3.5), the variance in the reference-value errors is given very approximately by

$$\sigma_{\Delta y}^2 \approx (0.07 \times 0.4)^2/5 + 0.01^2 = 2.6 \times 10^{-4}\,\mu\text{g}^2\,\text{cm}^{-4} \tag{6.1}$$

where the 5 in the denominator is the number of spectra per standard[2] and $0.01\,\mu\text{g}\,\text{cm}^{-2}$ is the uncertainty in the UV colorimetric determination. The reference-method errors are known to have significant heteroscedasticity, which is neglected in the present approach. A better method would be to use a more sophisticated expression derived from the EIV model that takes the heteroscedasticity of the errors into account. Finally, the residual variance σ_e^2 was estimated as suggested by Faber [96]; see Section 3.4.2.

Since the prediction error variance estimates obtained by all four methods have many (>30) degrees of freedom, the approximation $t_{0.975,\nu} \approx 2$ is acceptable. Accordingly, the 95 % confidence intervals

[1] The ordinate units of the derivative spectra are cm because the abscissa units for the original spectra are cm^{-1}. In the EIV formula, $\sigma_{\Delta X}^2$ is multiplied by the squared Euclidean norm of the regression vector, which always has units such that the product has the same units as $\sigma_{\Delta y}^2$.

[2] The first term in Equation 6.1 accounts for the fact that, because the averaging over the coupon surface is not perfect, the heteroscedasticity of the errors does not vanish when considering averaged spectra. Since a single variance estimate is required and most samples have 5 spectra, this number was used as the denominator.

have been calculated as

$$\hat{y}_u - 2\hat{\sigma}(PE_u) < y_u < \hat{y}_u + 2\hat{\sigma}(PE_u) \tag{6.2}$$

where $\hat{\sigma}(PE_u)$ is the estimated standard deviation of the prediction error.

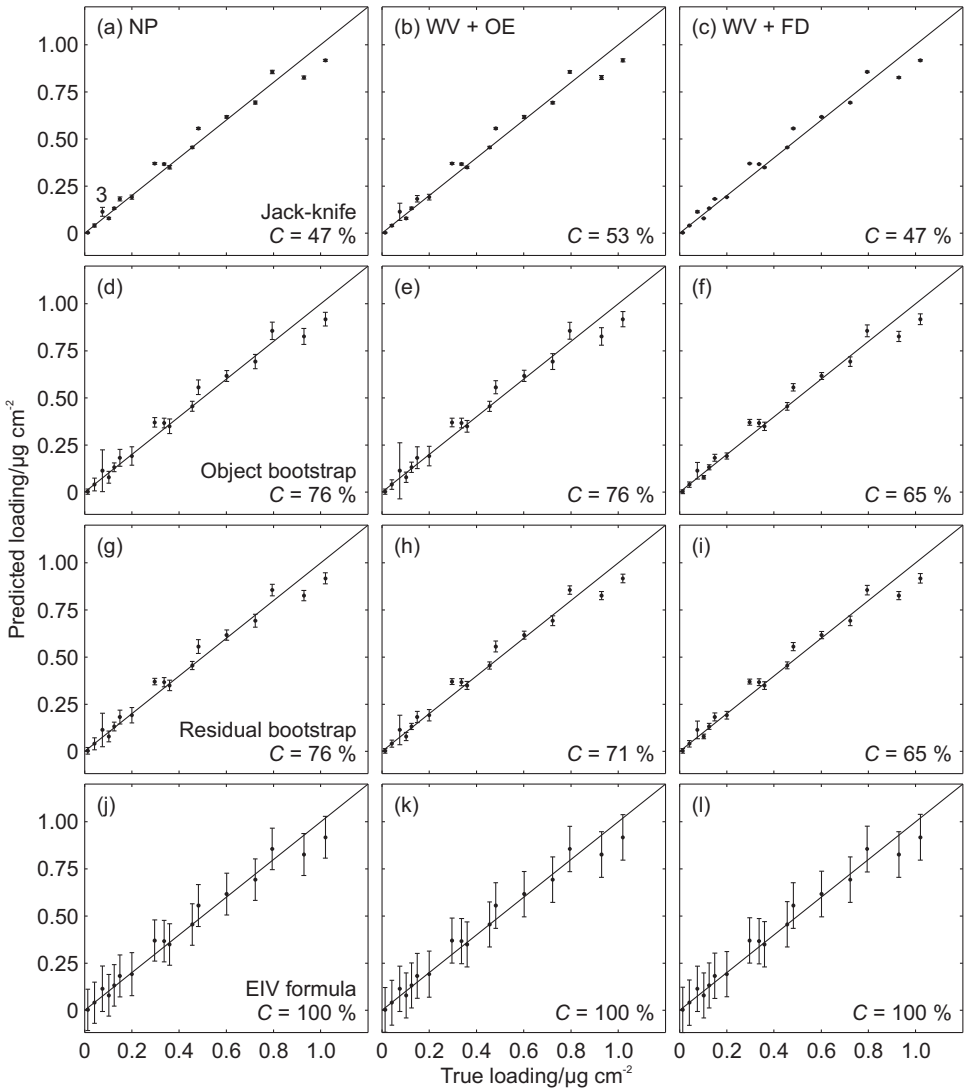

Figure 6.12: Averaged test-set predictions showing approximate confidence intervals. The first plot in each row is labelled with the method used to estimate the confidence intervals; the first in each column is labelled with the pre-treatment used (acronyms are defined in the text). The coverage percentages, *C*, are the percentages of true values that fall inside the intervals.

The predicted and true loadings and their confidence intervals are plotted for the four methods in Figure 6.12. There are major differences: the jack-knife appears to provide very optimistic estimates of the prediction error, leading to confidence intervals that are far too small. Contrastingly, the EIV formula appears to give confidence intervals that are too wide. The two bootstrap methods give similar results, with the object bootstrap estimating slightly greater prediction uncertainties. It is important

to note that the two bootstrap methods function in very different ways: in the object bootstrap, the composition of the calibration set is varied; while in the residual bootstrap, only the errors in the reference y values are varied.

To compare the nominal coverage percentage for the confidence intervals (95 %) with the observed coverage, it is necessary to take into account the error in the reference y values. For each sample, a 95 % confidence interval is constructed for the residual $e = y_u - y$; if this interval contains zero, the predicted and reference values are in agreement. The variance of the residual is $\hat{\sigma}_e^2 = \hat{\sigma}_{\Delta y}^2 + \hat{\sigma}^2(\mathrm{PE_u})$. Here, the heteroscedasticity of the reference-method errors can easily be taken into account, and $\hat{\sigma}_{\Delta y}^2$ is approximated by $(0.07y)^2/5 + 0.01^2$ (i.e. the actual value of y, rather than its mean, is used in the estimation of the error variance). The coverage percentage, C, is the percentage of samples for which the predicted value agrees with the measured value. From the values listed in Figure 6.12 (and the plots themselves), is is apparent that the jack-knife confidence intervals are too small. The bootstrap confidence intervals also appear slightly too small. The confidence intervals derived from the EIV formula give 100 % coverage, which is acceptable, given that there are only 17 samples. However, with such a small test set and high confidence percentage, comparing C to the nominal confidence percentage is only useful for identifying methods that provide confidence intervals that are too small; such a comparison cannot lead to a conclusion that the confidence intervals are too wide. It should also be noted that if the error variance estimate for the reference loadings is inaccurate (a distinct possibility given the lack of strict control over the spraying conditions), the coverage percentages will also be inaccurate.

One observation that highlights the value of resampling-based methods is that sample 3, indicated in Figure 6.12a and already identified as an outlier due to the presence of an unmodelled interferent, is recognised as having an unusually large prediction error by the jack-knife and both bootstrap methods.

In this case, the EIV formula shows little dependence on the sample leverage and is essentially reduced to the MSEC estimated from the calibration residuals. This situation occurs when $\hat{\sigma}_{\Delta X}^2$ and $\hat{\sigma}_{\Delta y}^2$ are small in comparison with $\hat{\sigma}_e^2$ and there are many calibration standards, so that the leverage is small for most test-set samples. More satisfactory results might be obtained with a more sophisticated model that takes into account the heteroscedasticity in the reference calibration-set y values. Such a model should lead to a lower (and more accurate) estimate of $\hat{\sigma}_e^2$ and smaller confidence intervals, but is beyond the scope of this work.

6.6 Importance of photometric noise

It was found in the previous section that the value of the estimate of the noise in the spectra, $\hat{\sigma}_{\Delta X}$, had little effect on the confidence intervals calculated by the EIV formula. The importance of the photometric noise can be evaluated by calibration experiments in which the spectroscopic matrix, $\mathbf{X}$, is corrupted with additional noise prior to the cross validation [96].

First, a more accurate estimate of the noise in these spectra was obtained, by fitting a straight line through a region of the spectrum containing no spectroscopic features and calculating the RMS of the residuals. The wavenumber range used was 1880–1730 cm^{-1}, a region containing no absorbance bands due to the analyte, but near to them (so the noise level is similar). The water vapour absorbance bands in this region are often much greater than the noise level, so the WV pre-treated spectra were used to determine the noise levels. The calculated RMS noise levels are plotted in Figure 6.13a. Since the water vapour lines are not always completely removed by subtraction, some of the noise estimates are larger than they should be; in fact, all the spectra with estimated noise levels $\gtrsim 3.0 \times 10^{-5}$ feature significant non-noise variation in the region chosen. Taking this into account, a slightly pessimistic estimate of the average RMS noise level is $\hat{\sigma}_{\Delta X} = 2 \times 10^{-5}$.

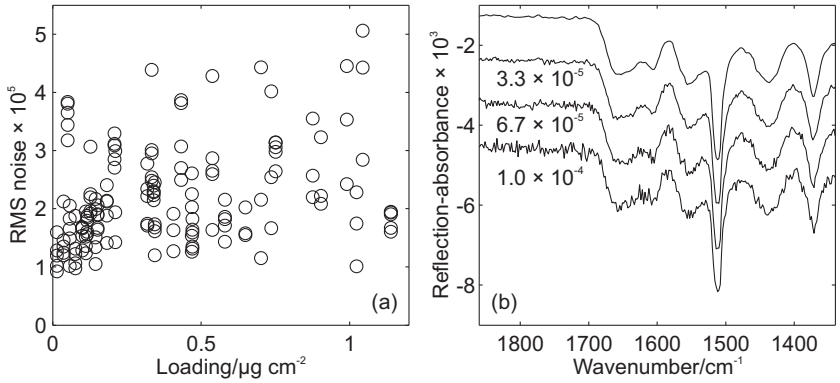

Figure 6.13: (a) Estimated noise levels calculated as the RMS of the residuals after subtraction of a straight line fitted through 1880–1730 cm^{-1} after subtraction of water vapour absorbance. (b) Spectrum of 0.9 $\mu\mathrm{g\,cm}^{-2}$ of acetaminophen on glass before (top) and after addition of various noise levels. The number by each spectrum is the standard deviation of the added noise.

To test the effect of increasing the noise, normally distributed noise was added to the spectra. Three noise levels, $\sigma^{+}_{\Delta X}$, were chosen: 3.3×10^{-5}, 6.7×10^{-5}, and 1.0×10^{-4}. For each noise level, a matrix of normally distributed random numbers drawn from a distribution with mean zero and standard deviation $\sigma^{+}_{\Delta X}$ was added to the $\mathbf{X}$ matrix prior to the cross-validation.[3] The noise additions and cross-validations

[3] Assuming the intrinsic noise in the spectra is $\hat{\sigma}_{\Delta X} = 2 \times 10^{-5}$, these additions result in noise levels of 3.9×10^{-5}, 7.0×10^{-5} and 1.0×10^{-4}, respectively. For the first-derivative spectra, the matrix of noise was first subjected to the same Savitzky-Golay filter as the spectra had been.

were repeated 20 times to obtain stable results. A typical spectrum (corresponding to a loading of 0.9 $\mu g\,cm^{-2}$), before and after addition of the various amounts of noise, is plotted in Figure 6.13b. The increase in noise is clearly visible at each step.

The RMSECV vs rank curves are plotted in Figure 6.14 and the RMSECV values at the previously determined optimal ranks are tabulated in Table 6.3. In general, the addition of noise increases the RMSECV, as would be expected. From Figure 6.14, the magnitude of the increase in the RMSECV generally increases as the rank is increased. This is also expected, since increasing the rank decreases bias in the predictions at the expense of increasing variance (see Section 3.3.2). At the optimal ranks (indicated by the vertical dotted lines in Figure 6.14, and for which the RMSECV values are given in Table 6.3), the increase in the RMSECV is slight: never more than 5 % when $\sigma^{+}_{\Delta X} = 6.7 \times 10^{-5}$ or 11 % when $\sigma^{+}_{\Delta X} = 1.0 \times 10^{-4}$ (five times the original noise level).

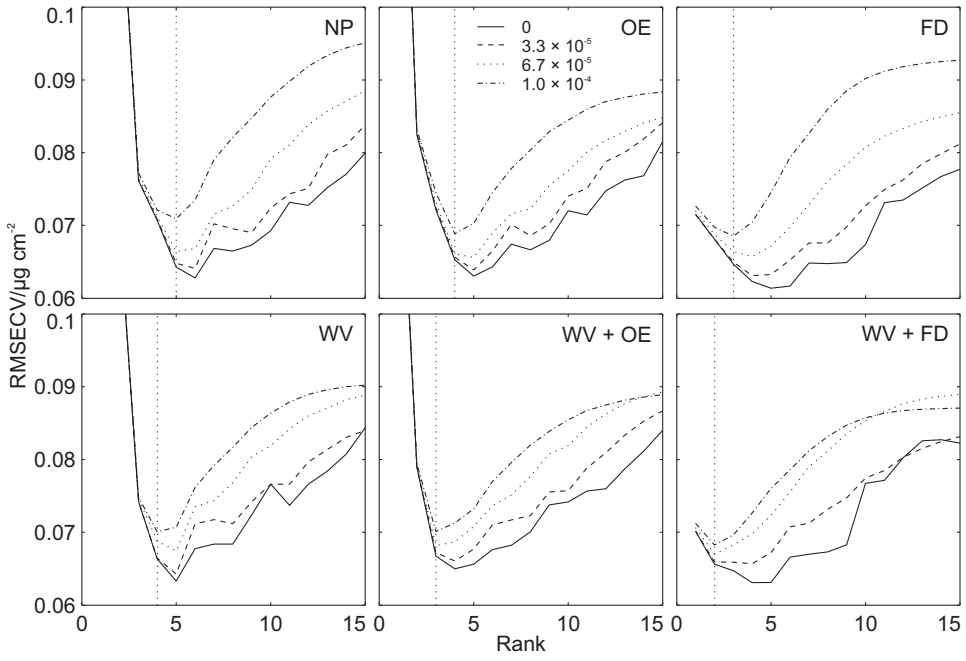

Figure 6.14: Cross-validation results before (solid line) and after addition of normally distributed noise. In each panel the optimum rank (Table 6.1) is indicated by the vertical dotted line. NP = no pre-treatment; OE = offset elimination; WV = water vapour subtraction; FD = first derivative.

The various pre-treatments all experience a similar degree of degradation at optimal rank, but with the NP model faring slightly worse than the others (perhaps because its optimal rank is greater). This situation is exaggerated at higher ranks (10–15), where the WV and OE pre-treatments provide a substantial reduction in the RMSECV when the noise level is high. These observations hint at a possible advantage for aggressive pre-treatments that is not apparent in the present study because the actual noise level is too low.

Table 6.3: Effect of photometric noise on the cross-validation results. The elements of the table are RMSECV values in $\mu g\,cm^{-2}$; A_{opt} is the optimal rank, determined previously (Table 6.1); $\sigma^{+}_{\Delta X}$ is the standard deviation of the noise that was added to the spectra. NP = no pre-treatment; OE = offset elimination; WV = water vapour subtraction; FD = first derivative.

Pre-treatment	NP	OE	FD	WV	WV + OE	WV + FD
A_{opt}	5	4	3	4	3	2
$\sigma^{+}_{\Delta X} = 0$	0.064	0.065	0.065	0.066	0.067	0.066
$\sigma^{+}_{\Delta X} = 3.3 \times 10^{-5}$	0.065	0.066	0.065	0.066	0.067	0.066
$\sigma^{+}_{\Delta X} = 6.7 \times 10^{-5}$	0.067	0.066	0.066	0.069	0.068	0.067
$\sigma^{+}_{\Delta X} = 1.0 \times 10^{-4}$	0.071	0.069	0.069	0.070	0.070	0.068

In general, these results support the assertion in the previous section that photometric noise is not a significant source of uncertainty in the present models. A consequence of this is that the spectra could be measured in a shorter time (by averaging fewer interferograms) without significantly impacting the results. For more complicated systems, with more interfering components (other APIs, excipients, etc.), models with more factors may be required, and the spectroscopic noise may take on greater importance.

6.7 Conclusions

A PLS model relating acetaminophen loading on a glass substrate to grazing-angle IRRA spectra has been constructed, having an RMSEP of ~0.07–0.08 $\mu g\,cm^{-2}$. This value implies detection limits well below the generally accepted limit for visual detection (~4 $\mu g\,cm^{-2}$), so the sensitivity of the method is certainly adequate for some cleaning validation applications. Since the apparent errors are heteroscedastic (increasing with increasing loading), the RMSEP would be smaller if the calibration range were restricted to lower loadings; the value found here is certainly not the smallest RMSEP that could be obtained for the current system. In fact, if the present results are used, but the RMSECV is calculated from the 14 standards with loadings <0.25 $\mu g\,cm^{-2}$, values around 0.04 $\mu g\,cm^{-2}$ are obtained. These RMSEP values are similar to those obtained in the studies [133, 134] mentioned in the Introduction to this chapter. The significance of this work is that it demonstrates that the IRRAS method works well with a glass substrate.

A variety of pre-processing steps were investigated, but none was found to improve the RMSECV or RMSEP significantly. They did reduce the number of PLS factors required, however; the combination of water vapour subtraction and first-derivative filtering reduced the optimal rank to 2, compared to 5 for untreated data. This study provides no decisive evidence either for or against the use of pre-treatments. There is no measurable improvement to the predictive ability of the model, but the spectra

are more easily interpreted: for example, in Figure 6.10, spectra that had been pre-treated by water vapour subtraction were plotted, because they more clearly revealed the phenomenon of interest than the untreated spectra. This improvement in interpretability is offset by the additional effort required to select and apply pre-treatments. In other applications, the leverage reduction due to pre-treatment might be a significant advantage of the use of pre-treatments. The rest of the results presented in this thesis were obtained without pre-treatments.

Two outliers were found in the calibration set by inspection of residual and leverage plots. Removing these outliers improved the RMSECV significantly and the RMSEP somewhat, demonstrating the importance of this step in model optimisation.

Confidence intervals for the test set were estimated by four methods, which were found to give very different results. The jack-knife procedure performed poorly, giving excessively narrow confidence intervals. Both the object and the residual bootstrap gave more reasonable, but probably slightly optimistic, results, while the approximate EIV formula gave unduly broad intervals. A possible reason for this is the failure to take into account the heteroscedasticity of the reference-method errors. Of these methods, the jack-knife is the easiest to implement, but can be ruled out on account of its inadequate results. The bootstrap methods are also quite simple to implement (and require no input besides the spectra and loadings), but may require a significant amount of computation if there are many calibration-set objects. In the present case, the bootstrap calculations took about 7 s using a 2.4 GHz Pentium processor and the software listed in Appendix C (but since the most time-consuming step of each iteration is the calculation of the PLS model, the calculation time is almost independent of the number of confidence intervals being calculated; the calculation time for the entire test set was 8 s). It should also be noted that PLS executes faster than most other chemometric algorithms, so bootstrapping with PCR or nonlinear methods would take considerably longer.

The formula-based approach is attractive because it requires very little computation and gives much more insight into the relative importance of the various sources of error. However, it requires independent error variance estimates for both $\mathbf{X}$ and $\mathbf{y}$. In general, since the purpose of the calibration is to replace a slow or inconvenient reference method with a rapid spectroscopic one, obtaining an error variance estimate for $\mathbf{X}$ is much more straightforward than measuring one for $\mathbf{y}$. Furthermore, the version of the EIV formula currently popular in the chemometrics literature (and the one used here) does not account for heteroscedastic errors. The theory required for the heteroscedastic case has been presented by Faber and Kowalski [74], but extracting a readily applied formula from their paper is not a trivial task. It may be simplified somewhat if the errors in the spectra are neglected: Section 6.6 above provides some evidence that this may be justified.

The best approach at present is probably to use both bootstrapping and the EIV formula during model validation, and then, provided that it performs reasonably, to use the EIV formula in routine application of the model.

Chapter 7

Residues of acetaminophen and aspirin on glass

7.1 Introduction

The previous chapter described the application of IRRAS to a model situation representing cleaning validation for a single API on a glass substrate. The work presented in this chapter (and in a recently submitted paper [137]) concerns the simultaneous quantification of two chemically similar APIs on glass. Aspirin and acetaminophen were selected as model compounds since they are inexpensive and safe to handle.

7.2 Experimental section

7.2.1 Materials

Aspirin (o-acetylsalicylic acid) and acetaminophen (4-acetamidophenol) were obtained from Sigma Aldrich and used without further purification. The solvents used were Milli-Q water, ethanol (solvent grade) and acetone (solvent grade), or mixtures of these components. Glass coupons ($15 \times 15\,\mathrm{cm}^2$) were cut from 3 mm-thick window (soda-lime float) glass and were roughened by bead-blasting on one side to prevent reflection from the back face.

7.2.2 Sample preparation

The spray method of sample preparation (see Section 4.3.2 and Ref. 135) was used for all results reported here. This gives more uniform coatings (in terms of the macroscopic distribution of the analyte)

than the smear technique, but requires an additional calibration step to determine the loadings. Solvents of various volatility were used (acetone, ethanol, water, and mixtures thereof), with the spraying distance adjusted, accordingly, between about 0.5–1 m. A less volatile solvent allows the airbrush to be held further from the coupon, facilitating the preparation of more homogeneous samples. Both APIs were dissolved (at various, known concentration ratios) in the same spraying solution. The loadings were determined rinsing the samples with ethanol and analysing the rinsate by UV colorimetry and CLS regression, as described in Section 4.3.3. The uncertainty (RMSEP) in the loadings is about $0.013\ \mu g\,cm^{-2}$.

7.2.3 IRRAS instrumentation and data collection

The grazing-angle IRRAS instrument was as described in Section 4.2. Transmission measurements were made using the internal sample compartment of the same spectrometer with a DTGS detector. The wavenumber range was 4000–1000 cm^{-1} and the resolution was 4 cm^{-1}. Single-beam background spectra (I_0) were obtained from clean glass coupons by averaging 100 interferometer scans (about 30 s). Five to ten sample spectra (I) were collected from different regions of each loaded coupon by averaging 50 interferometer scans (about 15 s) per spectrum. The IRRAS was calculated as $\log_{10}(I_0/I)$. Since there is a considerable (approximately 20 cm) beam path through laboratory air outside the spectrometer, absorbance bands arising from small changes in CO_2 and H_2O vapor concentrations often appear in the spectra. To test whether these features pose any problem for the chemometric methods used, a single background spectrum was used for each batch of 4 to 8 samples, rather than for each sample. Consequently, in most spectra, the atmospheric absorption bands are stronger than the bands due to the analyte, reinforcing the need for chemometric methods.

7.3 Results and discussion

7.3.1 IRRA spectra

Typical IRRAS of acetaminophen and aspirin on glass are shown in Figure 7.1 where they are compared with absorption spectra obtained in transmission from pressed KBr pellets. The effective loadings for the pellets were determined by dividing the mass of API (about 1 mg) by the pellet area (1.33 cm^2). The transmission spectra in Figure 7.1 have been scaled by the ratio of the IRRAS sample loading to the KBr pellet sample loading.

As discussed in Chapters 2 and 5, IRRAS can exhibit both positive and negative features, depending on the refractive indices of the substrate and film materials, the polarisation of the light and the inci-

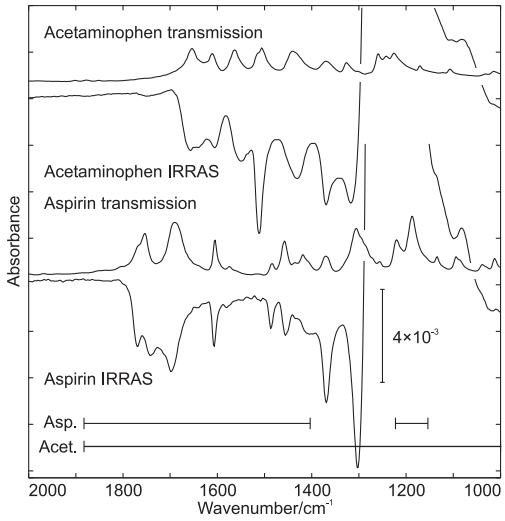

Figure 7.1: IRRA (glass substrate) and absorption (in transmission; KBr pellet) spectra of acetaminophen and aspirin. The transmission spectra have been scaled to the same effective loading as the IRRAS (acetaminophen 2.08 $\mu g\,cm^{-2}$; aspirin 2.95 $\mu g\,cm^{-2}$). The spectra have been offset along the ordinate for clarity; the abscissa scale is indicated. The horizontal bars labeled Asp and Acet indicate the wavenumber ranges used to build PLS-1 models for aspirin and acetaminophen, respectively.

dence angle. In addition, band shape distortions due to optical effects [138] and spectroscopic features due to the substrate can appear [139]. The strong positive feature at about $1280\,cm^{-1}$ in the IRRAS is due to the Si–O stretching mode of the substrate. As a consequence of this strong resonance, the real part of the refractive index crosses that of the incident medium (air; $n \approx 1$) while the imaginary part remains very small. This results in a deep minimum in the reflectance of the bare substrate and a correspondingly greater relative change in the reflectance due to the layer. To the blue of the substrate feature, the API bands are negative while to the red they are positive. These observations are consistent with the calculations in Section 5.2.5.

7.3.2 Model optimisation by cross-validation

The chemometric modeling was achieved by using the PLS regression method with cross- and test-set validations, as described in Chapter 3. The procedure entails two parts: first, generation of an optimised calibration model; and second, evaluation of the quality of the model. To ensure that the optimisation process did not bias the quality evaluation, the 45 samples were split into separate calibration (31 samples) and test (14 samples) sets, chosen to cover simultaneously the approximate loading range 0–2 $\mu g\,cm^{-2}$ for both compounds. The loadings are plotted in Figure 7.2. Using a larger proportion of the available samples for the test set would have improved the precision of the quality evaluation, but at the expense of a poorer model. For optimisation, model quality was judged in terms of the RMS

error of cross-validation (RMSECV). Once the model parameters had been selected, the RMS error of prediction for the test set (RMSEP) was calculated by applying the optimised model to the test set.

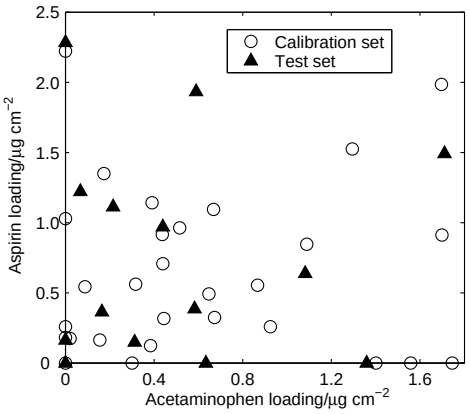

Figure 7.2: Loadings of acetaminophen and aspirin on glass. The 45 samples were split into a calibration set of 31 (open circles) and a test set of 14 (solid triangles). The target range was 0–2 μg cm^{-2} for each compound.

The models were built using the PLS-1 algorithm, in which the two analytes are modelled separately, potentially using different spectroscopic regions and ranks (numbers of PLS factors). The wavenumber ranges were selected by inspection of the spectra with the aim of maximising the incorporation of analyte bands (Figure 7.1) while minimising the influence of features that are uncorrelated with the analyte loading. Of several ranges that were trialled, those indicated in Figure 7.1 and Table 7.1 gave the lowest RMSECV values and were used for the results presented here. The range chosen for acetaminophen essentially encompasses the polar-group region (1800–1300 cm^{-1}) and part of the finger-print region. For aspirin, a narrower range was used and the prominent Si–O feature was excluded. Inclusion of the API O–H and N–H bands did not improve the RMSECV.

Several spectroscopic pre-processing procedures were investigated with the aim of building better-optimised models. Mean centring, which is commonly used with PLS but is unsuitable for some kinds of data [140], did not reduce the optimal rank or significantly reduce the RMSECV. The first-derivative, (quadratic Savitzky-Golay filter [85, 86, 136] with 15 smoothing points) reduced the optimal rank, but did not significantly change the RMSECV. These results are consistent with those of the previous chapter, and the results that follow were obtained without any pre-processing.

In the cross-validations, the spectra for each sample were treated collectively, being either included or left out together. The resultant RMSECV values, plotted against rank in Figure 7.3, exhibit a broad minimum between ranks of 8 and 12. The optimal ranks indicated by two methods (an F-test with $\alpha = 0.25$ and Martens' method [82] with $s = 0.02$; see Section 3.3.2) were 8 for each compound (Fig-

ure 7.3 and Table 7.1), corresponding to the left edge of the broad minimum.

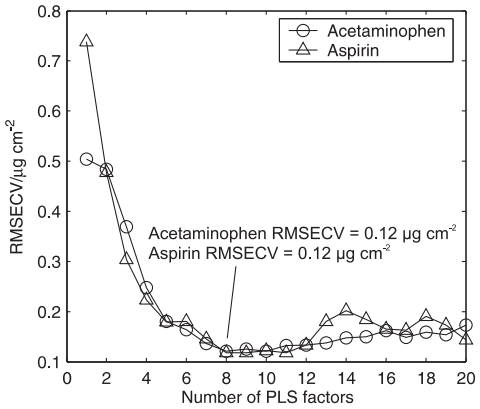

Figure 7.3: RMS errors of cross validation plotted against rank for acetaminophen (circles) and aspirin (triangles). The optimal model ranks (determined by the methods of Refs 70 and 82) are 8 for both APIs; the corresponding RMSECV values are indicated.

The cross-validation plots for acetaminophen and aspirin (Figures 7.4 and 7.5, respectively) at a rank of 8 are very similar and indicative of a satisfactory model. As the loadings increase, the spread of the predictions for individual spectra for each sample about the mean for the sample increases, but the deviations of the means from the diagonal ideal-fit (zero intercept and unit slope) line do not change significantly. This apparent heteroscedasticity is at least partly due to sample heterogeneity, as discussed below.

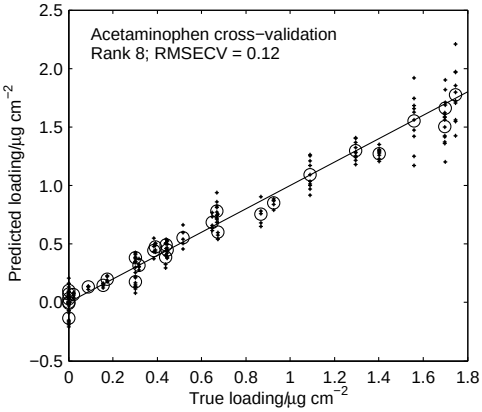

Figure 7.4: Predicted versus true loadings for the acetaminophen cross-validation with a rank of 8. Dots are predictions from individual spectra while open circles are mean predictions per sample. The solid diagonal (zero intercept and unit slope) is the line of perfect agreement.

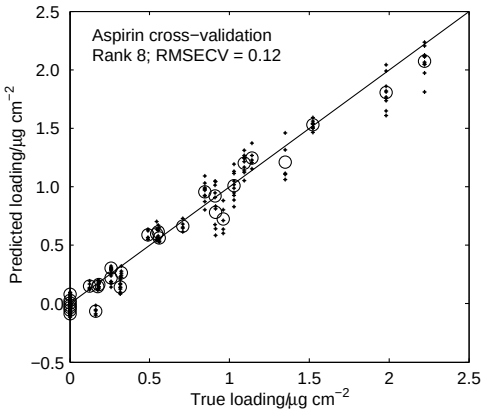

Figure 7.5: Predicted versus true loadings for the aspirin cross-validation with a rank of 8. Dots: predictions from individual spectra, open circles: mean predictions per sample. The solid diagonal (zero intercept and unit slope) is the line of perfect agreement.

7.3.3 Test-set validations and bias tests

The quality evaluations were performed by applying the calibration models to the test sets. The results for acetaminophen (Figure 7.6) essentially mirror those of the cross-validation. However, as shown in Figure 7.7a, the predictions for the test set of aspirin appear to show a systematic bias (referred to here as type-1 bias) to low values, with only two of the mean predictions falling above the ideal-fit line. A statistical test, based on the joint confidence region for the slope and intercept of a least-squares line fitted through the predicted and true loadings (see Appendix A.6), confirms that this bias is significant ($p_1 = 0.01 < \alpha$) at the $\alpha = 0.05$ level. By the same measure, type-1 bias is absent from the other rank-8 cross and test-set validations (Table 7.1).

A known cause of bias in PLS regression is the use of too few factors to account for all of the relevant variation in the spectra [74]. Consistent with this, when the rank for aspirin is increased, the type-1 bias is reduced. At a rank of 11, the right edge of the broad minimum in Figure 7.3, it is insignificant ($p_1 = 0.07$) and the RMSEP slightly decreased (Figure 7.7b and Table 7.1).

A second kind of bias (referred to here as type-2 bias) occurs when the loading of one compound has a systematic effect on the predicted loading of the other. This would not necessarily show up in the test for type-1 bias since the loadings of the two compounds are uncorrelated (Figure 7.2). However, it would be revealed by an equivalent test that investigates correlation between the residuals (differences between predicted and true loadings) for one compound and the loadings of the other, where the ideal-fit line now has zero slope and intercept. For acetaminophen (Figure 7.8), this test indicates no significant type-2 bias in either the cross or test-set validations at a rank of 8 (Table 7.1). For aspirin (Figure 7.9a), type-2 bias is insignificant in the rank-8 cross-validation, but it is significant ($p_2 = 0.002$) in the test-set

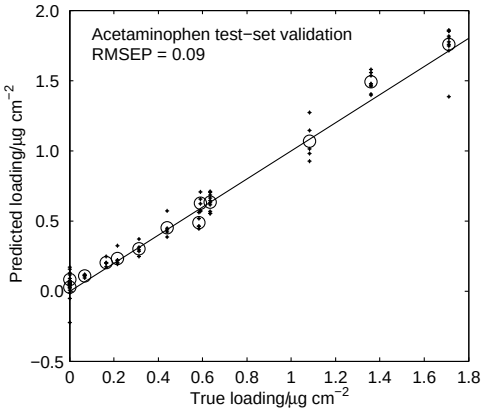

Figure 7.6: Predicted versus true loadings for the acetaminophen test-set validation with rank of 8. Dots: predictions from individual spectra, open circles: mean predictions per sample. The solid diagonal (zero intercept and unit slope) is the line of perfect agreement.

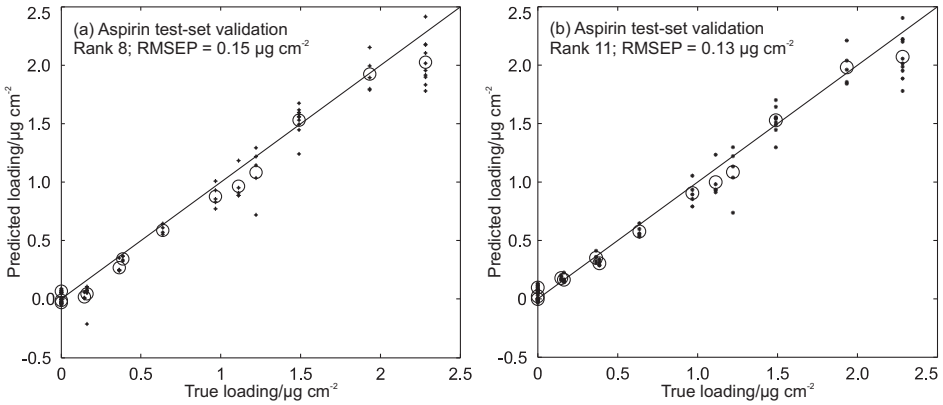

Figure 7.7: Predicted versus true loadings for the aspirin test-set validations with ranks of 8 (a) and 11 (b). Dots indicate predictions from individual spectra, while open circles show the mean predictions per sample. The solid diagonals (zero intercept and unit slope) are the lines of perfect agreement.

validation at that rank. As was the case for type-1 bias, the type-2 bias weakens as more factors are included and, for the 11-factor model (Table 7.1 and Figure 7.9b), $p_2 = 0.17$.

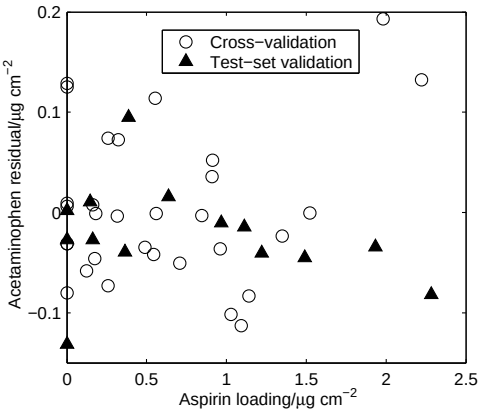

Figure 7.8: Loading residuals for acetaminophen in the cross- (open circles, $p_2 = 0.27$) and test-set (solid triangles, $p_2 = 0.51$) validations plotted against the aspirin loading.

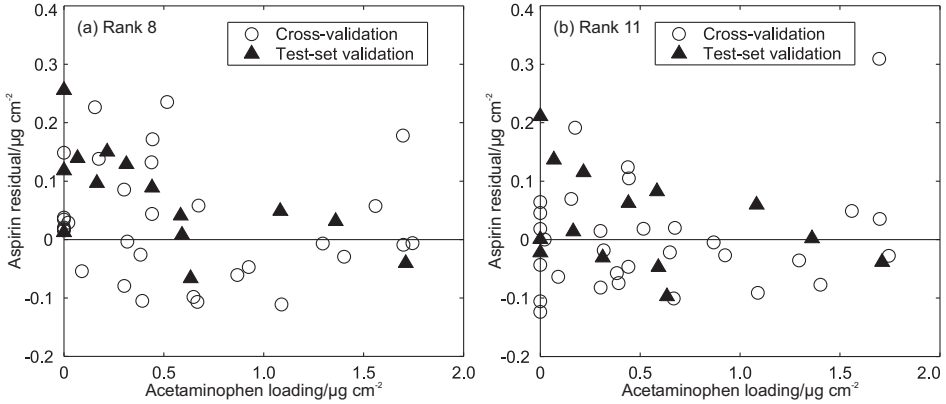

Figure 7.9: Loading residuals for the cross validations (open circles) and test-set (solid triangles) validations of aspirin plotted against acetaminophen loading for ranks of 8 (a) and 11 (b). The only significant type-2 bias revealed (at the $\alpha = 0.05$ level) in these data is for the rank-8 test-set validation, for which $p_2 = 0.002$.

7.3.4 Sample heterogeneity

An important limitation of the data presented here is that IRRA spectra are measured from several different regions of each sample, whereas the loading determined by the reference method is the average for the entire coupon. Although the latter can be determined with good precision, sample heterogeneity means that it may not accurately represent the true, local loading pertaining to any particular IRRAS measurement. As described in Section 4.3.5, the relative standard deviation of the local loadings has been estimated to be ~7 %. This error is equivalent to the presence of significant and heteroscedastic

Table 7.1: Model parameters and statistics for cross and test-set validations for acetaminophen and aspirin mixtures on glass.

	Acetaminophen	**Aspirin**	
Loading/$\mu g\,cm^{-2}$	0–1.7	0–2.2	
Wavenumber/cm^{-1}	1880–1000	1880–1398 & 1225–1157	
Rank	8	8	11
Cross validations:			
RMSECV/$\mu g\,cm^{-2}$	0.12	0.12	0.12
$RMSECV_{\lhd 1}/\mu g\,cm^{-2}$	0.09	0.10	0.08
R^2	0.96	0.97	0.97
p_1	0.10	0.14	0.67
p_2	0.27	0.21	0.63
Test-set validations:			
RMSEP/$\mu g\,cm^{-2}$	0.09	0.15	0.13
$RMSEP_{\lhd 1}/\mu g\,cm^{-2}$	0.06	0.08	0.06
R^2	0.97	0.96	0.97
p_1	0.24	0.01	0.07
p_2	0.51	0.002	0.17

errors in the reference loading values.

There are two consequences of these errors. The first is that unweighted regression methods will not give the best estimate of the regression vector; but due to the absence of appropriate algorithms for weighted multivariate regressions in most commercial software (and since the number of measurements is sufficient that improvements to the model would probably be minor) we have not attempted to address this issue. The second consequence is that the RMS prediction errors in Table 7.1 represent convolutions of contributions intrinsic to the spectroscopic (IRRAS) method with those arising from sampleheterogeneity and are therefore pessimistic estimates of the true accuracy of the spectroscopic method [91]. Other workers have attempted to correct this type of shortcoming by subtracting the estimated error in the reference values [74, 91], but while heteroscedastic reference method errors do not imply heteroscedastic errors in the IRRAS method, it is not immediately obvious how to generalise the technique to accommodate the heteroscedastic reference errors, and this approach has not been pursued here. This effects of reference-method errors in model validation are discussed in greater detail in Section 3.4.

7.3.5 Detection limits

A method for cleaning validation is more usefully characterised by a detection limit than by an RMSEP. Prediction uncertainties and figures of merit for multivariate calibration are topics of active research in chemometrics [90]. As discussed in Section 3.4.3, the detection limit in multivariate calibration cannot be specified on a per-model basis but must, formally, be calculated for every new sample. However, the treatment below shows that, provided the calibration set is sufficiently large, the detection limit is nearly constant for most samples, and a method-specific detection limit can be a useful indicator of its performance. It must be stressed that in actual application of a model, the detection limit should be calculated for every new measurement, by either the method presented in Section 3.4.3 or that given by Boqué et al [97].

If measurement errors are neglected, an approximate expression for the error variance of the predicted loading, y_u, for a new sample is [97]

$$V(y_u) \approx (1 + h_u) \times \text{MSEP} \tag{7.1}$$

where h_u is the leverage of the new sample, a weighted measure of its distance from the origin of the model space. A key consequence of Equation 7.1 is that the error associated with an estimated loading depends on contributions to the spectrum from all species, not just those from the analyte. However, if there are many degrees of freedom in the calibration set (many calibration standards and/or few factors) and the new sample is not unusual, the leverage is likely to be negligible ($h_u \ll 1$). In the present example, the mean test-set leverages are small (0.03 for the 8-factor acetaminophen model and 0.05 for the 11-factor aspirin model), although there are several spectra with leverages of 0.1 or greater.

For all samples with negligible leverage, the MSEP can be taken as the prediction error variance and the detection limit can be obtained from Equation 3.62:

$$L_D \approx (t_{1-\alpha,n} + t_{1-\beta,n}) \times \text{RMSEP} \tag{7.2}$$

where n is the number of samples in the test set. It must be emphasised that there is no guarantee that a low-loading sample will have low leverage, so while the detection limits calculated in this manner are indicative of the general performance of a model, the sample-specific detection limit, taking into account the leverage, should be calculated for all new samples.

With $\alpha = \beta = 0.05$, Equation 7.2 yields detection limits of approximately 0.3 and 0.4 $\mu g\,cm^{-2}$ for acetaminophen and aspirin respectively. But because of the significant contribution of error in the

reference loadings, the RMSEP values and hence these detection limit estimates are biased high. Due to the heteroscedasticity in the reference loading errors, this effect is exacerbated by the inclusion of samples with higher loadings. For this reason, the RMSECV and RMSEP values were re-determined after reducing the data sets to samples with loadings of $<1\ \mu g\,cm^{-2}$ in the analyte of interest, levels that are below those that permit naked-eye detection. The values are listed in Table 7.1 as $RMSE_{<1}$; they correspond to a detection limit of $\sim 0.2\ \mu g\,cm^{-2}$ for both APIs, a value that is more representative of the true capabilities of the technique.

7.4 Conclusions

The results presented here demonstrate that grazing-angle fiber-optic IRRAS can be used to quantify simultaneously two chemically similar APIs on a glass surface at loadings well below those visible to the naked eye, without making frequent background measurements or taking any special precautions regarding absorption by atmospheric gases. This is made possible by the use of PLS regression, although other multivariate inverse calibration methods are also likely be suitable. The main advantage of the method over the traditional swab-HPLC technique is that measurements can be made very rapidly (in 30 s or less) and *in situ*.

An important point emphasised by this work is the risk of under-fitting when employing the usual safeguards against over-fitting. These methods are intended to strike a balance between excessive bias due to under-fitting and excessive variance due to over-fitting but in this work seemed to err on the side of under-fitting, as illustrated by the significant bias encountered in the aspirin test-set validation at rank 8.

Later work will address the issues of heteroscedastic errors. In application to real samples, the error introduced by heterogeneity can be mitigated by making measurements at several different positions. For this reason, it is preferable to have an estimate of the uncertainty that pertains to homogeneous samples. In the first instance, additional effort will be devoted to improving the homogeneity of the standards to enable better characterisation of the errors intrinsic to the method and more thorough treatments of prediction intervals and detection limits. Secondly, it should be noted that other factors can contribute to heteroscedastic deviations. For example, as the loading increases, the risk of nonlinear spectroscopic response also increases: this would introduce another loading-dependent contribution to the error arising from lack of model fit. It may be possible to account for such effects by employing non-linear multivariate methods.

Later work will address the issue of separating the error intrinsic to the method from error due

to sampling of a heterogeneous standard. In application to real samples, the sampling error can be mitigated by making measurements at several different physical locations on the substrate surface, so it is important to have an accurate estimate of the error pertaining to a homogeneous sample. In the first instance, additional effort will be devoted to improving the homogeneity of the standards and to more thorough treatments of prediction intervals and detection limits. Finally, it should be noted that other factors can contribute to heteroscedastic deviations. For example, as the loading increases, nonlinear spectroscopic responses can introduce features that might be better described by approaches such as neural networks and polynomial PLS [141].

Chapter 8

Stainless steel substrates: effect of surface roughness

8.1 Introduction

Stainless steel is frequently used for contact surfaces in manufacturing processes [16]. Consequently, it is an important target for cleaning validation. While contact surfaces are likely to begin life highly polished, they are subject to wear, so it is important to evaluate the effect of surface roughness on any cleaning validation method. Can a single chemometric-IRRAS model give accurate contaminant loadings for a range of surface finishes? This chapter presents an empirical investigation of the effect of substrate roughness on the grazing-angle IRRAS of sodium dodecyl sulfate (SDS, a common surfactant) and acetaminophen deposited on stainless steel.

All the scanning electron microscope (SEM) images were collected by Michelle Hamilton, as were the SDS IRRA spectra [106]. The main results of the SDS work have been published elsewhere [142]: but in the first part of this chapter, certain features of the data that were not discussed in the earlier publication are elucidated. The results of the parallel study with acetaminophen are presented in the second part of this chapter and are also being prepared for publication elsewhere.

8.1.1 Experimental

Square $150 \times 150\,\mathrm{mm}^2$ coupons were cut from three finishes (described here as polished, smooth and rough) of ~0.73 mm thick, 316-grade stainless steel sheet (McMaster-Carr). The smooth steel has the standard mill finish; the rough material has a scoured appearance with no preferential grain; and the polished steel has a mirror finish, although close inspection reveals a slight directional grain. The sur-

faces of loaded and unloaded coupons were imaged using a Leica S440 scanning electron microscope (SEM) operating with a beam voltage of 10 kV. The beam current (I) was varied, according to the nature of the surface, from 3 nA for the polished coupons to 2 nA for smooth and 20 pA for rough.

Coupons were loaded with the analyte by the spray method (Section 4.3.2). SDS (BDH laboratory supplies) was dissolved in milli-Q water, while acetaminophen (Aldrich) was dissolved in ~1 mL of ethanol and diluted to ~20 mL with milli-Q water. The loadings were determined after IRRAS measurement by rinse analysis. Acetaminophen was measured by UV colorimetry, as described in Section 4.3.3. For SDS, a ^{1}H-NMR method was used. A brief description of this method is given by Hamilton et al. [142], and details will be presented in Michelle Hamilton's PhD thesis [106].

IRRA spectra were measured as described in Section 4.3.4, averaging 50 scans for each spectrum. A new background spectrum was measured (from a coupon of the same roughness) for each sample. The effect of varying the roughness of the background coupon will also be discussed in Michelle Hamilton's thesis [106].

8.2 Sodium dodecyl sulfate

In Ref. 142, it was shown that satisfactory chemometric models could be built for SDS on stainless steel coupons with three types of surface finish. In some cases, "combined-surface" models could be built from the standards for two surface finishes, in spite of marked differences in sensitivity between the surfaces. The purpose of the first section of this chapter is to investigate the properties of the spectra and of the chemometric methods that allow these successful calibrations.

8.2.1 Scanning electron micrographs

SEM images of the three surfaces coated with SDS by the spray method are shown in Figure 8.1. These micrographs reveal clear differences between the surface finishes. The polished coupons (top row of Figure 8.1) have areas that are flat (on the sub-μm scale) apart from relatively sparse, mm-length scratches that are separated laterally by tens of μm and have transverse dimensions (widths and depths) of ~1 μm. The smooth coupons (middle row) have a much greater density (separations <1 μm) of microscopic scratches and folds, which are shorter (tens of μm) than the polished-surface scratches, but have similar lateral dimensions. Less common are larger and much deeper scores, which run at slight diagonals across the mid-row images of Figure 8.1. On the rough-surface coupons, the deep, larger-scale scores are much more prevalent.

The spray technique is seen to provide a consistent and even spread of aerosol droplets across

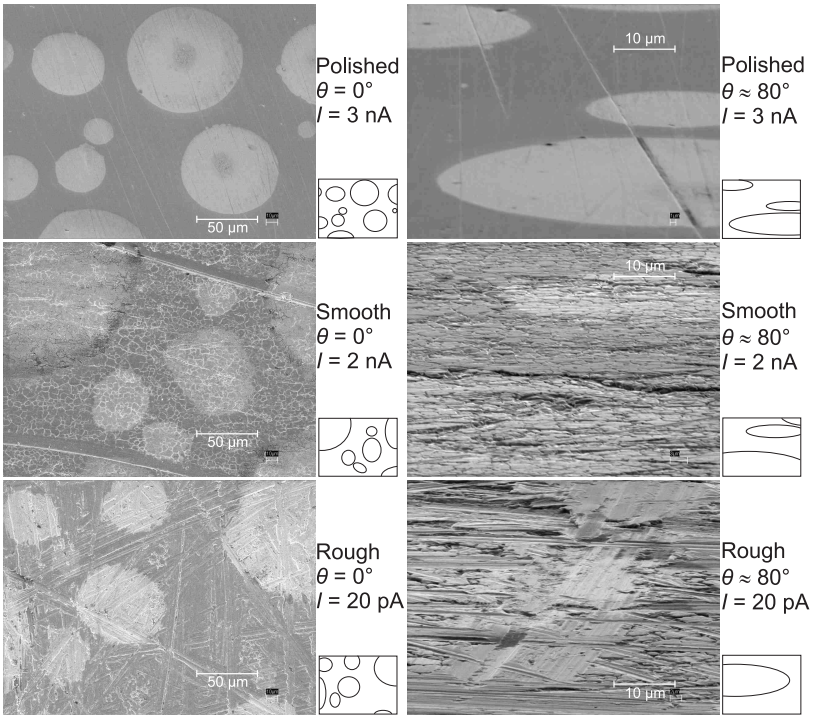

Figure 8.1: SEM images from three finishes of stainless steel after loading with ∼1 µg cm^{-2} of SDS by the spray technique. The schematic to the right of each photo indicates the location of the SDS spots. The images in the left column were taken perpendicular to the surface at ×1000 magnification. Those on the right were taken at ∼80° to the normal (the same as the IRRAS incidence angle) and at ×5000 magnification. The top-row images are for the polished finish, the middle row for the smooth finish and the bottom row for the rough finish. Beam currents are indicated by the parameter *I*.

the coupon surfaces. Evaporation of the solvent leaves a "polka-dot" patterning of near-circular SDS discs with diameters in the range 10–100 µm. The thickness of the SDS spots can be estimated as follows. Based on the concentration of the spraying solution and the duration of spraying, the loading is approximately 1 µg cm^{-2}. From a lower-magnification image of a larger area (see Figure 8.7 later in this chapter), it was estimated that 45% of the surface is actually covered, meaning that the loading for the covered area is ∼2.2 µg cm^{-2}. Assuming a density of 1.1 g cm^{-3} [143], the thickness is ∼20 nm. There is clear evidence, especially for the rougher surfaces, that part of each droplet flows into the scratches and grooves prior to complete solvent evaporation; this gives rise to the jagged disk edges that are particularly apparent in the smooth- and rough-surface micrographs.

8.2.2 Spectra

Representative spectra of SDS on the three surfaces are plotted in Figures 8.2 and 8.3. These spectra have been normalised by dividing the RA by the loading and taking the average over all samples for each surface, and then further treated by water vapour subtraction and polynomial baseline removal. In

the fingerprint region (Figure 8.2), the most obvious difference is in intensity: the three spectra have similar shapes, but the polished-surface spectrum is ~2 times as intense as the smooth-surface one, which, in turn, is somewhat stronger than the spectrum from the rough surface. The most obvious explanation of this observation is that, on the rougher surfaces, some of the analyte is present in small grooves and crevices that are inaccessible to infrared light, so that the effective loading is reduced.

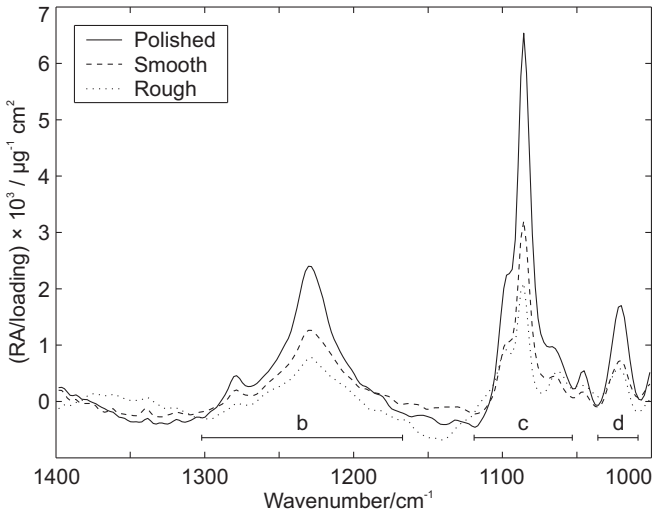

Figure 8.2: Mean concentration-normalised IRRAS for SDS on three types of stainless steel surface; fingerprint region. The ranges for the integrals are indicated by the horizontal bars.

In the C–H stretch region (Figure 8.3), however, a dramatic difference in shape for the polished-surface spectrum can be seen. The bands are (arguably) in essentially the same positions in all three spectra, but the intensity ratios are different for the polished surface. The bands at ~2960 cm^{-1} and ~2935 cm^{-1} are much more prominent, while that at ~2850 cm^{-1} is less so. The reason for these differences is not clear. It is known that C–H stretching bands in long alkyl chains are strongly dependent on molecular conformation [144, 145], so, if the preferred conformation is sensitive to the type of surface, different spectra may be seen on different surfaces. Another possible explanation lies in the fact that, if the molecules have a preferred orientation relative to the surface, the surface selection rule (see Chapter 2) reduces absorption by dipoles that are not normal to the surface. If the average orientation of the chains depends on the surface finish, so will the spectrum. There is insufficient evidence to evaluate either of these possibilities, however, and a thorough investigation is beyond the scope of this work.

The similarity in the "shape" of the spectra can be quantified by the correlation matrices, the elements of which are calculated according to Equation A.2 in Appendix A and which are presented in Table 8.1. In the fingerprint region, the surface has very little effect on the shape of the spectrum and R is large ($\gtrsim$0.90) for each pair of spectra. Contrastingly, in the C–H stretch region, R remains

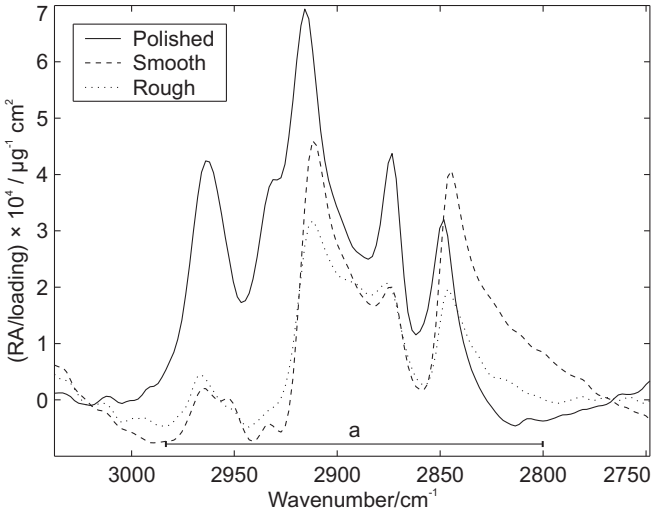

Figure 8.3: Mean concentration-normalised IRRAS for SDS on three types of stainless steel surface; C–H stretch region. The range for the band integral is indicated by the horizontal bar.

high (~0.90) when comparing the smooth- and rough-surface spectra, but is very much smaller when comparing the polished-surface spectrum with either of the others.

Table 8.1: Correlation coefficients (R) between IRRAS of SDS on different finishes of stainless steel.

	C–H stretch			Fingerprint		
	P	S	R	P	S	R
P	1	0.42	0.65	1	0.99	0.89
S		1	0.90		1	0.90
R			1			1

8.2.3 Band integrals

The obvious differences in intensity in Figures 8.2 and 8.3 can be quantified by measuring band integrals (without normalisation of the spectra). The bands or groups of bands to be integrated are indicated by the horizontal bars in Figures 8.2 and 8.3. Ideally, the bands in the C–H stretching region would be integrated individually, but due to baseline variations and other imperfections in the spectra, it is much easier to treat them as a single integral.

Integrals for the four bands are plotted as a function of SDS loading in Figure 8.4. Best-fit lines (with zero intercept) are also plotted. In all cases, the slope is greatest for the polished surface. The slopes for the other two surfaces are quite similar, with the smooth-surface slope being slightly greater in most cases.

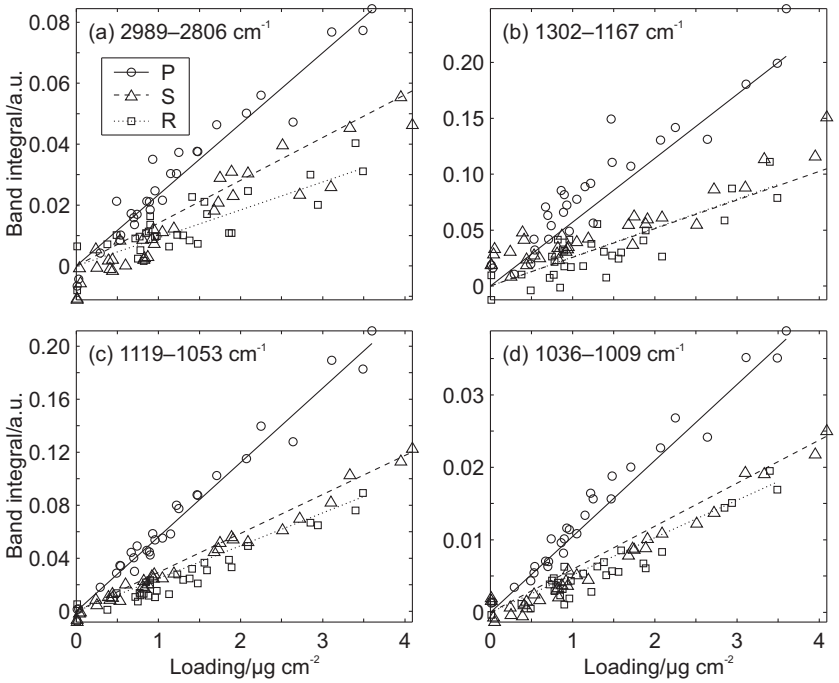

Figure 8.4: Integrals for four RA bands of SDS on stainless steel with three different finishes. Polished, smooth and rough surfaces are represented by circles, triangles and squares, respectively. The lines are for least-squares fits with zero intercept; solid, dashed and dotted lines are for polished, smooth and rough surfaces, respectively.

8.2.4 PLS modelling

The information from the correlation and band integral measurements above can be summarised:

- In the fingerprint region, all three surfaces give similar-shaped spectra.
- In the C–H stretch region, the smooth and rough surfaces give similar-shaped spectra, but the shape of the polished spectrum is quite different.
- In both regions, the following intensity relationships hold:

$$\text{RA(polished)} \gg \text{RA(smooth)} \gtrsim \text{RA(rough)}$$

These observations have some implications for several-surface PLS models. If the spectra from two surfaces have the same shape but different ratios of intensity to loading, it is impossible in principle to calibrate for both surfaces simultaneously: for any loading on the first surface, there will be a different loading on the second surface that will give an identical spectrum. On the other hand, if there is a difference in shape, PLS can exploit it to generate a working model. This phenomenon is illustrated below by a simple simulation using a single Gaussian band. The "change of shape" between three

"surfaces" is modelled as a shift in wavelength of the peak maximum, which is accompanied by a change of the band integral for a given loading.

The spectra for the three surfaces at a loading of one unit are given in Figure 8.5a. Each spectrum is a row of a matrix **S**, and the PLS model is built using these three spectra and a vector of ones for **y** (see Section 3.2.4). The RMS fit error (RMSEC without correction for the degrees of freedom) is 0.47 units for a one-factor model, 0.07 for a two-factor model, and 0 for a three-factor model. The three-factor PLS model will therefore give the correct loading for a sample prepared on *any* of the three surfaces despite the fact that the spectra from each surface are significantly different.

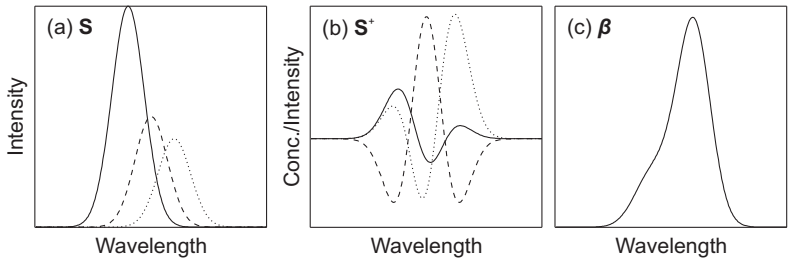

Figure 8.5: Simulated spectra. (a) Spectra (rows of **S**) at an arbitrary, common loading; (b) Individual regression vectors (columns of $\mathbf{S}^+$); (c) Combined regression vector.

The explanation behind this observation is straightforward. Since the spectra differ in shape, there is a component of each spectrum that is orthogonal to the subspace spanned by the other two spectra. These components can be found as the columns of the pseudoinverse of **S**, given by (see Section 3.2.1)

$$\mathbf{S}^+ = \mathbf{S}^{\mathrm{T}} \left(\mathbf{S}\mathbf{S}^{\mathrm{T}} \right)^{-1} \tag{8.1}$$

Each of the columns of $\mathbf{S}^+$ (plotted in Figure 8.5b) is orthogonal to the rows of **S** (Figure 8.5a) corresponding to the spectra from the other surfaces. Taking the dot product of a spectrum ($\mathbf{x}$) and the ith column of $\mathbf{S}^+$ will give the correct loading if $\mathbf{x}$ is from the ith surface, and zero otherwise. The vector $\boldsymbol{\beta}$, formed by adding the columns of $\mathbf{S}^+$, will give the correct loading for any spectrum, regardless of the surface:

$$\begin{aligned} \boldsymbol{\beta} &= \mathbf{s}_1^+ + \mathbf{s}_2^+ + \mathbf{s}_3^+ \\ \mathbf{x}_q\boldsymbol{\beta} &= \mathbf{x}_q(\mathbf{s}_1^+ + \mathbf{s}_2^+ + \mathbf{s}_3^+) \\ &= \mathbf{x}_q\mathbf{s}_q^+ \end{aligned}$$

where q denotes the surface, and is 1, 2 or 3; $\mathbf{s}_i^+$ is the ith column of $\mathbf{S}^+$ and $\mathbf{x}_q$ is a new spectrum on one of the three surfaces. Comparing $\boldsymbol{\beta}$ (plotted in Figure 8.5c) with the regression vector from the

three-factor PLS model reveals that they are identical within rounding error.

However, a calibration model relying on this kind of feature in the data may not be ideal. While this approach works perfectly for noise-free data, it is very sensitive to noise and other perturbations. The reason for this sensitivity is that the significant overlap between the spectra means that only a small component of the total signal is actually selective for the analyte. If the spectra can be reliably classified as originating from one or the other surface beforehand, more-robust single-surface models with fewer factors could be used. Essentially, the spectrum of the analyte on a different surface acts as an interfering species with a spectrum unusually well correlated with the analyte spectrum.

Based on these observations, a few predictions can be made about the performance of mixed-surface models for SDS:

- Smooth and rough samples may be almost freely interchangeable, since the intensity difference is small (i.e. the smooth and rough surfaces appear as one surface). In this case, a combined model with slightly higher prediction errors than the individual models is expected.
- A combined model that includes polished-surface standards and is limited to the fingerprint region should fail, since there is a dramatic intensity difference but very little difference in shape.
- A combined model including the polished-surface standards and including (or limited to) the C–H stretch region might work, since the difference in shape is significant. However, the RMSECV may be somewhat higher than for the individual-surface models.

Testing these predictions requires building a number of models and evaluating them by cross-validation. The data set consists of 23 polished, 27 smooth, and 26 rough-surface samples. Eight spectra were measured for each sample but, for simplicity, the averaged spectra are used for this work. Two wavenumber ranges are used: the "fingerprint" range (1300–1000 cm^{-1}), and the C–H stretch range (3020–2800 cm^{-1}). The polished, smooth and rough surfaces are abbreviated P, S and R, respectively; the combined models considered are PS, SR and PSR. Each model is characterised by a wavenumber range (or ranges) and the surfaces used.

For each model, a cross validation is conducted and several statistics are calculated. The model complexity is chosen by Martens' method (See Section 3.3.2) with $s = 0.005$.[1] In addition to the RMSECV, the slope (m) and intercept (b) of the best-fit line between the true loadings and the predicted loadings are calculated. The significance level (p-value) required for the point (0, 1) to lie on the perimeter of the joint confidence region for (b, m) is also determined, as described in Appendix A.6.

[1] This small value of s discriminates only weakly against more complex models, so overfitting is a potential problem. However, the present work is concerned with the fit of the models to the available data rather than with actual predictive use of the models.

For the combined models, each of these statistics is calculated both for the complete pool of standards, and also for each surface individually. These statistics should reveal where predictions relating to one or another surface are significantly biased, even if the overall model results seem reasonable. The RMSECV(0) (see Section 3.3.2) is 0.51, 0.65 and 0.54 $\mu g\,cm^{-2}$ for the polished, smooth and rough standards, respectively (0.54 $\mu g\,cm^{-2}$ overall). Here, only the tabulated results are presented, since they describe well the trends in the cross-validation plots (which are available elsewhere [106, 142]).

Results for the single-surface models are presented in Table 8.2. They show that satisfactory models can be built for any surface using either or both of the wavenumber ranges. The P models generally have slightly lower RMSECV than the others, as do models utilising the fingreprint region. For P and S, combining the regions gives the best results.

Table 8.2: Cross-validation results for the single-surface models for SDS on stainless steel. P, S, and R indicate polished, smooth and rough, respectively. The fingerprint region is 1300–1000 cm^{-1} and the C–H region is 3020–2800 cm^{-1}. A_{opt} is the optimum rank determined by Martens' method with $s = 0.005$; RMSECV is in $\mu g\,cm^{-2}$; b and m are the slope and intercept of the best-fit line through the true and predicted loadings; p is the p-value for comparison of (b, m) with (0, 1). Values of $p < 5 \times 10^{-3}$ are reported as zero.

		Fingerprint	C–H	Both
P	A_{opt}	6	4	5
	RMSECV	0.07	0.11	0.06
	b	-0.02	-0.02	0.02
	m	1.01	1.01	0.98
	p	0.64	0.85	0.66
S	A_{opt}	5	11	4
	RMSECV	0.09	0.10	0.08
	b	0.01	-0.06	0.02
	m	0.99	1.04	0.98
	p	0.85	0.15	0.65
R	A_{opt}	7	6	5
	RMSECV	0.10	0.16	0.11
	b	0.04	0.09	0.04
	m	0.96	0.91	0.96
	p	0.51	0.40	0.60

Most of the interesting results are presented in Table 8.3, for the two-surface models. For the PS model using only the fingerprint region, the RMSECV is much higher than for either of the single-surface models. Isolating the results for each surface, it is seen that the slope for the polished predictions is much greater than one, while that for the smooth predictions is much less: this is precisely what would be expected on the basis of the intensity difference. The small p-values indicate that the bias is significant. When the C–H stretch region is used instead, the results are much better. The prediction

error is still somewhat higher than for the single-surface models, but the predictions for each surface are much less biased. Using both regions gives a model comparable in performance to the single-surface ones, but at somewhat higher optimal rank.

The SR model using the fingerprint region appears to perform only slightly worse than the single-surface models, but closer inspection reveals that the slope for the rough-surface predictions is substantially less than one. Adding the C–H region improves it somewhat, but, from the single-surface results, this region appears to be less useful for the rough-surface samples. These results are consistent with the slightly greater intensity of the smooth-surface samples and the similarity between the shapes of the spectra on the smooth and rough surfaces.

The PSR model results (Table 8.4) are generally consistent with the PS and SR results. None of the models provides satisfactory prediction for rough-surface samples: in all cases $b > 0$ and $m > 1$. The same patterns in prediction of polished and smooth samples are seen as for the PS model: the predictions are severely biased (high for P, and low for S) when the fingerprint region is used, but are essentially unbiased when both the fingerprint and C–H ranges are used. The RMSECV is substantially greater, however.

8.2.5 Conclusions

For SDS deposited on stainless steel having a particular surface finish, the loading can be effectively modelled as a function of the IRRAS, even if the surface is rough. However, increasing surface roughness decreases the intensity of the IRRAS. If the wavenumber range is restricted to the fingerprint region, this change in intensity precludes the incorporation of roughness into the model as an unquantified interferent without destroying the predictive ability of the model. However, the relative intensities of the C–H stretching bands also appear to depend on the roughness. This causes a variation of the shape of the spectrum with the surface roughness that allows a combined polished + smooth-surface model to perform well. It is not clear what causes the change in the relative intensities, though, and the phenomenon should not be relied upon until it can be explained or shown to be reproducible. It should be noted that in this particular case, the three surfaces can readily be distinguished from the spectra: if a polished coupon is used as the background, the spectra of the smooth and rough coupons will have strong sloping or curved baselines. These features could be used by an algorithm to select the appropriate model to apply, avoiding the reliance on the change in shape of the spectra.

Table 8.3: Cross-validation results for the two-surface (PS and SR) models for SDS on stainless steel. The label XY (X) indicates statistics calculated from the X-surface subset of the X- and Y-surface combined-model cross-validation results; the parenthesised values are where the optimal rank for the subset differs from that for the combined results. P, S, and R indicate polished, smooth and rough, respectively. The fingerprint region is 1300–1000 cm^{-1} and the C–H region is 3020–2800 cm^{-1}. A_{opt} is the optimum rank determined by Martens' method with $s = 0.005$; RMSECV is in $\mu g\,cm^{-2}$; b and m are the slope and intercept of the best-fit line through the true and predicted loadings; p is the p-value for comparison of (b, m) with (0, 1). Values of $p < 5 \times 10^{-3}$ are reported as zero.

		Fingerprint	C–H	Both
PS (all)	A_{opt}	7	10	8
	RMSECV	0.25	0.14	0.09
	b	0.03	-0.08	0.02
	m	0.94	1.05	0.98
	p	0.62	0.05	0.59
PS (P)	A_{opt}	7	10 (4)	8
	RMSECV	0.21	0.15 (0.12)	0.09
	b	-0.04	-0.08 (-0.09)	0.06
	m	1.14	1.04 (1.05)	0.96
	p	0.04	0.42 (0.06)	0.22
PS (S)	A_{opt}	7 (6)	10 (11)	8 (7)
	RMSECV	0.27 (0.25)	0.13 (0.12)	0.09 (0.09)
	b	0.05 (0.05)	-0.09 (-0.07)	0 (0.02)
	m	0.83 (0.80)	1.05 (1.04)	0.98 (0.95)
	p	0.02 (0)	0.09 (0.17)	0.51 (0.04)
SR (all)	A_{opt}	7	8	9
	RMSECV	0.12	0.20	0.10
	b	0.04	-0.04	0.04
	m	0.96	1.00	0.96
	p	0.33	0.27	0.34
SR (S)	A_{opt}	7 (5)	8 (6)	9 (5)
	RMSECV	0.12 (0.12)	0.20 (0.19)	0.10 (0.10)
	b	0 (-0.02)	-0.13 (-0.18)	-0.01 (-0.07)
	m	1.02 (1.04)	1.11 (1.15)	1.02 (1.09)
	p	0.78 (0.51)	0.12 (0.01)	0.82 (0)
SR (R)	A_{opt}	7	8	9 (10)
	RMSECV	0.11	0.20	0.11 (0.10)
	b	0.13	0.12	0.12 (0.07)
	m	0.87	0.84	0.89 (0.93)
	p	0	0.05	0.02 (0.15)

Table 8.4: Cross-validation results for the three-surface model for SDS on stainless steel. The label PSR (X) indicates statistics calculated from the X-surface subset of the three-surface combined-model cross-validation results; the parenthesised values are where the optimal rank for the subset differs from that for the combined results. P, S, and R indicate polished, smooth and rough, respectively. The fingerprint region is 1300–1000 cm^{-1} and the C–H region is 3020–2800 cm^{-1}. A_{opt} is the optimum rank determined by Martens' method with $s = 0.005$; RMSECV is in $\mu g\,cm^{-2}$; b and m are the slope and intercept of the best-fit line through the true and predicted loadings; p is the p-value for comparison of (b, m) with (0, 1). Values of $p < 5 \times 10^{-3}$ are reported as zero.

		Fingerprint	C–H	Both
PSR (all)	A_{opt}	9	6	9
	RMSECV	0.27	0.25	0.19
	b	0.09	0.03	0.13
	m	0.90	0.93	0.87
	p	0.18	0.23	0
PSR (P)	A_{opt}	9 (2)	6 (5)	9 (3)
	RMSECV	0.36 (0.23)	0.17 (0.15)	0.17 (0.12)
	b	-0.10 (-0.03)	0.06 (0.11)	0.06 (0.18)
	m	1.22 (1.14)	1.01 (0.94)	0.99 (0.85)
	p	0.17 (0.07)	0.11 (0.11)	0.31 (0)
PSR (S)	A_{opt}	9 (10)	6 (6)	9 (8)
	RMSECV	0.23 (0.22)	0.23	0.13 (0.12)
	b	0.05 (0.03)	-0.22	-0.02 (-0.02)
	m	0.86 (0.91)	1.19	1.02 (1.03)
	p	0.04 (0.27)	0.01	0.84 (0.74)
PSR (R)	A_{opt}	9 (11)	6 (7)	9 (12)
	RMSECV	0.20 (0.19)	0.31 (0.29)	0.25 (0.23)
	b	0.31 (0.21)	0.41 (0.35)	0.45 (0.37)
	m	0.69 (0.77)	0.51 (0.57)	0.56 (0.64)
	p	0 (0)	0 (0)	0 (0)

8.3 Acetaminophen

Sixty-nine samples (24 on polished and smooth and 21 on rough substrates) were prepared by the spray method, as described in Section 4.3.2. Eight spectra were then measured from each sample, as described in Section 4.3.4. The acetaminophen loadings were determined by the UV colorimetric method described in Section 4.3.3.

8.3.1 Scanning electron micrographs

Figure 8.6 shows scanning electron micrographs of the three surfaces loaded with ~2 μg cm^{-2} of acetaminophen. The left-hand images are viewed normal to the surface at a working distance of 8 mm; those on the right are inclined at the IRRAS angle of $\theta \approx 80°$ with working distances of 18 ± 5 mm. The differences between the surface finishes themselves are evident, as described above in Section 8.2.1.

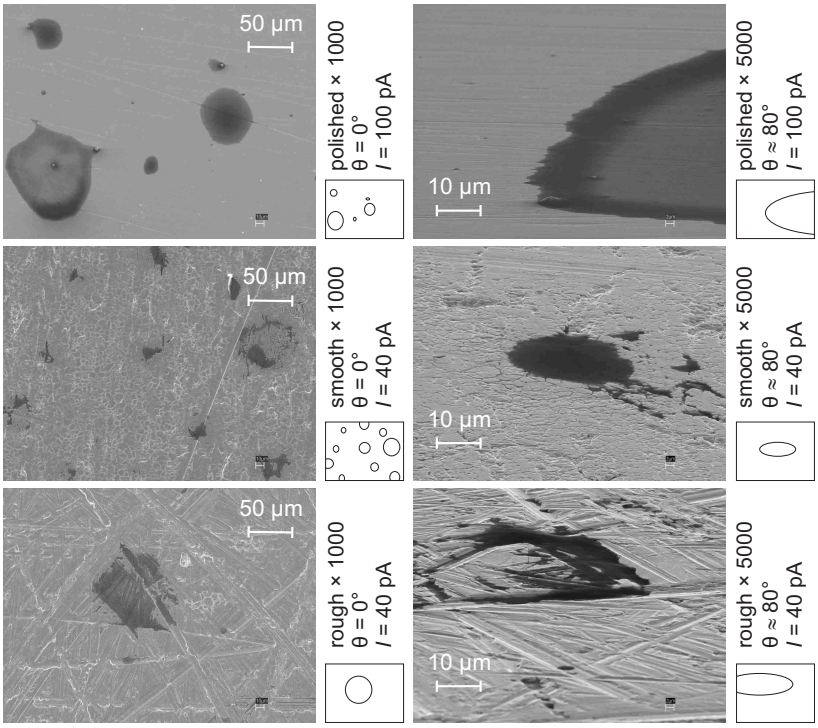

Figure 8.6: SEM images from three finishes of stainless steel after loading with ~2 μg cm^{-2} of acetaminophen by the spray technique. The schematic to the right of each photo indicates the location of the acetaminophen spots. The images in the left column were taken perpendicular to the surface at ×1000 magnification. Those on the right were taken at ~80° to the normal and at ×5000 magnification. The top-row images are for the polished finish, the middle row for the smooth finish and the bottom row for the rough finish. Beam currents are indicated by the parameter *I*.

SDS and acetaminophen are compared in Figure 8.7, in which lower-magnification images of the two analytes on the polished surface are shown. The thickness of the acetaminophen spots can be

estimated in the same way as for SDS. Approximately 25 % of the surface is covered, and the density of acetaminophen is ~1.3 g cm^{-3} [143], so the average thickness of the spots is approximately 60 nm, about three times thicker than for SDS (but with roughly twice the loading). From Figure 8.6, the spots of acetaminophen have domed profiles that are particularly evident for the images of the larger spots on the polished surface. Acetaminophen also appears to show a greater propensity than SDS to fill the surface scratches and grooves, especially on the smooth and rough surfaces. If, as suggested above, the reduction in IRRAS intensity with increasing surface roughness is due to material residing in grooves that are inaccessible to the infrared radiation, then this last observation leads to the prediction that the reduction in intensity will be greater for acetaminophen than for SDS, because a larger proportion is present in the inaccesible regions.

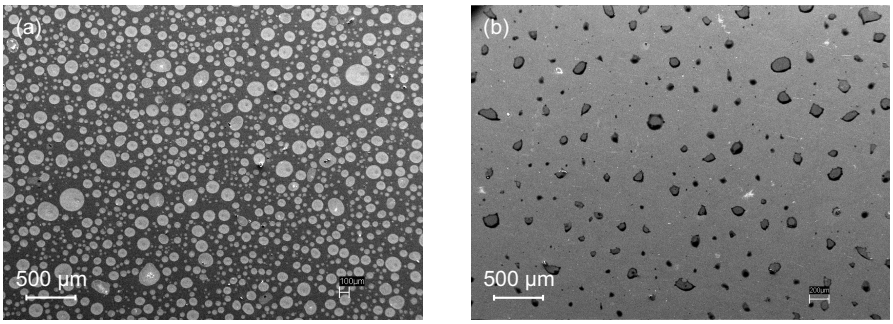

Figure 8.7: Low-magnification SEM images of (a) ~1 µg cm^{-2} SDS and (b) ~2 µg cm^{-2} acetaminophen on polished stainless steel. The analytes were deposited by the spray method. For SDS, the beam voltage was 10 kV and the current was 3 nA; for acetaminophen, the beam voltage was 10 kV and the current was 0.1 nA.

8.3.2 Spectra

IRRA spectra for the finger-print region of acetaminophen at loadings of ~0.4 and ~2.6 µg cm^{-2} on each finish of stainless steel are given in Figure 8.8. To simplify comparison, they have been modified in several ways from those used in the chemometric modelling. First, the signal-to-noise ratios have been improved by averaging over all the spectra for each sample. Second, strongly curved baselines (due to variations in the surface roughness [106]) have been corrected by subtraction of a low-order polynomial. And third, absorption bands due to atmospheric water vapour have been subtracted by using a reference spectrum scaled by a derivative-minimisation algorithm.

The most striking point of comparison between these spectra is that the intensity at a loading of ~0.4 µg cm^{-2} is much greater on the polished surface than on either of the other surfaces: in fact, it is more comparable to the intensities of ~2.6 µg cm^{-2} spectra on the smooth and rough surfaces. It is also clear that the relationship between intensity and API loading is far from linear for the polished surface.

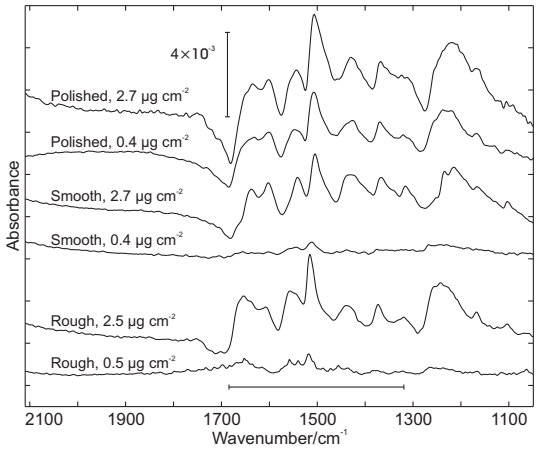

Figure 8.8: Representative IRRAS of acetaminophen on stainless steel. Spectra of two loadings on each of the surfaces are plotted, as indicated on the figure. Baseline correction is by polynomial fitting and water vapour has been reduced by subtraction of a reference spectrum. For clarity, the spectra have been offset along the ordinate; the intensity scale is indicated. The wavenumber range used for the PLS regression is indicated by the horizontal bar at the bottom of the figure.

More subtly, there are shifts of the wavenumbers of the maxima that depend not only on the roughness of the surface but also on loadings. A slight shift to the red is seen as the loading increases. Based on the calculations in Chapter 5 and the estimated film thickness, nonlinearity would not be expected for these loadings. However, the estimated thickness does not take into account the possibility that the thickness of the spots is not constant. Since acetaminophen is a crystalline material, it seems very likely that the larger spots are also thicker, and this may be the cause of the nonlinearity.

8.3.3 Band integrals

To get a better impression of the intensity behaviour, the estimated area under the ~1520 cm^{-1} band is plotted as a univariate measure of intensity against loading in Figure 8.9. (The choice of other features and variations in the method of integration gave similar results.) Despite the scatter of the results, it is clear that the different surfaces show very different IRRA responses. Nonlinearity is evident for all, but is particularly acute for the polished surface at loadings greater than ~1.0 µg cm^{-2}. Furthermore, in the ranges where the response is approximately linear (indicated by the straight lines in Figure 8.9), the slope depends strongly on the surface finish, with the sensitivity decreasing with increasing roughness. Comparing Figure 8.9 with the SDS integrals plotted in Figure 8.4, it can be seen that the dependence of the slope on the surface finish is much more acute for acetaminophen, as predicted on the basis of the SEM images.

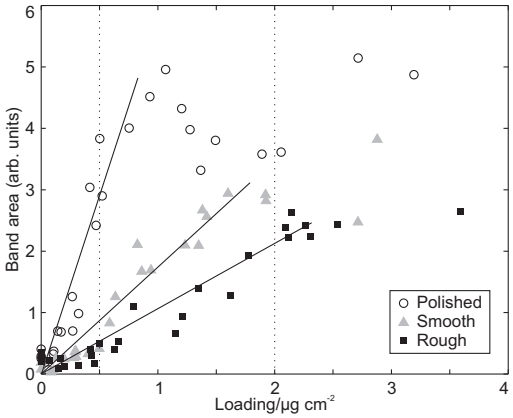

Figure 8.9: Integral of the ~1520 cm^{-1} IRRAS band of acetaminophen on stainless steel as a function of loading. The integral was calculated by the trapezium method after subtraction of a reference water vapor spectrum and using a straight baseline that intersects the spectrum at each end of the integration range. The three solid lines indicate the regions over which the band integral is approximately proportional to the loading. The cut-off points for the low- and high-loading ranges are indicated by the vertical dotted lines at 0.5 and 2.0 μg cm^{-2}, respectively.

8.3.4 PLS modelling

In light of the complicated nonlinearities, chemometric calibrations were conducted using twelve subsets of data over two loadings ranges (indicated in Figure 8.9): a higher-loading range of 0–2.0 μg cm^{-2}, where the upper limit approaches the point at which contamination is visible to the naked eye; and a lower-loading range of 0–0.5 μg cm^{-2}, which is more relevant to real situations of pharmaceutical cleaning validation. For each loading range, the three individual surface finishes (P, S and R) were considered separately, along with three combinations: PS, SR, and PSR. The purpose of this investigation is to obtain an overview of the effects of the surface roughness, so somewhat less attention than is usual will be paid to the optimisation steps.

The wavenumber range used was 1690–1320 cm^{-1}. The optimal number of PLS factors for each model was selected using Martens' method (see Section 3.3.2), with $s = 0.02$. Model quality was evaluated as the RMSECV obtained from leave-one-out cross validations in which the eight spectra from each sample were treated together as a unit.

Low-loading range

There are 14 polished-, 12 smooth- and 13 rough-surface samples in the target loading range (see Figure 8.9). RMSECV(0) is 0.16, 0.16 and 0.20 μg cm^{-2} for the polished-, smooth-, and rough-surface samples respectively, and 0.17 μg cm^{-2} overall.

The RMSECV curves are plotted in Figure 8.10. The cross-validation statistics for the six models

at optimal rank are presented in Table 8.5, and the corresponding plots of the cross-validation predicted loadings are given in Figure 8.11.

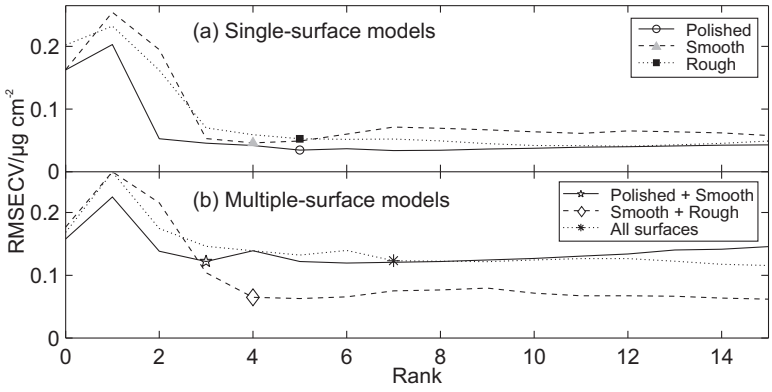

Figure 8.10: RMSECV vs rank for the low-loading ($< 0.5\,\mu g\,cm^{-2}$) acetaminophen on stainless steel datasets. (a) Individual surfaces. (b) Combinations of surfaces. The optimal rank for each model is indicated by a symbol.

All the single-surface models perform adequately: the RMSECV is fairly low and there is no evidence of bias (the p values are large). There are some potentially outlying samples for the polished and rough surfaces, which would be investigated in a more thorough optimisation.

The situation is quite different for the combined-surface models, however. The RMSECV for the smooth samples in the PS model is actually greater than RMSECV(0), and severe bias is evident in the predicted-loadings plot. Increasing the rank does not improve the results. The performance for the polished-surface samples in this model is much better, but still significantly poorer than in the polished-only model.

The RS model stands out from the others in the RMSECV plot (Figure 8.10). It is the only combined-surface model to perform comparably to the single-surface models. There is some evidence of bias in the form of over-estimation at low loadings and under-estimation at high loadings, but this is only statistically significant at the $\alpha = 0.05$ level for the smooth-surface samples.

When all three surfaces are combined, the results are uniformly poor. The RMSECV values are only slightly lower than the RMSECV(0) values.

High-loading range

When all samples with loadings $<2.0\,\mu g\,cm^{-2}$ are considered, RMSECV(0) for the polished-, smooth- and rough-surface datasets is 0.57, 0.63 and 0.57 $\mu g\,cm^{-2}$, respectively, and 0.58 $\mu g\,cm^{-2}$ overall.

The RMSECV vs rank plots are presented in Figure 8.12a (single-surface models) and Figure 8.12b (combined-surface models). These plots are qualitatively very similar to the low-loading ones, but the

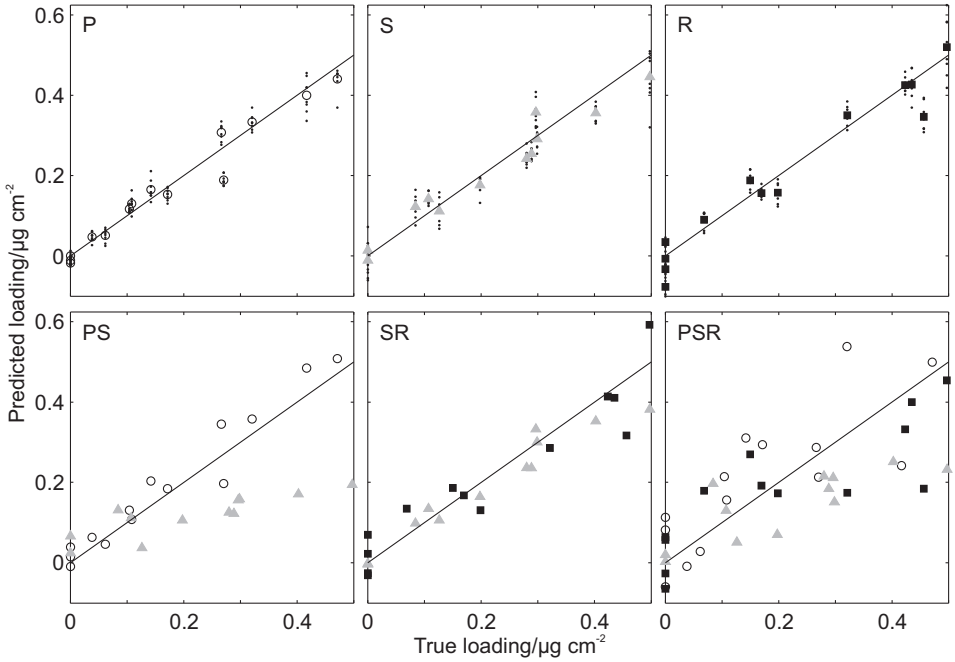

Figure 8.11: Cross-validation predictions for the lower-loading (<0.5 µg cm^{-2}) acetaminophen on stainless steel datasets. P, S, R: Single-surface models with predictions shown for all individual spectra. PS, SR, PSR: Combined-surface models with averaged (per-sample) predictions shown.

Table 8.5: Cross-validation statistics for acetaminophen on stainless steel (loadings <0.5 µg cm^{-2}). RMSECV is calculated from per-spectrum predictions, and is in µg cm^{-2}; b and m are the slope and intercept of the best-fit line through the true and (averaged) predicted loadings; p is the p-value for comparison of (b, m) with (0, 1). Values of $p < 5 \times 10^{-3}$ are reported as zero.

	A_{opt}	RMSECV	b	m	p
P	5	0.035	0	0.96	0.61
S	4	0.046	0.02	0.88	0.19
R	5	0.053	-0.01	0.99	0.69
PS (all)	3	0.120	0.03	0.66	0.01
PS (P)	3	0.058	0.01	1.07	0.12
PS (S)	3 (1)	0.17 (0.25)	0.06 (0.07)	0.29 (-0.11)	0 (0)
SR (all)	4	0.065	0.01	0.88	0.10
SR (S)	4	0.056	0.02	0.83	0.02
SR (R)	4	0.072	0.01	0.92	0.69
PSR (all)	7	0.120	0.05	0.65	0
PSR (P)	7 (2)	0.12 (0.11)	0.06 (0.07)	0.90 (0.68)	0.40 (0.02)
PSR (S)	7	0.130	0.05	0.45	0
PSR (R)	7 (4)	0.12 (0.12)	0.04 (0.05)	0.66 (0.52)	0.06 (0)

RMSECV is generally higher. The predicted loadings are plotted in Figure 8.13 and the cross-validation statistics are listed in Table 8.6. The PS and PSR models perform very poorly, while the performance of the SR model is comparable to that of the single-surface models (all of which perform similarly). If the plots in Figure 8.13 are examined closely, it can be seen that, in every case, the accuracy of the predictions in the low-loading range (0–0.5 $\mu g\,cm^{-2}$) is actually quite poor when the full range is used for calibration. This is a consequence of the nonlinearity, and is in contrast to the situation with SDS and also with the results for acetaminophen on glass (Chapters 6 and 7).

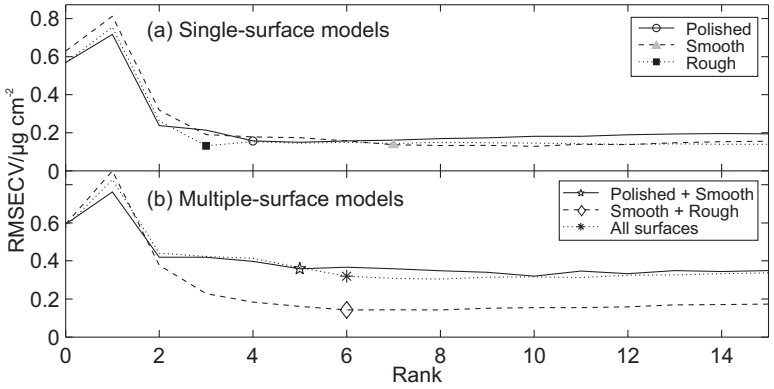

Figure 8.12: RMSECV vs rank for the higher-loading (<2 $\mu g\,cm^{-2}$) acetaminophen on stainless steel datasets. (a) Individual surfaces. (b) Combinations of surfaces. The optimal rank for each model is indicated by a symbol.

While fitting a straight line through the predictions in Figure 8.13 does not indicate any bias (the p-values, listed in Table 8.6, are all large), an interesting pattern can be seen in all of the plots. The predictions are too small at low loadings, too large at intermediate loadings, then become too small again at higher loadings. A similar pattern of scatter about the best-fit line can also be observed in the band-integral plots (Figure 8.9). Increasing the rank does not ameliorate this effect.

This presence of structure (with respect to the loading) in the residuals suggests that nonlinear regression might provide some improvement. Quadratic PLS[2] was applied to these data, and the results are compared with those of the linear PLS model in Figure 8.14 (in which the cross-validation residuals are plotted) and Table 8.6. The most dramatic difference is the reduction in the optimal rank for the smooth-surface model: however, the RMSECV is somewhat larger, so it is difficult to say whether the reduction in rank is due to the quadratic inner relationship or simply to random variation increasing the RMSECV at higher rank. In general, the structure in the residuals is very slightly reduced, but still evident. The cross-validation statistics in Table 8.6, almost without exception, favour the linear

[2] In polynomial PLS each latent variable is calculated in the standard way, but the "inner relationship" between the X scores $\mathbf{T}$ and the Y scores $\mathbf{U}$ is modelled with a polynomial instead of a straight line. For this work, the polynomial PLS function from the PLS Toolbox [146] for MATLAB was used.

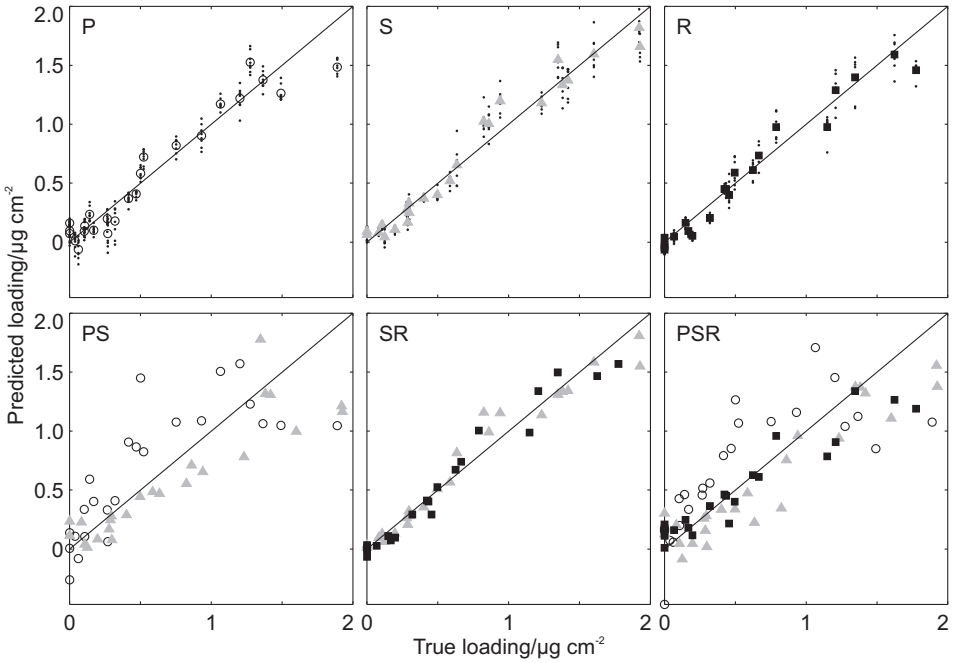

Figure 8.13: Cross-validation predictions for the higher-loading (<2.0 µg cm^{-2}) acetaminophen on stainless steel datasets. P, S, R: Single-surface models with predictions shown for all individual spectra. PS, SR, PSR: Combined-surface models with averaged (per-sample) predictions shown.

PLS model. Cubic and quartic PLS models similarly failed to provide substantial improvements. This nonlinearity is presumably due to some optical effect related to the thickness of the film, but since these films are so poorly characterised (as they would be in a "cleaned" pharmaceutical reactor), meaningful comparison with theory is difficult.

8.3.5 Conclusions

Nonlinearity

All surfaces are subject to severe nonlinearity in the IRRAS at loadings $\gtrsim$0.5 µg cm^{-2}, but the effect is particularly strong for the polished surface. This nonlinearity is not observed over the same loading range for SDS. Polynomial PLS does not seem to offer a significant improvement over linear PLS for this system. Accurate prediction of low-loading samples appears to require a model constructed from only low-loading standards. In practice, a two-model system could be useful. In the first step, the loading is predicted approximately from the full-range model; if the estimated loading is small, its loading is predicted more precisely using a low-loading model.

Interestingly, this dramatic nonlinearity was not encountered in an earlier study [133] in which an unspecified API was deposited on aluminium coupons, nor in the work with SDS [142]—the data discussed in Section 8.2 above are only a subset of the SDS spectra, and from the full data set, it can be

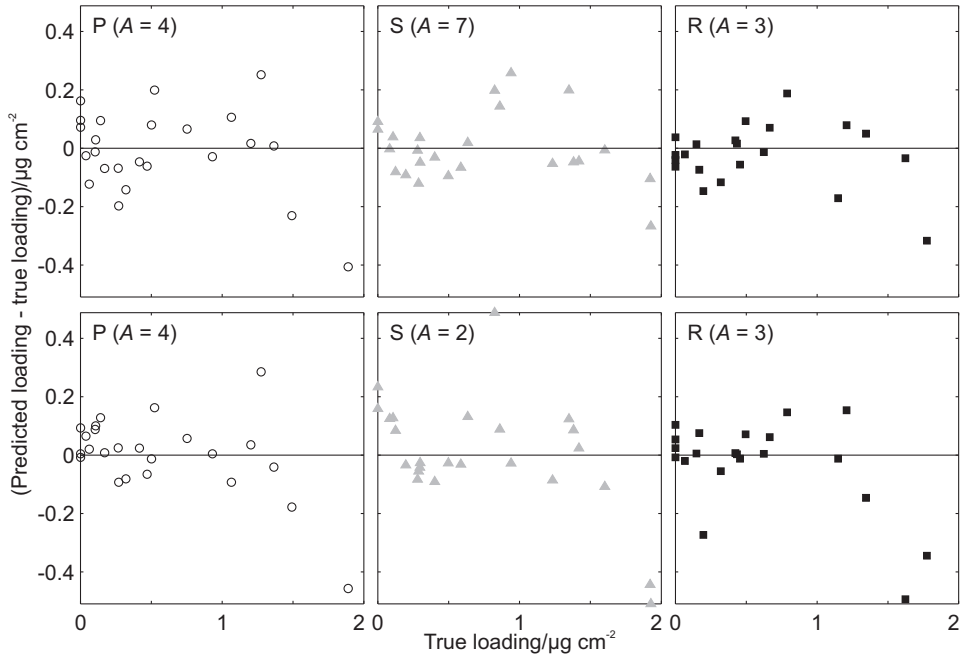

Figure 8.14: Cross-validation residuals (predicted loading minus true loading) for the higher-loading (<2.0 μg cm^{-2}) single-surface models. The top row of plots is for linear PLS; the second row is for quadratic PLS. The optimal ranks (A) are indicated.

Table 8.6: Cross-validation statistics for acetaminophen on stainless steel (loadings 0–2 μg cm^{-2}). RMSECV is calculated from per-spectrum predictions, and is in μg cm^{-2}; b and m are the slope and intercept of the best-fit line through the true and (averaged) predicted loadings; p is the p-value for comparison of (b, m) with (0, 1). Values of $p < 5 \times 10^{-3}$ are reported as zero. Q-PLS denotes a quadratic PLS model.

	A_{opt}	RMSECV	b	m	p
P	4	0.16	0.03	0.93	0.38
S	7	0.14	0.02	0.97	0.76
R	3	0.13	0	0.96	0.44
P (Q-PLS)	4	0.15	0.07	0.89	0.07
S (Q-PLS)	2	0.21	0.12	0.84	0.04
R (Q-PLS)	3	0.18	0.06	0.85	0.03
PS (all)	5	0.36	0.16	0.72	0
PS (P)	5 (6)	0.38 (0.35)	0.23 (0.24)	0.77 (0.81)	0.10 (0.04)
PS (S)	5	0.33	0.05	0.72	0
SR (all)	6	0.14	0.02	0.96	0.29
SR (S)	6	0.15	0.05	0.93	0.29
SR (R)	6 (5)	0.13 (0.13)	-0.01 (-0.04)	0.98 (0.99)	0.67 (0.16)
PSR (all)	6	0.32	0.14	0.73	0
PSR (P)	6 (7)	0.41 (0.38)	0.29 (0.28)	0.71 (0.76)	0.03 (0.02)
PSR (S)	6	0.28	-0.01	0.81	0
PSR (R)	6	0.22	0.11	0.71	0

seen that the linearity extends to much higher loadings, perhaps ~16 $\mu g\,cm^{-2}$ or more [106]. A possible reason for the nonlinearity is the tendency of acetaminophen to form larger crystals rather than to spread out over the surface. Larger crystals could cause nonlinearity either through the reflectance effects described in Chapters 2 and 5 or through scattering. A more thorough characterisation of the standards was attempted [106], using profilometry, but found to be impracticable because of the baseline roughness of even the polished surfaces. Perhaps a better approach would be to prepare samples by the spray method using a very smooth metallic substrate, such as a metal film evaporated onto a glass slide, and to characterise these by profilometry or atomic force microscopy.

Effects of surface roughness

Generally, IRRAS intensity has been shown to decrease as the surface roughness increases. This effect is attributed to a portion of the analyte residing in recessed regions of the surface that cannot be reached by the infrared-wavelength radiation. The extent of the decrease, for the surface finishes studied here, is much greater for acetaminophen than for SDS, an observation consistent with the greater tendency of acetaminophen to fill the grooves and crevices of the surface (as revealed by the SEM studies). In principle, the decrease in intensity is a significant obstacle to constructing a single chemometric model that can give the correct loading for different grades of surface finish. This is demonstrated by the poor performance of the mixed-surface models for acetaminophen (with the exception of the smooth + rough model, for which two surfaces the IRRAS intensity is similar). For SDS, however, a change in shape in the C–H region of the spectrum between polished and unpolished surfaces permits a mixed-surface calibration.

A limitation of this study is that the differences in surface finish between P, S and R are large compared to what would be encountered in practice. A useful extension of this work would be to compare different grades of "polished" finish, preferably with a quantitative measure of the roughness of the substrate.

Chapter 9

Conclusions and future work

9.1 General conclusions

The work described in this thesis and in recent publications [133, 134, 135, 137, 142] shows that fibre-optic grazing-angle IRRAS, when combined with factor-based inverse regression methods, such as PLS, is applicable to some problems in cleaning validation. These studies have demonstrated that this combination of methods is both selective and sensitive. IRRAS has been shown to be applicable to the determination of residues on metallic and glass substrates; preliminary investigations into extending these results to some organic-polymer substrates are discussed below.

The potential for selectivity was shown in Chapter 7, where calibrations were achieved for aspirin and acetaminophen simultaneously, and also in Refs 135 and 134, which describe calibrations for APIs in the presence of a cleaning compound and some excipients, respectively. Atmospheric water vapour is pervasive, but its effects can be compensated either by the methods described in Appendix B or by allowing PLS to model it implicitly.

From the present work, it is somewhat difficult to get an accurate estimate of the sensitivity of the method. All the reported RMSE values are biased high by the presence of errors in the reference loading values, due mostly to heterogeneity of the validation standards. For all of the studies carried out in the project, an RMSE of ~0.1 $\mu g\,cm^{-2}$ or less was readily obtainable. This error corresponds to an approximate detection limit (as defined in Chapter 7) of $L_D \approx 0.3\ \mu g\,cm^{-2}$. From the cleaning validation examples in the literature, this is a useful L_D: it is lower than both the commonly accepted visible residue limit of 1–4 $\mu g\,cm^{-2}$ [7] and several published examples of acceptable residual limits [18, 19]. Furthermore, because of the upward bias in the RMSE estimate, the true L_D is somewhat lower than this value; further work to address this discrepancy is suggested below. From Refs 13, 18 and 19, it appears that, where a UV detector can be used, HPLC provides lower detection limits. However, IR-

RAS can be effective for compounds without UV chromophores, such as SDS, and may be particularly useful for measuring residuals from cleaning compounds.

While IRRAS is generally more intense with metallic substrates than with dielectrics, the RMSECV values (and hence, detection limits) presented here for acetaminophen on polished stainless steel and on glass are similar. There are several reasons for this. The most important is that a significant portion of the RMSECV is due to the errors in the reference loading values (arising mostly from heterogeneity of the standards). Because of this, the SNR of the spectra has only a small effect on the apparent prediction error: see, for example, Section 6.6, where a fivefold increase in the noise level led to only an 11 % increase in the RMSECV. Another factor is the nonlinearity that was encountered with the steel substrate but not with the glass.

The most dramatic advantage of IRRAS over HPLC-based methods is the rapidity of the measurements. Developing a calibration model can be fairly time-consuming, taking on the order of a week or two, but this only needs to be done once for a given analyte/substrate combination. Each *in situ* IRRAS measurement of a new sample takes only a few seconds.

9.2 Substrate considerations

9.2.1 Model validity for several metals

A question that could be of practical importance is whether a model constructed using one kind of metallic substrate could be used with another kind of metallic substrate. The IRRAS intensity depends on the optical constants of the metal substrate. Very highly reflective metals, such as aluminium and gold, give very similar results, but iron, at higher wavenumbers, gives rather different results (lower intensity). It may be possible to use calibration-transfer techniques (see Section 9.3.3 below) to modify a model relevant to one metallic substrate to be valid for another. A scenario more likely to be of importance is transfer from one grade of stainless steel to another. The different grades have different percentages of the constituent metals, so will have different optical properties. It is likely that these differences will be small compared to the contrast between, say, aluminium and iron, and it is probable that the calibration transfer would be quite straightforward.

9.2.2 Surface finish

Surface finish has been shown to be very important in the case of stainless steel. As discussed in Chapter 8, the IRRAS intensity is significantly greater on highly polished surfaces. For SDS, the shape of the IRRAS spectrum exhibits an interesting dependence on the surface roughness, which allows a

combined-surface model to be built successfully, despite the intensity difference. For acetaminophen, however, there was no such change in shape, and models combining polished-surface data with spectra from rougher surfaces were unsuccessful. The work described here involved much greater differences in roughness than would likely be encountered in practice, and more work is required both to identify the extent of roughness required to significantly affect the intensity and to investigate methods for dealing with this problem.

Attempting to build a single PLS model to cope with surfaces of varying roughness may not be the best approach. Since the roughness also affects the slope and curvature of the baseline (because of the wavelength dependence of the scattering of light from the surface), it may be possible to use the baseline to determine the roughness. Then, either an appropriate model could be selected from several possibilities, or calibration-transfer techniques (see Section 9.3.3 below) could be used to correct the prediction from one general model.

9.2.3 Nonlinearity; Brewster-angle measurements

The reason for the severe nonlinearity encountered in Chapter 8 for acetaminophen on stainless steel is not clear. It may be due to the film being (locally) quite thick, introducing nonlinearity and distortion of the types discussed in Chapter 5. If this is the case, using *p*-polarised light incident near the Brewster angle for the film should alleviate the nonlinearity, at the cost of some sensitivity. As the incidence angle is increased towards grazing incidence from the Brewster angle, the sensitivity increases and the thickness threshold for linear response decreases, so some intermediate angle may provide the best compromise between sensitivity for very thin films and linearity for thicker ones. In any case, the variable angle probe should be modified to incorporate a polariser so that these ideas can be tested.

It is important to note that this nonlinearity was observed for acetaminophen, but not for SDS: there would be no advantage to using Brewster-angle measurements for SDS. Future work should expand the range of analytes considered and evaluate the possibility of a compact probe with adjustable incidence angle.

9.2.4 Other substrates of interest (preliminary studies)

The work presented in this thesis focused on glass and metallic substrates, as these are of particular importance in the pharmaceutical industry. However, other materials are also commonly used, such as silicone or ethylene-propylene-diene-monomer (EPDM) rubber for gaskets and seals, and poly(tetrafluoroethene) (PTFE) for reactor linings. In addition, if applications outside of the pharmaceutical industry are identified, they will have their own substrates of interest. Preliminary studies were conducted in

this group for acetaminophen on EPDM, PTFE and poly(methylmethacrylate) (PMMA) substrates.

9.2.5 EPDM rubber

EPDM is a terpolymer of ethylene, propylene and a non-conjugated diene such as dicyclopentadiene or ethylidene norbornene. It has good resistance to acids, bases and polar solvents, and is used for hoses and seals in the pharmaceutical and food industries. A reflectance spectrum of an EPDM sample is plotted in Figure 9.1 along with an RA spectrum of acetaminophen on an EPDM substrate. The substrate has C–H stretching bands around 2900 cm^{-1} and many bands at longer wavelength due to alkene stretching modes various bending modes. Peaks in the RA spectrum that correspond to acetaminophen absorbance are indicated by dashed lines; those that correspond to the substrate are indicated by dotted lines (to the red of 1200 cm^{-1} both species have bands and the assignment is difficult).

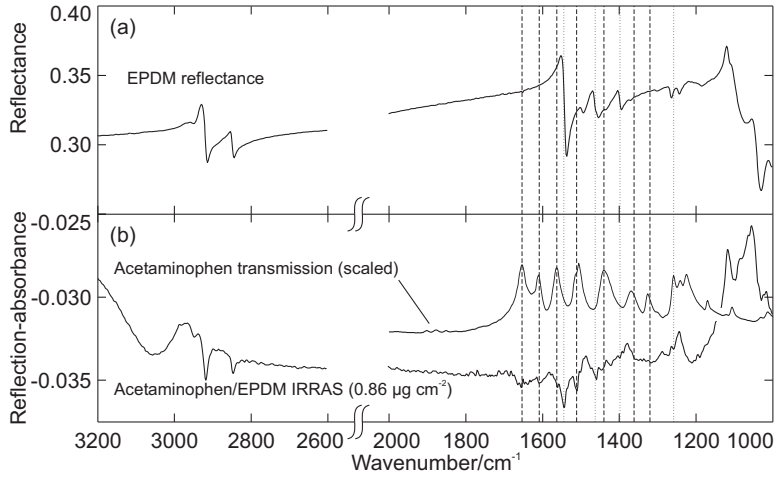

Figure 9.1: (a) Reflectance spectrum of an EPDM rubber sample relative to polished stainless steel. (b) IRRAS (80°) of 0.86 µg cm^{-2} of acetaminophen on EPDM, compared with a transmission spectrum of acetaminophen. Dotted lines indicate RA bands corresponding to substrate reflectance features; dashed lines indicate RA bands corresponding to acetaminophen absorbance bands.

Results from a preliminary calibration study for acetaminophen on EPDM are presented in Figure 9.2. Standards were prepared by the smear method from ethanol solution. Clearly, the quality of the model is substantially inferior to that of the others presented in this thesis. Comparing the RMSECV values to RMSECV(0), though, reveals that the model does have some predictive ability. From Figure 9.2b, it appears that the calibration is linear up to about 3 µg cm^{-2}. Repeating the calibration with the high-loading samples removed (Figure 9.2c) improved the results somewhat, but they are still rather poor.

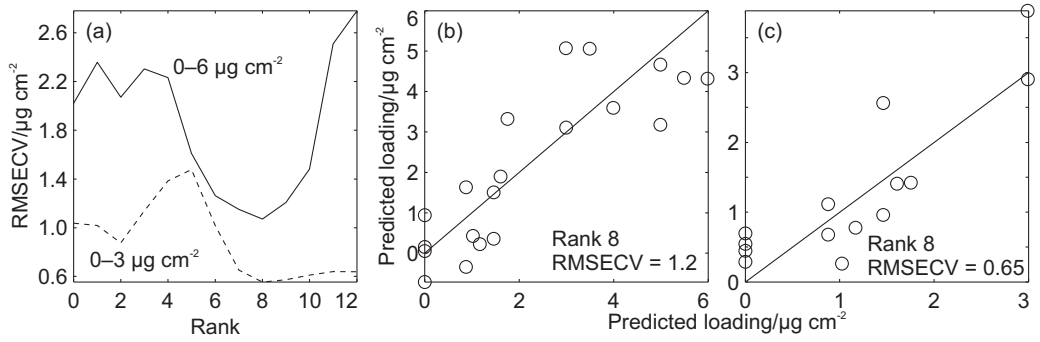

Figure 9.2: Cross-validation results for acetaminophen on EPDM. (a) RMSECV as a function of rank. (b) Predicted vs true loadings for the full-range model at optimal rank. (c) Predicted vs true loadings for the 0–3 µg cm^{-2} model at optimal rank. RMSECV is in µg cm^{-2}.

9.2.6 Poly(methylmethacrylate)

While poly(methylmethacrylate) (PMMA) is not widely used in the pharmaceutical industry, it is an extremely common material elsewhere. Physically, it resembles glass, being rigid, flat and smooth. As an organic material, however, it has quite different optical properties in the infrared [147].

Joshua Lehr, a BSc honours student in this research group, conducted a calibration experiment for acetaminophen on PMMA coupons using the methods developed in this thesis [148]. A reflectance spectrum of the substrate and some typical reflection-absorption spectra are plotted in Figure 9.3. The dominant band in the substrate reflectance spectrum is due to the carbonyl stretching mode at ~1740 cm^{-1}; this mode absorbs very strongly ($k \approx 0.36$) and has a correspondingly intense dispersion in the refractive index [147]. To the red of this band there is a region free of substrate bands (~1700–1500 cm^{-1}) in which several bands due to acetaminophen can be seen in the IRRAS spectra.

Cross-validation results are plotted in Figures 9.4a and b. This calibration is clearly better than that for EPDM, evidenced by the smaller RMSECV for a comparable range of loadings (and the stability of the RMSECV at higher ranks) and the better agreement between the predicted and true loadings. There are several possible reasons for the poor performance of the EPDM model. The most obvious explanation is that most of the acetaminophen bands are severely overlapped by bands due to the substrate, which change in intensity and shape in a somewhat unpredictable way as the acetaminophen loading is varied. Other possibilities include the extent of scattering from the surface: comparing the reflectance spectra in Figures 9.1 and 9.3, the reflectance of PMMA is seen to be much higher, even though the two materials have similar refractive indices. Finally, the EPDM coupons are not as perfectly flat as the PMMA, which could result in variation in the incidence angle between measurements. This issue could be resolved by attaching the EPDM coupons to a flat substrate.

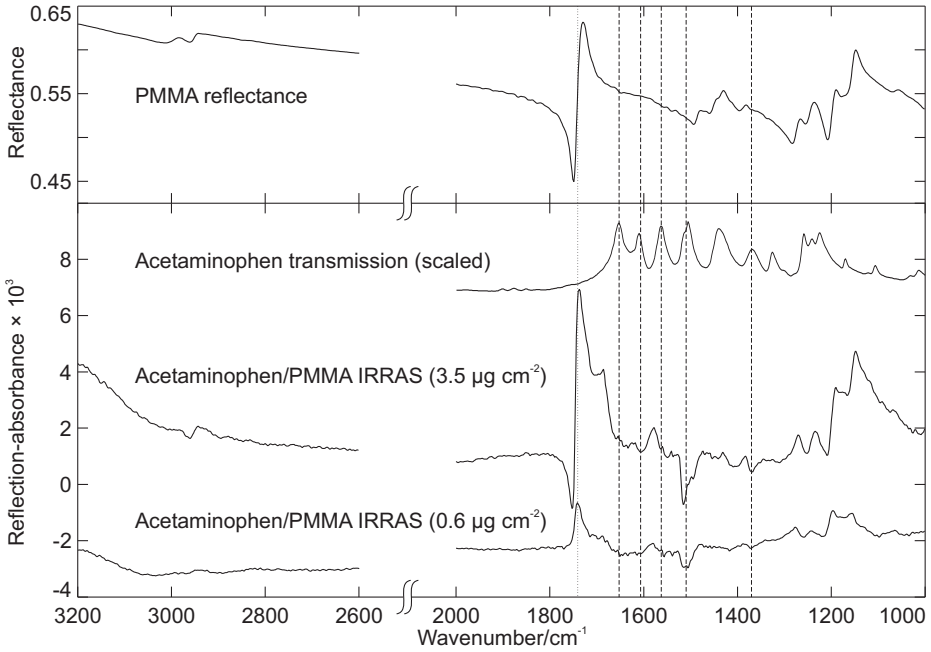

Figure 9.3: Spectra pertaining to IRRAS of acetaminophen on PMMA. From top: PMMA reflectance, acetaminophen absorbance, acetaminophen/PMMA IRRAS (3.5 and 0.6 $\mu g\,cm^{-2}$, from Ref. 148). The dashed lines indicate IRRAS bands attributed to the film; dotted lines indicate IRRAS bands attributed to the substrate.

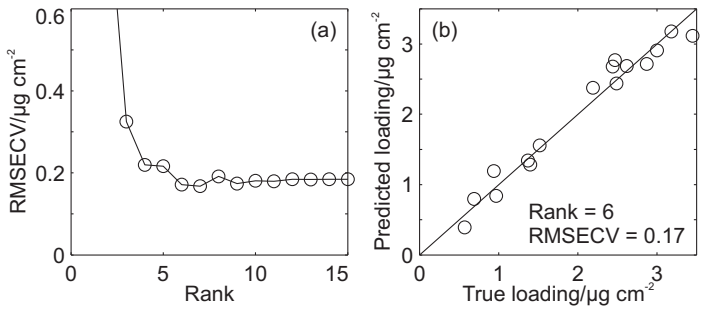

Figure 9.4: Cross-validation results for acetaminophen on PMMA. (a) RMSECV vs rank. The optimal rank is 6. (b) Predicted vs true loadings at the optimal rank; RMSECV = 0.17 $\mu g\,cm^{-2}$.

9.2.7 Poly(tetrafluoroethene)

Because of its non-stick properties and chemical inertness, poly(tetrafluoroethene) (PTFE) is used extensively in the pharmaceutical industry for reactor linings and other contact surfaces. Preliminary efforts have been made to produce a calibration on PTFE coupons, but were not successful [148]. The problems appeared to be mostly due to the physical nature of the coupons, which were slightly curved and had a visible pattern of longitudinal ridges; for details, see Ref. 148. In light of the importance of this substrate, demonstrating a successful IRRAS calibration on PTFE would be valuable.

9.3 Chemometric and sampling issues

9.3.1 Heterogeneity of calibration standards

The heterogeneity of the standards has proven to be a more significant problem than was anticipated at the outset of this project. In particular, it is difficult to draw conclusions about the precision of the IRRAS method: are the observed errors due to the IRRAS measurements, or can they be attributed to errors in the reference method? Given the theoretically predicted nonlinear behaviour of the IRRAS with respect to film thickness, it is certainly conceivable that the errors are not entirely due to the heterogeneity of the standards. The best solution to this problem would be to produce perfectly homogeneous standards, so that the errors in the reference method are known to be very small. Further work on improving the aerosol deposition method, including the use of a medical nebuliser rather than an air-brush, is ongoing and will be presented elsewhere [106]. It is important to emphasise that the heterogeneity referred to here is on the scale of millimetres to centimetres. Heterogeneity at smaller scales, such as that postulated above to be responsible for the nonlinearity seen for acetaminophen on steel, is an important property of the sample, and is likely to depend strongly on the substrate, the analyte and the manner in which the analyte is deposited. Ideally, large-scale heterogeneity would be greatly reduced, while small-scale heterogeneity would be controlled, with the aim of optimising the method so that it is robust to variations in the small-scale heterogeneity.

If a method for producing perfectly homogeneous standards cannot be found (or is too time-consuming for routine use), an alternative is to strictly define a set of parameters for the spray method (distance, spraying time, etc.) and directly measure the heterogeneity of the resulting standards. A first attempt at this process was described in Section 4.3.5, where some ideas for improvements are also given. The multinomial model for the errors described in Appendix A.1.2 may prove to be useful for describing the dependence of the heterogeneity on parameters such as the spraying time and size of the sampling error.

At the start of this work, it was thought that, in addition to allowing greater precision in the loading determination, using coupons larger than the infrared footprint and measuring several spectra from each would be a more efficient way to obtain spectra spanning the range of concentrations of the interfering species. However, it was found that there were significant correlations in the interferent concentrations between spectra for a single standard (see Section 3.5 for a theoretical discussion of this effect and Section 6.3 for an experimental demonstration). Consequently, it seems that there is little to be gained from taking several measurements from a single sample. It may be simpler to produce homogeneous standards when using smaller coupons, having dimensions similar to the illuminated area. However, since the illumination is not uniform, the use of smaller coupons would not eliminate the requirement for homogeneity (or characterisation of the heterogeneity).

9.3.2 Confidence intervals for predictions

It is very important to provide an estimate of the uncertainty of an analysis result, preferably in the form of a confidence interval with a specified coverage percentage. Several methods were discussed in Chapter 6, and it was found that bootstrap methods gave the best results. Unfortunately, the bootstrap is quite a time-consuming procedure, taking anywhere from a few seconds to a few minutes on a contemporary desktop PC (depending on the number of variables, the number of objects in the calibration set, and the number of bootstrap replicates). A formula-based approach would give results much more rapidly. An approximate error variance formula by Faber et al. [96] was tested and found to give confidence intervals that were too large. It is likely that this was due principally to an overestimate of the MSEC (caused by the reference-method errors) being used in the formula. A modified version of the formula taking into account the heteroscedastic reference-method errors would likely perform better.

9.3.3 Calibration transfer and model updating

Topics that are very important in process control, but which were not investigated in this thesis, include calibration transfer and model updating. The capability to transfer chemometric models between instruments is valuable because it greatly simplifies the calibration procedure if a model can be built using one spectrometer and then used on others. Model transfer has been discussed extensively in the literature: see, for example, Ref. 149 for a recent review and Ref. 150 for an example with mid-IR instruments. The most common problems are to do with differences in the types of baseline, the spectroscopic response and wavelength shifts. An issue related to the last is that, since each FTIR spectrometer will have, in general, a slightly different effective HeNe laser wavenumber (see Section 4.1.4), the abscissa spacing will be different. So, even if two instruments are calibrated correctly, some kind of

interpolation will be required to match the wavenumber scale of one to the other. All of these problems in calibration transfer can be addressed by various transformations of the model or the spectra, but model transfer between two fibre-optic FT-IRRAS systems has not yet been demonstrated.

Model updating is the practice of adding new data to a calibration model already in use, to account for the appearance of interferences (new species or baseline behaviour) that were not present in the initial calibration [151]. For example, if an IRRAS measurement were made, and a new interferent were present, this situation would be detected by unusually large spectroscopic residuals (see, for example, Section 6.4). The analyte loading predicted by the model would be considered unreliable, and the area would be swabbed and the swab analysed by an appropriate reference method. The measured spectrum could then be added to the original calibration model [152], so that future measurements would be unaffected by the new interferent. Efficient updating with only one or a few spectra containing the new interferent can be achieved by weighting the new sample more heavily in the model [153].

Appendix A

Statistical miscellany

A.1 Covariance matrices and weighted regression

A.1.1 Covariance and correlation

The covariance between two variables X_1 and X_2 is given by [154]

$$\mathrm{cov}(X_1, X_2) = \frac{\sum_{i=1}^{n} \left[(X_{1,i} - \bar{X}_1)(X_{2,i} - \bar{X}_2)\right]}{(n-1)} \tag{A.1}$$

where n is the number of samples and a bar denotes the mean. The correlation coefficient, commonly named R, is a scaled version of the covariance:

$$R = \frac{\sum_{i=1}^{n} \left[(X_{1,i} - \bar{X}_1)(X_{2,i} - \bar{X}_2)\right]}{(n-1)s_{X_1} s_{X_2}} \tag{A.2}$$

where s is the sample standard deviation.

A common assumption in chemometrics is that the residuals (or errors in $\mathbf{y}$) are (in addition to being normally distributed with mean zero) independently and identically distributed, *iid*. If this is the case, a single variance, σ_e^2 is enough to characterise the distribution of the residuals. Generally, however, the errors are not *iid*. For example, the variance of the errors might increase with increasing y. If the variances of the errors are not equal, the errors are said to be heteroscedastic. Additionally, there may be correlations between the errors, in which case they are not independently distributed.

The error variance-covariance matrix $\mathbf{V}_e$ describes the statistical structure of the errors, accounting for the general case of heteroscedastic, correlated errors. $\mathbf{V}_e$ has dimensions $n \times n$, where n is the number of samples. For *iid* errors, $\mathbf{V}_e = \mathbf{I}\sigma_e^2$. If the errors are heteroscedastic but not correlated, $\mathbf{V}_e$ will be diagonal. Non-zero off-diagonal elements indicate correlated errors.

A.1.2 Multinomial model for the spray method

An example relevant to this work involves samples measured in two ways. For each standard, a single measurement is made of the loading by UV colorimetry. This measurement is a precise estimate of the mean loading, but is subject to an error, characterised by σ^2_{UV}. This variance was estimated by colorimetric measurements of gravimetrically prepared standard solutions to be about 0.01 $\mu g\,cm^{-2}$.

Subsequently, each standard is measured by grazing-angle IRRAS. Spectra are measured from several (typically ~5–10) places on the coupon. Since the analyte is not perfectly homogeneously distributed over the surface of the coupon, the local loading for any particular measurement may be significantly different from the mean loading. This introduces another kind of error into the loading values, which is expected be proportional to the loading. This error has been estimated (see Section 4.3.5) to have a standard deviation of about 7 % of the loading; that is, $\sigma^2_{\mathrm{SE}} \approx (0.07y)^2$, where "SE" stands for sampling error.

In our chemometric models, all of the spectra from a given standard are assigned the measured mean loading for the standard. The errors we are interested in, therefore, are the differences between the local loadings and the measured mean loadings. Since a single mean-loading measurement is made for the standard, any error in the mean loading is the same for all spectra measured from that standard: this implies correlation. The error arising from sample heterogeneity, on the other hand, is assumed to be uncorrelated from measurement to measurement. Strictly, this cannot be true, since the total amount of analyte on the standard is effectively fixed and a significant fraction of the area is measured: if the local loading for one measurement is much higher than the mean, the local loading for the next measurement is likely to be lower than the mean. This means that the covariance terms are negative.

The experiment can be approximated by a multinomial model [118]. The coupon is divided into n regions of equal area, and then sprayed with N droplets of aerosol, all of which contain the same amount of analyte and each of which has probability $p = 1/n$ of landing in any given region. After spraying, the variable X_i represents the number of droplets that landed in region i. The mean of X_i is

$$\bar{X}_i = N/n \tag{A.3}$$

and the variance is

$$V(X_i) = \frac{N}{n}\left(1 - \frac{1}{n}\right) \tag{A.4}$$

The covariance between the number of droplets in any two regions is given by

$$\mathrm{cov}(X_i, X_j) = -N/n^2 \tag{A.5}$$

where $i \neq j$. N can be estimated by conducting experiments in which the local loadings of the n regions are measured directly (for example, by marking the regions on the plate and swabbing them; see Section 4.3.5). If the coupons are sprayed for the same length of time and the loading is adjusted by varying the concentration, N is constant. This model is only expected to be approximate, so N should be viewed as a parameter to fit rather than the literal number of aerosol droplets that adhere to the coupon.

The mean, variance and covariance calculated above are in units of "number of drops". To change to loading units, the following relationship is used:

$$y_i = X_i \frac{m}{A_i} \tag{A.6}$$

where m is the mass of analyte per droplet and $A_i = A_{\text{tot}}/n$ is the area of region i (A_{tot} being the total area of the coupon). It follows that the mean loading in region i is given by

$$\bar{y}_i = \bar{X}_i \frac{m}{A_i} = \bar{y} \tag{A.7}$$

and the variance by

$$V(y_i) = V(X_i)\frac{m^2}{A_i^2} = \frac{N}{n}\left(1 - \frac{1}{n}\right)\frac{m^2}{A_i^2} \tag{A.8}$$

The covariance is then

$$\text{cov}(y_i, y_j) = -\frac{Nm^2}{n^2 A_i^2} \tag{A.9}$$

Since $\bar{y} = Nm/A_{\text{tot}} = Nm/nA_i$, the variance and covariance expressions can be further simplified to

$$V(y_i) = \bar{y}^2 \frac{n-1}{N} \tag{A.10}$$

and

$$\text{cov}(y_i, y_j) = -\bar{y}^2/N \tag{A.11}$$

respectively. Equations A.10 and A.11 can be used to populate the diagonal and off-diagonal elements of the covariance matrix of the effective reference-method errors for the measurements from a single standard. (If the variance σ^2_{UV} is significant, it should be added to every element of the covariance matrix.) The error covariance matrix for several standards would have several of these smaller matrices strung along the diagonal (assuming that there is no correlation between standards).

A.1.3 Weighted regression

If the errors in the dependent variable are not *iid*, standard least-squares methods no longer give the optimal solutions [66]. If the error variance-covariance matrix $\mathbf{V}_e$ is known, however, the data can be transformed so that the errors become *iid*, and least-squares regression may be used.

First, $\mathbf{P}$, the matrix square root of $\mathbf{V}_e$, must be found (by a Cholesky decomposition [61], for example). Then, $\mathbf{Y}$ and $\mathbf{X}$ are premultiplied by $\mathbf{P}$ and the regression is carried out as normal. Draper and Smith [66] show how this transformation results in data with *iid* errors in the dependent variable. A discussion of several forms of weighted regression applicable to various error structures in both $\mathbf{X}$ and $\mathbf{Y}$ is given by Andrews et al. [103].

A.2 Studentised residuals

The studentised, or unit normal deviate, residuals, e^s, are calculated by dividing the residuals by their estimated variance [66]:

$$s_e^2 = \frac{\sum_{i=1}^{n}(e_i - \bar{e})^2}{n - df} \tag{A.12}$$

$$e_i^s = e_i / s_e^2 \tag{A.13}$$

The residuals are said to be externally studentised if s_e^2 is recalculated for each residual by considering all the other residuals (much like a cross validation).

A.3 *F*-tests

F-tests can be used to test compare the standard deviations of two populations [155]. The null hypothesis is that the two variances are equal, while the alternative hypothesis depends on the type of test:

H_0: $\sigma_1 = \sigma_2$

H_a: $\sigma_1 < \sigma_2$ for a lower one-tailed test

$\sigma_1 > \sigma_2$ for an upper one-tailed test

$\sigma_1 \neq \sigma_2$ for a two-tailed test

The test statistic is $F = s_1^2/s_2^2$, where s_1^2 and s_2^2 are the sample variances. The more this ratio deviates from unity, the stronger the evidence that the population variances are unequal. At the significance level α, the null hypothesis is rejected if

$F < F_{(1-\alpha,\nu_1,\nu_1)}$ for a lower one-tailed test

$F > F_{(1-\alpha,\nu_1,\nu_2)}$ for an upper one-tailed test

$F < F_{(1-\alpha/2,\nu_1,\nu_2)}$ for a two-tailed test

$F > F_{(1-\alpha/2,\nu_1,\nu_2)}$

where $F_{(\alpha,\nu_1,\nu_2)}$ is the tabulated critical value at significance level α of the F distribution with ν_1 and ν_2 degrees of freedom (which are the degrees of freedom associated with s_1^2 and s_2^2, respectively).

A.4 Uncertainty in MSEP estimates

If the prediction errors are normally distributed, the sum of the squared errors of prediction (SSEP) should follow a χ^2 distribution [96]. Consequently, the mean of the MSEP is proportional (if the prediction errors are *iid*) to n and the variance to $2n$: the relative error is given by

$$\frac{\sigma(\text{MSEP})}{\text{MSEP}} = \frac{\sqrt{2n}}{n} = \sqrt{\frac{2}{n}} \tag{A.14}$$

Using the linear approximation

$$\sigma(\text{RMSEP}) \approx \frac{\partial \text{RMSEP}}{\partial \text{MSEP}} \sigma(\text{MSEP}) = \frac{\sigma(\text{MSEP})}{2\text{RMSEP}} \tag{A.15}$$

the relative error in the RMSEP is given by

$$\frac{\sigma(\text{RMSEP})}{\text{RMSEP}} \approx \frac{1}{2} \frac{\sigma(\text{MSEP})}{\text{MSEP}} = \sqrt{\frac{1}{2n}} \tag{A.16}$$

Consequently, if a relative standard deviation of 10 % in the RMSEP is desired, about 50 test-set standards are required. The situation for heteroscedastic, correlated errors is more complex, but the above approach might still provide a reasonable estimate.

A.5 Degrees of freedom of complex variance estimates

A variance estimate that is a weighted sum of other variance estimates is called a complex variance estimate:

$$\hat{V} = \sum_{i=1}^{N} a_i \hat{V}_i \tag{A.17}$$

where the a_i are the weights. Satterthwaite's rule [98] gives an estimate of ν, the number of degrees of freedom of $\hat{V}$.

$$\nu \approx \frac{\hat{V}^2}{\sum_{i=1}^{N} \frac{a_i^2 \hat{V}_i^2}{\nu_i}} \tag{A.18}$$

In the particular case of the approximate prediction error variance formula introduced in Section 3.4.2 (Equation 3.56),

$$\begin{aligned} V(\mathrm{PE_u}) &\approx (n^{-1} + h_\mathrm{u})(\hat{V}_e + \hat{V}_{\Delta y} + \|\boldsymbol{\beta}\|^2 \hat{V}_{\Delta X}) + \hat{V}_e + \|\boldsymbol{\beta}\|^2 \hat{V}_{\Delta X} && \text{(A.19)} \\ &= (1 + n^{-1} + h_\mathrm{u})\hat{V}_e + (n^{-1} + h_\mathrm{u})\hat{V}_{\Delta y} + (n^{-1} + h_\mathrm{u} + 1)\|\boldsymbol{\beta}\|^2 \hat{V}_{\Delta X} && \text{(A.20)} \end{aligned}$$

where the n^{-1} terms should be omitted if mean centring is not used. Applying Equation A.18 to this case gives

$$\nu \approx \frac{\hat{V}(\mathrm{PE_u})^2}{(1 + n^{-1} + h_\mathrm{u})^2 \hat{V}_e^2/\nu_e + (n^{-1} + h_\mathrm{u})^2 \hat{V}_{\Delta y}^2/\nu_{\Delta y} + (n^{-1} + h_\mathrm{u} + 1)^2 \|\boldsymbol{\beta}\|^4 \hat{V}_{\Delta X}^2/\nu_{\Delta X}} \tag{A.21}$$

The degrees of freedom for the individual variance estimates depend on the number of replicate measurements. According to Burdick and Graybill [98], Satterthwaite's rule is reasonable if the ν_i are all similar or all large, but can produce an unrealistically small estimate if there are large differences between them. In any case, the importance of the number of degrees of freedom should not be overstated. Comparing, for example, $t_{0.95,10} = 1.81$ and $t_{0.95,\infty} = 1.64$, it is apparent that unless ν is very small, a large error in ν will cause only a small error in the size of the confidence interval.

A.6 Bias testing based on joint confidence regions

When a model with a single parameter (e.g. a line with a fixed intercept) is fitted to data, a confidence interval can be derived from the uncertainty in the parameter. However, when multiple parameters are fitted, their errors may be correlated. For example, if the slope is too great, the intercept will tend to be too small, and vice versa. For this reason, the uncertainty in the parameters should be expressed as a joint confidence region. In general, the joint confidence region for n parameter estimates is an n-dimensional ellipsoid. For the case of fitting a straight line $y = b + mx$ through a series of N points (x, y), the formula for the ellipse (Equation A.23) is given by Mandel and Linnig [156] in terms of various quantities calculated during least-squares fitting (Equations A.22).

$$
\begin{aligned}
S &= \sum x; \quad Y = \sum y \\
Q &= \sum x^2; \quad L = \sum y^2; \quad P = \sum xy \\
\Delta &= NQ - S^2 \\
\hat{m} &= \tfrac{NP-SY}{\Delta}; \quad \hat{b} = \tfrac{QY-SP}{\Delta} \\
s^2 &= \tfrac{1}{N-2}\left(L - \tfrac{Y^2}{N} - \tfrac{\Delta}{N}\hat{m}\right)
\end{aligned}
\tag{A.22}
$$

The equation of the ellipse is

$$
N(b-\hat{b})^2 + 2S(b-\hat{b})(m-\hat{m}) + Q(m-\hat{m})^2 = 2Fs^2 \tag{A.23}
$$

where F is the critical value of the variance ratio with 2 and $N-2$ degrees of freedom at the desired confidence level (α). In practice, it is much easier to plot an ellipse parametrically. The ellipse is defined by its two axes β and γ, its rotation angle ϕ, and its centre $(\hat{b}, \hat{m})$. The angle θ is varied from zero to 2π and the corresponding $(b-\hat{b}, m-\hat{m})$ points are calculated and plotted. For an ellipse in the form of Equation A.23, the relevant equations are [157]

$$
\beta = \sqrt{\frac{-4Fs^2 + 4NQFs^2}{(S^2 - NQ)\left[(Q-N)\sqrt{1 + \frac{4S^2}{(N-Q)^2}} - N + Q\right]}} \tag{A.24}
$$

$$
\gamma = \sqrt{\frac{-4Fs^2 + 4NQFs^2)}{(S^2 - NQ\left[(N-Q)\sqrt{1 + \frac{4S^2}{(N-Q)^2}} - N + Q\right]}} \tag{A.25}
$$

$$
\phi = -\frac{1}{2}\cot^{-1}\left(\frac{Q-N}{2S}\right) \tag{A.26}
$$

$$
b - \hat{b} = \beta\cos\theta\cos\phi - \gamma\sin\theta\sin\phi \tag{A.27}
$$

$$
m - \hat{m} = \beta\cos\theta\sin\phi + \gamma\sin\theta\cos\phi \tag{A.28}
$$

When comparing concentrations estimated using a new analytical method against accurate reference values, the best-fit line should have zero intercept and unit slope. If a confidence region constructed around the measured slope and intercept does not contain the point $(b_0, m_0) = (0, 1)$, the new method is exhibiting bias significant at the α confidence level. MATLAB code to generate a list of points to plot a joint confidence ellipse is listed in Appendix C as `JCR`.

To find the value of α at which the bias is just significant (called the p-value), an iterative scheme has been implemented. The shape and position of the ellipse do not change when α is varied (just its size). The first step is to find the angle, θ_0, corresponding to the intercept of the ellipse with the line drawn from the centre of the ellipse through (b_0, m_0). Then, α can be varied to minimise the distance

between the point on the ellipse with angle θ_0 and the point (b_0, m_0).

The notation is simplified if we take the ellipse to be centred at $(0, 0)$, and $\hat{b}$ and $\hat{m}$ to have been subtracted from b_0 and m_0. The line between the centre of the ellipse and (b_0, m_0) is then given by

$$m = \frac{m_0}{b_0} b \tag{A.29}$$

and the aim is to solve Equations A.29, A.27, A.28 for θ. Taking the ratio of b and m,

$$\frac{b_0}{m_0} = \frac{\beta \sin\phi \cos\theta + \gamma \cos\phi \sin\theta}{\beta \cos\phi \cos\theta - \gamma \sin\phi \sin\theta} \tag{A.30}$$

and solving for θ,

$$\tan\theta_0 = \frac{\beta}{\gamma} \frac{m_0 \cos\phi - b_0 \sin\phi}{m_0 \sin\phi + b_0 \cos\phi} \tag{A.31}$$

Care must be taken to obtain θ_0 in the correct quadrant. The $\tan^{-1}$ function in mathematical software usually returns angles in the interval $(-\pi/2, \pi/2)$, while an angle between 0 and 2π is required, so additional information is needed to decide on the quadrant. This information can be obtained by drawing a line from the point on the ellipse with $\theta = 0$ to that with $\theta = \pi$ and bearing in mind that θ increases anticlockwise. The equation of the line is given by $m = b \tan\phi$. The required adjustment to $\tan^{-1}(\tan\theta_0)$ depends on whether (b_0, m_0) falls above or below this line, and is given in the table below.

	$\tan^{-1}(\tan\theta_0) < 0$	$\tan^{-1}(\tan\theta_0) > 0$
$m_0 > b_0 \tan\phi$	$+\pi$	0
$m_0 < b_0 \tan\phi$	$+2\pi$	$+\pi$

Once θ_0 has been determined, α may be varied to minimise the distance between (b_0, m_0) and (b, m). The minimisation is carried out in MATLAB by a golden section search [158], with the starting interval $[0, 1]$ (which is guaranteed to bracket α). MATLAB code for this procedure is listed in Appendix C as `JCR_P`.

Appendix B

Water vapour in mid-infrared spectroscopy

A potential problem with open-path IRRAS (and in infrared spectroscopy in general) is the appearance of water vapour absorption bands in the spectra due to small changes in humidity or path length between recording the background and sample single channels. These bands have very sharp lines and make quantification of other species whose bands they overlie difficult (by traditional univariate means, at least). In the laboratory, it is possible to deal with this problem experimentally by careful purging, but in industrial settings this is frequently impractical, and some software-based approach must be taken. In principle, inverse regression methods, such as PLS regression, should be able to model the water-vapour absorbance implicitly, provided that it is present in the calibration set and is uncorrelated to the analytes. It may be better, however, to remove the water vapour bands before the calibration. This pre-processing step would have the additional advantages of rendering the spectra more amenable to inspection and of enabling calculation of band integrals. The effect of water vapour removal as pre-processing in PLS regression was tested in Chapter 7, where it was found often to reduce the optimal dimensionality of the model, but not the RMSECV at optimal rank. This appendix describes several algorithms for minimising water vapour bands.

B.1 Appearance of the water vapour infrared absorption bands

Water has three vibrational modes: a symmetric stretch at 3657 cm^{-1}, an antisymmetric stretch at 3756 cm^{-1}, and a bending mode at 1595 cm^{-1}. As can be seen in, for example, Figure 4.9, rotational structure in the bands is evident.

The single-beam "background" spectrum B from the spectrometer is the blackbody emission curve

of the infrared source modified by the absorbance A_w of the water vapour, the reflectance R_0 of the uncontaminated substrate, and the response curve of the detector. B_0 will be used here to mean the background single channel in the absence of any absorption by water and before reflection from the substrate. The single-beam "sample" spectrum, S, consists of the same blackbody curve modified by the absorbance $A_w + \delta A_w$ of a possibly slightly different concentration of water vapour and the reflectance R of the contaminated substrate. The reflection-absorbance spectrum A is calculated as an absorbance spectrum, and the quantity A_s derives from the change in reflectance caused by the presence of the contamination.

Thus, the single-channel background and sample spectra are given by

$$B = R_0 \times B_0 \times 10^{-A_w} \tag{B.1}$$

$$S = R \times B_0 \times 10^{-(A_w + \delta A_w)} \tag{B.2}$$

If the conditions of the experiment are reasonably constant on the timescale over which a spectrum is measured, then it is realistic to assume that δA_w is just A_w times a constant, and the final absorbance spectrum is

$$\begin{aligned} A &= \log_{10}(B/S) \\ &= \log_{10}(R_0/R) + \delta A_w \\ &= A_s + \delta A_w \end{aligned} \tag{B.3}$$

The problem is the appearance, in the absorbance spectrum, of the water vapour absorbance δA_w.

B.2 Algorithms for reducing water vapour absorption bands

B.2.1 Derivative minimisation subtraction of a reference spectrum

The most obvious method of dealing with these unwelcome bands is to subtract an appropriately scaled reference water vapour absorbance spectrum A_w^0 from the sample absorbance spectrum.

$$A_s = A - bA_w^0 \tag{B.4}$$

To use this approach, b and A_w^0 must be determined. A_w^0 is straightforward to measure experimentally, since it is simple to induce a change in the ambient concentration of water vapour (for example, by flushing the sample area with air that has been bubbled through water). The scaling factor b can

be determined by trial and error, with visual inspection of the resultant spectrum, but obviously this is time-consuming and not entirely reproducible. Another method is to find b algorithmically by minimising some function $f(b)$, provided that the minimum of the function coincides with the true value of b. Software packages such as GRAMS [159] typically offer an auto-subtract algorithm [160] in which $f(b)$ is the integral of the absolute value of the derivative of the resultant spectrum;

$$f(b) = \int \left| \frac{\mathrm{d}\left(A(\bar{\nu}) - bA^0_{\mathrm{w}}(\bar{\nu})\right)}{\mathrm{d}\bar{\nu}} \right| \mathrm{d}\bar{\nu} \tag{B.5}$$

where $\bar{\nu}$ is the frequency in wavenumbers. For a digitised spectrum, the integral is replaced by a sum and the derivative by a difference:

$$f(b) = \sum_{i=1}^{n-1} \left| \frac{d_{i+1} - d_i}{\bar{\nu}_{i+1} - \bar{\nu}_i} \right| \tag{B.6}$$

where $d_i = A_i - bA^0_{\mathrm{w,i}}$.

The idea is essentially that the spectrum will be smoothest when the correct amount of the interferent is subtracted, and that this will be reflected in a minimum in the magnitude of the derivative integrated over some region of the spectrum. However, the minimum of this function does not in general correspond to the correct value of b unless

$$\left| \frac{\mathrm{d}(A - bA^0_{\mathrm{w}})}{\mathrm{d}\bar{\nu}} \right| = \left| \frac{\mathrm{d}A}{\mathrm{d}\bar{\nu}} \right| - \left| b \frac{\mathrm{d}A^0_{\mathrm{w}}}{\mathrm{d}\bar{\nu}} \right|$$

which is not generally true when bands in A^0_{w} overlap with those of A. However, since the water vapour absorbance spectrum consists of many sharp lines, the effect of the lines that lie near sample bands is outweighed by that of the lines that don't, and the estimate of b should be close to the true value. The range of frequencies to consider should be chosen to minimise overlap between bands due to water vapour and bands due to the analyte. A function has been written in MATLAB to carry out the minimisation of Equation B.5 (`UNWIGGLE`; see Appendix C), and the results obtained are quite reliably as good as those from the trial and error approach. This is a relatively "safe" method because, if the reference spectrum doesn't match the interfering spectrum (or if there is no interference), b will be calculated to be small since adding any of the reference spectrum to the sample spectrum will tend to increase $f(b)$. Another benefit is that b provides an estimate of the relative concentration of water vapour.

A general problem with the reference-spectrum subtraction method is that it necessitates measuring an additional spectrum, and relies on this spectrum differing from the interfering spectrum only by a scaling factor. For this to be true the underlying spectrum must be the same and the measurement of

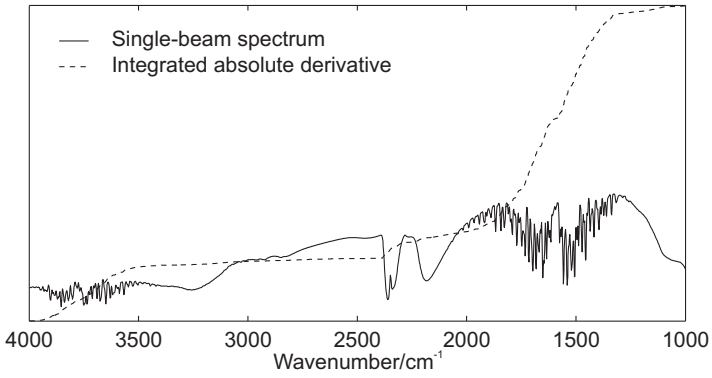

Figure B.1: A typical single-beam spectrum (solid line) and the cumulative sum of the absolute value of its derivative (dashed line; not to scale)

it must be reproducible. There is some uncertainty as to whether it is reasonable to expect either of these criteria to be the true for water: changes in temperature may affect the rotational populations, and small frequency shifts due to spectrometer instability may make significant differences because the lines are narrower than the instrument resolution. For these reasons, and to minimise the amount of measurement required, a method that doesn't rely on a reference spectrum may be preferable.

B.2.2 Derivative minimisation subtraction of the background

Because the instrument used in this work is not purged and there is a significant path length through laboratory air, water vapour absorbance is significant and the large number of sharp lines make the overwhelmingly dominant contribution to the magnitude of the derivative of the single-channel spectra, as illustrated in Figure B.1. This can be exploited to allow the removal of most of the water absorption from the sample spectrum at the cost of introducing some other, hopefully less harmful, artefacts.

If b is found so that

$$b = 1 + \frac{\delta A_w}{A_w} \tag{B.7}$$

then a modified absorbance spectrum can be calculated:

$$\begin{aligned} A' &= \log_{10}(B^b/S) \\ &= \log_{10}\left[\frac{{R_0}^b {B_0}^b 10^{-bA_w}}{RB_0 10^{-(Aw+\delta A_w)}}\right] \\ &= \log_{10}\frac{{R_0}^b {B_0}^b}{RB_0} \\ &= A_s + (b-1)\log_{10}(R_0 B_0) \end{aligned} \tag{B.8}$$

Comparing Equations B.3 and B.8 it is apparent that the water vapour absorbance δA_w has been

replaced by the artefact $(b-1)\log_{10}(R_0B_0)$. If something is known about R_0B_0 then such artefacts may be partially correctable. In particular, if R_0B_0 is reasonably constant over the region of interest, the artefacts take the form of a constant offset, which can be dealt with easily.

Because the majority of the intensity of the derivative of the single channel spectra arises from the water vapour contribution (see Figure B.1), a close approximation to b can be found by minimising

$$f(b) = \int_{x_0}^{x_1} \left| \frac{\mathrm{d}A'(b)}{\mathrm{d}x} \right| \mathrm{d}x \tag{B.9}$$

This minimisation can be carried out using the auto-subtract algorithm described above; essentially, it is the same as the reference spectrum method except that $\log_{10}(S)$ takes the place of the absorbance spectrum, $\log_{10}(B)$ is used as the reference water vapour spectrum and the definition of b is slightly different.

Obviously this is not an ideal solution to the problem because it involves introducing new artefacts at the same time as removing old ones (see Figure B.2), but it is likely that the introduced artefacts are the lesser of the two evils, certainly as far as qualitative analysis of the spectrum is concerned.

B.2.3 Estimating the background

In principle, a reference water vapour spectrum can be obtained from the background spectrum if the background spectrum in the absence of water (B_0R_0) is known.

$$A_w^0 = \log_{10}(B_0R_0/B) \tag{B.10}$$

If the spectrometer is well purged or evacuated, then the background may be measured directly, and A_w^0 will be the absorbance of all atmospheric gases, not just water vapour. In practice water vapour and CO_2 are likely to be the significant absorbing gases. The spectrum of CO_2 does not interfere significantly with that of water vapour. If direct measurement of B_0R_0 is not possible, it may be feasible to estimate it by fitting some function to B. This is difficult because the water bands usually obliterate all information about the baseline over a significant frequency range. Another possibility is to calculate low-resolution versions of B from the original interferogram; but, rather than eliminating the water bands entirely, this replaces them with a broad, featureless band. This approach is certainly difficult to automate, given that R_0 varies widely depending on the nature of the substrate (metal, glass, rubber, etc.)

B.2.4 Smoothing-based methods

Generalised approaches based on smoothing [161] provide another way to deal with interference by water vapour, and can be effective if the sample bands are well resolved and the interfering bands are much sharper than the sample bands. The idea behind these methods is to calculate a moving average for the spectrum, and then, by comparison with the original spectrum, to calculate a new moving average from which points that differ by some threshold from their smoothed values are omitted. These methods are particularly suited to isolated spike noise, but have been applied successfully to the removal of water vapour lines from infrared spectra [162]. Preliminary investigations indicate that these techniques are unsuitable for our spectra because the instrument resolution is not sufficiently greater than the sample bands.

B.2.5 OPUS atmospheric compensation

The OPUS [105] software from Bruker provides an "atmospheric compensation" feature that claims to remove bands due to atmospheric water and/or carbon dioxide. The algorithm is undocumented, and requires input of both B and S, but no reference spectrum. It is somewhat unstable—usually significant improvement is made, but sometimes spectra are dramatically degraded (as in Figure B.2a). The algorithm is quite sensitive to wavenumber scale errors (see Section 4.2.6), so may not be safe to use in an unsupervised manner. It is quite possible that the constant wavenumber ratio error due to the miscalibration of our instrument (see Section 4.2.6) is responsible for this behaviour, and that small changes due to factors other than laser misalignment do not cause problems.

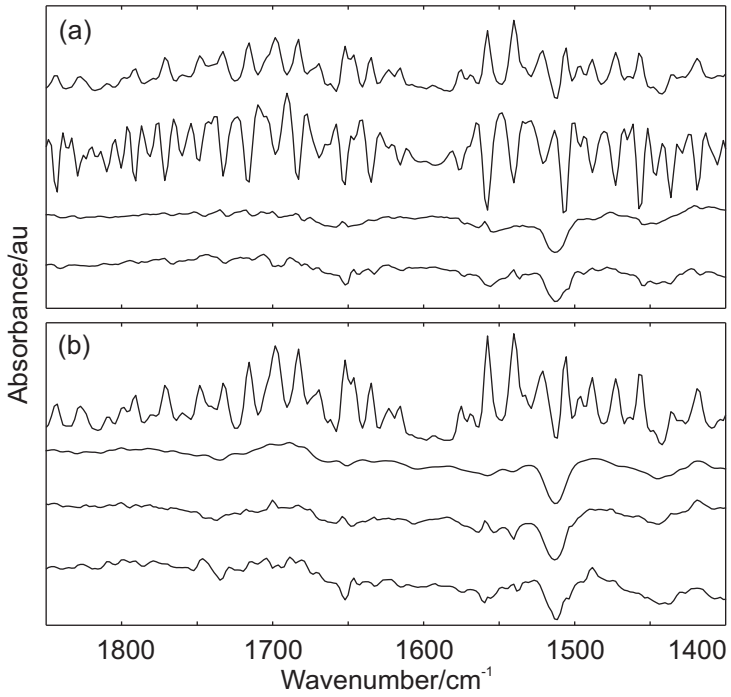

Figure B.2: IRRAS of acetaminophen on stainless steel, showing the effect of several water-vapour removal algorithms. (a) A spectrum for which OPUS atmospheric compensation fails; (b) one for which it works well. In each case the treatments are, in descending order: none, OPUS, reference spectrum auto-subtraction, background auto-subtraction.

Appendix C

MATLAB and other code listings

C.1 Introduction and instructions

This appendix contains code listings for some the software developed as part of this work for building and validating chemometric models. Most of the software is written in the MATLAB language, and is designed to be executed from the MATLAB command-line.

The main advantage of this software over commercial packages such as QUANT2 (part of OPUS) is the flexibility that the MATLAB environment provides: a set of spectra is represented as an array, and may be manipulated by any of the standard MATLAB functions.

The instructions given here are brief, but all the functions listed here begin with a comprehensive block of help comments.

C.1.1 Converting spectra from OPUS

To load spectra in MATLAB, they first must be converted to delimited text files. Unfortunately, due to bugs and limitations in software, converting a large number of files from the OPUS format is somewhat difficult. Before describing how this has been done, a brief summary of the relevant file formats might be useful:

OPUS uses an undocumented, binary file format.

SPC is the binary format used by ThermoGalactic's GRAMS/AI software. The specification is available from ThermoGalactic, and free utilities exist to convert from SPC to plain text.[1] The PLS Toolbox for MATLAB published by Eigenvector [146] includes a function for loading SPC files.

[1] See, for example, SHOWSPC, released by the US EPA, and currently available from http://www.epa.gov/ttn/emc/ftir/showspc.html.

JCAMP-DX is a file format specification [163] in which the data file is plain text but the spectroscopic data are stored in a slightly compressed manner.

Spectra measured in OPUS are automatically saved in its native format. OPUS has a "Save as" function that can export spectra to SPC or to plain text ("data point table"), but this function must be invoked for each individual spectrum, which is extremely time-consuming. A much faster alternative is to use the Import/Export module of Galactic's GRAMS/AI software, which can convert OPUS files to SPC and can operate on many files at once.

However, GRAMS' converter has the limitation that only the first data block in the OPUS file will be converted, and then only if it is of a type recognised by GRAMS. Unfortunately, while the "absorbance spectrum" and "sample single-channel spectrum" block types are recognised, the "background single-channel spectrum" is not. As a final complication, the ASCII (plain text) files created by GRAMS are problematic to load in MATLAB. To circumvent this problem, and to allow storage in a slightly more efficient format, a progamme has been written to allow MATLAB to read JCAMP-DX files (see `DX2CSV` and `DXREAD` below).

The following procedure has been used to convert the single-beam spectra to a MATLAB-readable format:

1. Make a copy of all the OPUS-format files to be converted.
2. In OPUS, open all of the spectra.
3. Select all the absorbance data blocks and invoke the "Delete block" function. The first data block in each file should now be the sample single-channel (SSC) spectrum.
4. Save all the files, then unload them.
5. In GRAMS, import all the files using the OPUS filter. This will create an SPC file for each OPUS file, containing only the sample single-beam spectrum.
6. In GRAMS, export the SPC files to JCAMP-DX format.
7. Return to OPUS and re-open all the files.
8. Manually save all the background spectra in either data point table or JCAMP-DX format. This is usually less tedious than for the sample spectra, since there is normally one unique background per block of eight or so spectra.

C.1.2 Chemometric tools

Once the spectra have been loaded, they can be plotted and processed using standard MATLAB functions. Some additional functions have been written for tasks relevant to chemometric pre-processing:

- `UNWIGGLE` carries out derivative-minimisation subtraction of one spectrum (usually water vapour) from another [160, 164].
- `SAVGOL` smooths spectra by the Savitzky-Golay convolution method [85, 86, 136], and can also be used to calculate derivative spectra.
- `POLYBLSUB` can be used to subtract a low-order polynomial baseline from spectra.

The main tool for chemometric model optimisation is `CROSSVAL`. This function carries out a cross validation using PCR, PLS or their polynomial variants, as described in Chapter 3. It supports "per-spectrum" and "per-sample" cross validations in which sequential blocks of objects are left out; other cross-validation patterns may be obtained by permuting the rows of the input matrices beforehand as desired.

For actual use of PLS models, the function `PLS` should be used. This simply calls `IMP_PLS_K1` (an efficient kernel PLS algorithm [72]) and calculates some additional quantities. Predictions can be made with `PLSPRED`. Code for PCR models is simple and can be copied from `CROSSVAL` if necessary.

Several methods for estimating confidence intervals are discussed in Section 3.4.2. Code for the jack-knife (`JK_CI`), object (`BSOBJ_CI`) and residual (`BSRES_CI`) bootstraps, along with an approximate errors-in-variables-model variance formula (`PREDVAR`) are listed below.

One approach to testing models for bias is to model the predicted values as a linear function of the true values and to determine the joint α-confidence region for the slope and intercept; see Appendix A.6. The functions `JCR` and `JCR_P` can be used to plot joint confidence regions and to calculate p-values, respectively.

C.1.3 Programmes for IRRAS calculations

Two programmes were used for the bulk of the calculations in Chapters 2 and 5. `REFL3` calculates the Fresnel reflection coefficients for a three-phase system, while the electric field intensities are determined by `FIELDS3`.

C.2 MATLAB code listings

C.2.1 Spectrum processing tools

DXREAD

This function requires the Python JCAMP reader listed in Section C.3. The M-file should be edited to reflect the path of the Python script.

```
function S=dxread(dxfile,dx2csvpath,tempfile)

% function S=dxread(dxfile,dx2csvpath,tempfile)
%   Runs the Python script dx2csv.py to convert JCAMP files to CSV and
%   then loads the CSV file into S. The first column of S is the
%   abscissa, the second the ordinate.
%
%   Inputs:
%       dxfile is the JCAMP-DX file to load
%       dx2csvpath is the filename including path of the python script
%       tempfile is the temporary .csv file to create (default is
%         documents/spectra/dx2csvtemp.csv).  This name must contain
%         only one sort of slash used only to separate folder levels.

if isunix;
    pre='~/'; del='!rm ';
else
    pre='d:/'; del='!del ';
end

if nargin<3;
    tempfile=[pre,'documents/spectra/dx2csvtemp.csv'];
if nargin<2;
    dx2csvpath=[pre,'documents/uni/python/dx2csv.py'];
end; end

evalstr=['!python ',dx2csvpath,' ',dxfile,' ',tempfile];
eval(evalstr)

S=load(tempfile);

eval(sl([del,tempfile]));

function str=sl(str)
    if isunix;
        str(find(str=='\'))='/';
    else
        str(find(str=='/'))='\';
    end
```

UNWIGGLE

```
function [Xout,x]=unwiggle(X,S,range,tol,maxits)
% function [Xout,b]=unwiggle(X,S,range,tol,maxits)
%   Unwiggles a vector by subtracting the optimum amount of another
%   vector from it.
%
%   Xout = X - bS, where b is calculated iteratively on the basis of
```

```
%   minimising the absolute integral of the derivative of Xout in the
%   range specified by range.
%
%   X and S must be vectors of the same length
%   range is a vector specifying which elements in X are to be
%     considered. If unspecified or empty, the full spectrum is used.
%   tol = convergence criterion (default is sqrt(eps)^1.5e-8)
%   maxits = max no. of iterations (default 1000)

% The function to minimise is f(b) = sum(abs(diff(X-b*S))).
% 1. Far away from the minimum f(b) is linear, so the intercept of
%    lines fitted through two points on either side of the minimum, far
%    away, should be near the minimum. This provides a good initial
%    guess.
% 2. The minimum is bracketed by choosing an interval about the guess
%    and enlarging it in Golden steps until the minimum is located.
% 3. The minimum is then found precisely by a Golden search.
%
% See Press et al, Numerical Recipes in C 2nd ed, Cambridge (1992).
% Brent's method would probably work very well for this problem.

% Process input arguments and assign defaults
if nargin<5; maxits=1000;
if nargin<4; tol=sqrt(eps);
if nargin<3; range=[];
if nargin<2
    error('At least two input arguments must be supplied.')
end; end; end; end;

% Golden ratio
phi=1.61803398874989;

% Save full vectors to calculate Xout later, then select range
X0=X; S0=S;
if ~isempty(range)
    X=X(range); S=S(range);
end

% Initial guess for x
q=norm(X)/norm(S); %(in case S and X are scaled very differently)
xtest=q*[-6 -4 4 6]';
diffs=Fvec(X,xtest,S);

m1=(diffs(2)-diffs(1))/(xtest(2)-xtest(1)); c1=diffs(1)-m1*xtest(1);
m2=(diffs(4)-diffs(3))/(xtest(4)-xtest(3)); c2=diffs(3)-m2*xtest(3);

xinit=(c2-c1)/(m1-m2);

% The starting interval is empirically determined to optimise the
% tradeoff between wasting time bracketing the minimum and getting as
% close as possible before starting the search.
q=q*1e-3;

% Bracket the minimum
% a and b are the left and right bracketing values
a=xinit-q; n=0;
if F(X,a,S)<F(X,xinit,S);
    % search for a
    b=xinit; xinit=a; a=a-q;
    while(F(X,a,S)<F(X,xinit,S))
        n=n+1;
        if n>maxits; error('Failed to bracket minimum.'); return; end
```

```
        a=a-phi*(b-a);
    end
else
    % search for b
    b=xinit+q;
    while(F(X,b,S)<F(X,xinit,S))
        n=n+1;
        if n>maxits; error('Failed to bracket minimum.'); return; end
        b=b+phi*(b-a);
    end
end

x=0; xold=tol+1; n=0;

% "Golden search" for the minimum
while abs(x-xold)>tol
    n=n+1;
    xold=x;
    if n>maxits;
        warning('Maximum number of iterations exceeded!'); break;
    end
    x=(b+a*phi)/(phi+1);
    x2=(b+x*phi)/(phi+1);
    if F(X,x2,S)<F(X,x,S);
        a=x; b=b;
    else
        a=a; b=x2;
    end
end

Xout=X0-x*S0;

function diffs=F(X,x,S)
    % Evaluate F for scalar x
    diffs=sum(abs(diff(X-x*S,1,2)),2);

function diffs=Fvec(X,x,S)
    % Evaluate F for vector x
    diffs=sum(abs(diff(ones(length(x),1)*X-x*S,1,2)),2);
```

SAVGOL

```
function Yout=savgol(Y,N,M,D,dx,pad)
% function Yout=savgol(Y,N,M,D,dx,pad)
%   Applies Savitzky-Golay smoothing/differentiation to the rows of Y.
%
%  N = order of polynomial to use
%  M = number of points either side of central point to use
%  D = order of derivative to take (0 = smooth only, default)
%  If D>0 then the output is scaled by 1/dx^D, where dx is the spacing
%  of the points.
%  pad governs treatment of the end points
%       0 -> first and last M points are truncated
%       1 -> first and last M points are set to the first/last smoothed
%            value
%       2 -> polynomials of order N are fit through first/last 2M+1
%            points and used to calculate the first/last M points
%            (default)

if nargin<6; pad=2;
if nargin<5; dx=1;
```

```
    if nargin<4; D=0;
    if nargin<3; M=9;
    if nargin<2; N=2; end; end; end; end; end
23
    %Check for valid arguments
    if N>2*M
        error('Invalid input arguments: N must be < 2M+1')
    end
28
    if D>N
        error('Invalid input arguments: D must be <= N')
    end

33  % Get the coefficients
    C=savgolcoeff(N,M,D);

    Y=Y'; % Algorithm below operates on columns
    Yout=zeros(size(Y,1)-2*M,size(Y,2));
38
    % Perform the convolution
    for ii=1:size(Yout,1);
        Yout(ii,:)=C*Y(ii:ii+2*M,:);
    end
43
    % Pad the vector to its original length if requested
    if pad==1
        % Pad with constants
        Yout=[ones(M,1)*Yout(1,:);Yout;ones(M,1)*Yout(end,:)];
48  elseif pad==2
        % Fit polynomial
        Yout=[zeros(M,size(Yout,2));Yout;zeros(M,size(Yout,2))];
        for hh=1:size(Y,2)
            a=polyfit([1:2*M+1]',Y(1:2*M+1,hh),N);
53          for ii=1:D; a=polyder(a); end
            Yout(1:M,hh)=polyval(a,1:M)';
            a=polyfit([-M:M]',Y(end-2*M:end,hh),N);
            for ii=1:D; a=polyder(a); end
            Yout(end-M+1:end,hh)=polyval(a,1:M)';
58      end
    end

    % Scaling for derivatives
    if D>0; Yout=Yout/dx^D; end
63
    Yout=Yout';

    %--------------------------------------------------------------------
    function C=savgolcoeff(N,M,D)
68  % function C=savgolcoeff(N,M,D)
    %
    % Calculates the convolution coefficients for smoothing data for the
    % Savitzky-Golay least-squares method (as presented by Steinier et al,
    % Anal. Chem. 44(11) 1906 (1972)). The coefficients are normalised,
73  % and C is a row vector.

    X=[-M:M]'*ones(1,N+1);
    P=ones(2*M+1,1)*[0:N];
    X=X.^P;
78  T=inv(X'*X)*X';
    C=T(D+1,:)*factorial(D);
```

POLYBLSUB

```
function Xout=polyblsub(X,n,x)

% function Xout=polyblsub(X,n,x)
%   A polynomial of order n is fitted to each row of X and then
%   subtracted.
%   x indicates the column indices to use for the fitting (to avoid CO2
%   bands, etc.) If not specified or 0, the baseline is fitted to the
%   whole spectrum.

if nargin<3; x=0;
if nargin<2; p=1; end; end

[rows,cols]=size(X); B=zeros(rows,cols);
if (~x) | isempty(x); x=1:cols; end

for ii=1:rows
    [p s mu]=polyfit(x,X(ii,x),n);
    B(ii,:)=polyval(p,1:cols,[],mu);
end

Xout=X-B;
```

CROSSVAL

```
function [Yp, RMSECV, RMSECVavg, R2, R2avg, Ypavg, Xr, Xp]= ...
                    crossval(X, Y, regmethod, A, nrep,nout, meancentre)

%function [Yp RMSECV RMSECVavg R2 R2avg Ypavg Xr Xp]= ...
%                    crossval(X, Y, regmethod, a, nrep,nout, meancentre)
%
% Carries out a multivariate leave-some-out cross validation, to
% assist with rank selection and model optimisation. Several factor-
% based algorithms are available and datasets with semi-independent
% replicate spectra are supported.
%
% Inputs:
%   X (spectra in rows)
%   Y (mixture concentration profiles in rows)
%   regmethod: 'pcr', 'polypcr' (polynomial pcr), 'pls', 'polypls'
%   A is the maximum rank (model complexity) to calculate. For polypcr
%     or polypls, A must have two elements; the first is the polynomial
%     order.
%   nrep = number of replicate spectra per sample; default = 1. If
%     samples have differing numbers of spectra, nrep must be a vector.
%   nout = number of samples to leave out each time; default = 1
%   if meancentre~=0 then X and Y are mean-centred; default = 0
%
% Outputs:
%   Yp is an array with dimensions (n,a,c), where n is the number of
%     spectra (rows of X) and c is the number of components (columns of
%     Y).
%   The rows of RMSECV are the RMS prediction error for a compound as a
%     function of rank.
%   RMSECVavg has the same dimensions as RMSECV but is calculated on a
%     per-sample (rather than per-spectrum) basis from the averaged
%     predictions.
%   R2 and R2avg are the determination coefficients, following the same
%     form as RMSECV and RMSECVavg.
%   Ypavg is an array with dimensions (nsamps,a,c) calculated by
```

```
%     averaging the predictions of all replicates for each sample.
%   Xr is a matrix of spectral residuals (not implemented for all
%     algorithms). Dimensions n by a, with each element the RMS
%     residual.
%   Xp is an array (n,m,a) of fitted spectra.
%
% Note that PLS2 is used when Y is 2-dimensional; i.e. a single model
% is used to predict the concentrations of all species simultaneously.
% To use PLS1, run once per compound, giving 1-dimensional Y.

% Notes
% Polynomial PCR is very slow since it recalculates the eigenvectors
%   for each rank.
% Functions used: imp_pls_k1, imp_pls_nipals, polypls, polyplspred,
%                 stretchy, avgrows, errorstats,

% 1. Process input arguments
% Check for enough arguments and supply default values.
if nargin<7; meancentre=0; end
if nargin<6; nout=1; end
if nargin<5; nrep=1; end
if nargin<4
    error('At least four input arguments are required!');
end

% Check for vector a if using qpcr
if strcmp(regmethod,'polypcr') | strcmp(regmethod, 'polypls');
    if length(A)<2;
        error('For polypcr or polypls A must have two elements.');
    end
    d=A(1); A(1)=[];
end

% Get sizes
[nx m]=size(X); [ny c]=size(Y);

% Special case where Y has one entry per sample instead of per spectrum
if nargin>5 & nx==(ny*nrep) & length(nrep)==1
    warning('Assuming Y has one entry per sample...')
    Y=stretchy(Y,nrep); n=ny*nrep;
else n=ny;
end

% Standardise nrep to be a column vector of sample sizes.
if length(nrep)>1 % Vector nrep supplied: check it matches
    if sum(nrep)~=n
       error('Vector nrep doesn''t match data (sum(nrep)~=n)!')
    end
    nsam=length(nrep);
    if size(nrep,2)>size(nrep,1); nrep=nrep'; end
else % Constant nrep; check and convert to vector
    nsam=n/nrep;
    if rem(nsam,1)
        error('n is not an even multiple of scalar nrep.');
    end
    nrep=ones(nsam,1)*nrep;
end

if rem(nsam,nout)
    warning('number of samples is not an even multiple of nout.')
end
```

```
    %2. Set up variables for the leave-some-out loop
99
    % List of start indices (in terms of spectra) for samples
    saminds=[1;cumsum(nrep(1:end-1))+1];
    % Number of groups to be left out
    nloops=floor(nsam/nout);
104 % Initialise output arrays
    Yp=zeros(n,A,c);
    Xr=zeros(n,A);
    Xp=zeros(n,m,A);

109 % 3. Carry out the validation
    for ii=1:nloops
        % Get indices of spectra to leave out
        % jj is position in terms of samples, indout is indices of spectra
        %   to omit
114     jj=(ii-1)*nout+1;
        if ii<nloops
            indout=saminds(jj):saminds(jj+nout)-1;
        else
            indout=saminds(jj):n;
119     end
        % Create submatrices
        Xnew=X; Xnew(indout,:)=[]; Ynew=Y; Ynew(indout,:)=[];
        % Put left-out X spectra in a new matrix to simplify mean-centring
        Xout=X(indout,:);
124
        if meancentre
            meanX=mean(Xnew); meanY=mean(Ynew);
            Xnew=Xnew-repmat(meanX,[size(Xnew,1),1]);
            Xout=Xout-repmat(meanX,[size(Xout,1),1]);
129         Ynew=Ynew-repmat(mean(Y),[size(Ynew,1),1]);
        end

        % For each algorithm, there is code to:
        % 1. Calculate the model
134     % 2. Determine Yp, Xp, and Xr
        switch regmethod
            case {'pls','nipals'}
                switch regmethod
                    case 'pls'; [W P Q]=imp_pls_k1(Xnew,Ynew,A);
139                 case 'nipals'; [W P Q]=imp_pls_nipals(Xnew,Ynew,A);
                end
                Tp=Xout*W*inv(P'*W);
                for kk=1:A
                    Yp(indout,kk,:)=Tp(:,1:kk)*Q(:,1:kk)';
144                 if nargout>6
                        Xpred=Tp(:,1:kk)*P(:,1:kk)';
                        Xr(indout,kk)=sqrt(sum((Xout-Xpred).^2,2));
                        Xp(indout,:,kk)=Xpred;
                    end
149             end
            case 'pcr'
                [T S P]=svds(Xnew,A); T=T*S;
                B=T\Ynew;
                Tp=Xout*P;
154             for kk=1:A
                    Yp(indout,kk,:)=Tp(:,1:kk)*B(1:kk,:);
                    if nargout>6
                        Xpred=Tp(:,1:kk)*P(:,1:kk)';
                        Xr(indout,kk)=sqrt(sum((Xout-Xpred).^2,2));
159                     Xp(indout,:,kk)=Xpred;
```

```
                end
            end
        case 'polypcr'
            [T S P]=svds(Xnew,A); T=T*S;
            for kk=1:A
                T2=[ones(size(T,1),1) T(:,1:kk)];
                for ll=2:d;
                    T2=[T2 T(:,1:kk).^ll];
                end
                B=T2\Ynew;
                Tp=Xout*P;
                Tp2=[ones(size(Xout,1),1) Tp(:,1:kk)];
                for ll=2:d;
                    Tp2=[Tp2 Tp(:,1:kk).^ll];
                end

                Yp(indout,kk,:)=Tp2*B;
                if nargout>6
                    Xpred=Tp(:,1:kk)*P(:,1:kk)';
                    Xr(indout,kk)=sqrt(sum((Xout-Xpred).^2,2));
                    Xp(indout,:,kk)=Xpred;
                end
            end
        case 'polypls'
            [P,Q,W,T,U,b]=polypls(Xnew,Ynew,A,d);
            for kk=1:A
                [Yp(indout,kk,:) Xpred]=polyplspred(Xout, b(:,1:kk),...
                                P(:,1:kk), Q(:,1:kk), W(:,1:kk), kk);
                if nargout>6
                    Xr(indout,kk)=sqrt(sum((Xout-Xpred).^2,2));
                    Xp(indout,:,kk)=Xpred;
                end
            end
    end
    % Un-mean-centre Yp if necessary
    if meancentre
        for kk=1:c
            Yp(indout,:,kk)=Yp(indout,:,kk)+meanY(kk);
        end
    end
end

% 4. Calculate the prediction error, averaged predictions, and
%    determination coefficient.
for ii=1:c
    [RMSECV(ii,:) R2(ii,:) RMSECVavg(ii,:) R2avg(ii,:)]= ...
        errorstats(Y(:,ii), Yp(:,:,ii),nrep);
    Ypavg(:,:,ii)=avgspec(Yp(:,:,ii),nrep);
end
```

IMP_PLS_K1

```
function [W,P,Q,beta]=imp_pls_k1(X,Y,A)
% function [W P Q beta]=imp_pls_k1(X,Y,A)
%   Improved PLS kernel algorithm no. 1 from Dayal and MacGregor,
%   J. Chemom. 11, 73-85 (1997).
%   Inputs:
%       X = matrix of predictor variables (N by M)
%       Y = matrix of response variables (N by K)
%       A = number of factors
%   Outputs:
```

```
%       W, P, Q = weights, X and Y loadings
%       beta = regression vector
%
% To predict, use Y=X*beta (= X*W*inv(P'*W)*Q')

[N K]=size(Y);
M=size(X,2);
[W R P]=deal(zeros(M,A));
Q=zeros(K,A);

XY=X'*Y;

for ii=1:A
    if K==1 % PLS-1
        w=XY;
    else    % PLS-2
        [C,D]=eig(XY'*XY);
        q=C(:,find(diag(D)==max(diag(D))));
        w=XY*q;
    end
    w=w./norm(w);
    r=w;
    for jj=1:ii-1
        r=r-(P(:,jj)'*w)*R(:,jj);
    end
    t=X*r;
    tt=(t'*t);
    p=(X'*t)./tt;
    q=(r'*XY)'./tt;
    XY=XY-(p*q').*tt;
    W(:,ii)=w;
    P(:,ii)=p;
    Q(:,ii)=q;
    R(:,ii)=r;
end

if nargout>3; beta=R*Q'; end
```

IMP_PLS_NIPALS

```
function [W,P,Q,beta,T,U]=imp_pls_nipals(X,Y,A,convcrit,maxits)
% function [W P Q beta]=imp_pls_nipals(X,Y,A,convcrit,maxits)
%   Improved PLS NIPALS algorithm no. 1 from Dayal and MacGregor,
%   J. Chemom. 11, 73-85 (1997).
%   Inputs:
%       X = matrix of predictor variables (N by M)
%       Y = matrix of response variables (N by K)
%       A = number of factors
%       convcrit = convergence criterion (default 1e-13)
%       maxits = maximum iterations per factor (default 1000)
%   Outputs:
%       W, P, Q = weights, X and Y loadings
%       beta = regression vector
%       T, U = X and Y scores
%
% To predict, use Y=X*beta (= X*W*inv(P'*W)*Q')

% This version has some minor improvements and corrects the error on
% (their) line 23.

if nargin<5; maxits=1000;
```

```
if nargin<4; convcrit=1e-13; end; end

[N K]=size(Y);
M=size(X,2);

[W R P]=deal(zeros(M,A));
Q=deal(zeros(K,A));
[T U]=deal(zeros(N,A));

for ii=1:A
    u=Y(:,find(std(Y)==max(std(Y))));
    dif=1; its=0;
    while dif>convcrit
        u0=u;
        w=X'*u/(u'*u);
        w=w/sqrt(sum(w.^2));
        r=w;
        if ii>1
            for jj=1:ii-1
                r=r-(P(:,jj)'*w)*R(:,jj);
            end
        end
        t=X*r;
        q=Y'*t/(t'*t);
        u=Y*q/(q'*q);
        dif=sqrt(sum((u0-u).^2))/sqrt(sum(u.^2)); its=its+1;
        if its>maxits;
            warning(['Convergence failed for factor ', num2str(ii)]);
            break
        end
    end
    p=(t'*X/(t'*t))';
    W(:,ii)=w; R(:,ii)=r; P(:,ii)=p; Q(:,ii)=q; T(:,ii)=t; U(:,ii)=u;
    Y=Y-t*q';
end

if nargout>3; beta=R*Q'; end
```

PLS

This function wraps `IMP_PLS_K1` and calculates some other useful quantities.

```
function [model,stats]=mypls(X,Y,a,meancentre)
%function [model stats]=mypls(X,Y,a,meancentre)
%  Calculates a PLS model and some statistics.
%  Input:
%    X has spectra in rows, Y has concentrations
%    a is the number of factors to use
%    meancentre can be 1 or 0 (default)
%
%  Output:
%    model has fields W (weights), P, Q (loadings), T (scores) and B
%                     (regression vector)
%    stats has fields Xpred, Ypred, (predicted values)
%                     Xres, Yres,   (residuals)
%                     RMSEP         (no degrees of freedom correction)
%                     h             (leverage w/o centring term)
%
% Use PLSPRED for predictions.
```

```
model.type='pls';

if nargin<4; meancentre=0; end

if meancentre;
    model.meancentre=1;
    model.Xmean=mean(X);
    model.Ymean=mean(Y);
    X=msub(X, model.Xmean);
    Y=msub(Y, model.Ymean);
else
    model.meancentre=0;
end

[W P Q B]=imp_pls_k1(X,Y,a);
T=X*W*inv(P'*W);

model.P=P; model.W=W; model.Q=Q; model.B=B; model.T=T;
model.X=X; model.Y=Y;

% The leverage doesn't include term for centering error
stats.h=diag(T*inv(T'*T)*T');
stats.Xpred=T*P';
stats.Ypred=X*B;
stats.Yres=Y-stats.Ypred;
stats.Xres=X-stats.Xpred;
if meancentre
    stats.Ypred=madd(stats.Ypred, model.Ymean);
    stats.Xred=madd(stats.Xpred, model.Xmean);
end
stats.RMSEP=sqrt(sum(stats.Ycalresid.^2)./size(Y,1));
```

PLSPRED

```
function [Ypred,stats]=pls2pred(X,model,a)
% function [Ypred,stats]=pls2pred(X,model,a)
%   Predicts concentrations for spectra X using the PLS struct, model.
%   If a third argument is given, only the first a PLS factors are
%   used.
%
%   The output struct stats has the following fields:
%       Xres = spectroscopic residuals
%       h = leverage (with no centring term)

meancentre=model.meancentre;

W=model.W; P=model.P; Q=model.Q; Tcal=model.T;

if meancentre
    X=msub(X,model.Xmean);
end

if nargin>=3;
    W=W(:,1:a); P=P(:,1:a); Q=Q(:,1:a); Tcal=Tcal(:,1:a);
end

T=X*W*inv(P'*W);
Ypred=T*Q';

stats.Xres=X-T*P';
```

```
stats.h=diag(T*inv(Tcal'*Tcal)*T');

if meancentre
    Ypred=madd(Ypred,model.Ymean);
end
```

POLYPLS

```
function [P,Q,W,T,U,b,Xr,Yr]=polypls(X,Y,A,n)
% function [P,Q,W,T,U,b,Xr,Yr]=polypls(X,Y,A,n)
%   Polynomial PLS.
%   Inputs:
%       X = spectra in rows
%       Y = concentrations
%       n = polynomial order
%       A = number of PLS factors
%   Outputs:
%       W, P, Q, T, U = as in PLS (but actual values will be slightly
%                       different)
%       b = polynomial coefficients (one column per factor)
%       Xr, Yr = X and Y residuals
%
%   Make predictions for new standards with POLYPLSPRED.

% For each factor:
%   1. Calculate the PLS factor via NIPALS
%   2. Determine the polynomial 'inner relationship' b
%   3. Update X in the usual fashion
%   4. Update Y in accordance with the polynomial inner relationship

for ii=1:A
    [W(:,ii) P(:,ii) Q(:,ii) z T(:,ii) U(:,ii)]=imp_pls_nipals(X,Y,1);
    b(:,ii)=polyfit(T(:,ii),U(:,ii),n)';
    X=X-T(:,ii)*P(:,ii)';
    Y=Y-polyval(b(:,ii),T(:,ii))*Q(:,ii)';
end

Xr=X; Yr=Y;
```

POLYPLSPRED

```
function [Yp, Xp]=polyplspred(X,b,P,Q,W,A)
% function [Yp, Xp]=polyplspred(X,b,P,Q,W,A)
%   Prediction for polynomial PLS
%   Inputs
%       X = spectra in rows
%       W, P, Q, b are from POLYPLS
%       A = number of factors to use
%   Outputs
%       Yp = Predicted concentrations
%       Xp = Fitted spectra

if nargin<6
    A=size(W,2);
    if nargin<5
        error('At least five input arguments are required')
    end
end

Yp=0; X0=X;
```

```
for ii=1:A
    T(:,ii)=X*W(:,ii);
    X=X-T(:,ii)*P(:,ii)';
    Yp=Yp+polyval(b(:,ii),T(:,ii))*Q(:,ii)';
end

Xp=X0-X;
```

ERRORSTATS

```
function [RMSE,R2,RMSEavg,R2avg]=errorstats(Ytrue,Ypred,nrep)
% function [RMSE R2 RMSEavg R2avg]=errorstats(Ytrue,Ypred,nrep)
%
% Calculates RMSE and R2 for both individual and averaged predictions
% (where nrep is the number of replicates per sample, and may be a
% vector). Ytrue may be a vector or a matrix in which each column is a
% variable and each row an observation. If Ytrue is a vector then
% Ypred may have any number of columns; otherwise it must have the same
% dimensions as Ytrue.

% 1. Process input arguments and check for consistency
if nargin<3
    if nargin<2; error('Two input arguments are required.'); end
    nrep=1;
end

%   Check for consistent array dimensions
[rows1 cols1]=size(Ytrue);
[rows2 cols2]=size(Ypred);
if (rows1~=rows2) | (cols1>1 & cols2~=cols1)
    error(['Invalid arrays: Ytrue is ', ...
           num2str(rows1), 'x', num2str(cols1),...
           ' while Ypred is ',...
           num2str(rows2), 'x', num2str(cols2), '.'])
end

%   Make nrep a vector and check for consistency
if length(size(nrep))>2 | min(size(nrep))>1
    error('nrep must be a scalar or a vector.')
end
if length(nrep)==1
    if rem(rows1,nrep);
        error('Rows of Y not a multiple of nrep'); end
    nrep=nrep*ones(rows1/nrep,1);
else
    if sum(nrep)~=rows1;
        error('sum of nrep does not equal rows of Y'); end
    if size(nrep,2)>size(nrep,1); nrep=nrep'; end
end

% 2. Calculate the non-averaged prediction statistics
if cols1>1; residuals=Ytrue-Ypred;
else;       residuals=-msub(Ypred,Ytrue);
end
PRESS=sum(residuals.^2);
RMSE=sqrt(PRESS/rows1);
R2=1-PRESS./sum(msub(Ytrue,mean(Ytrue)).^2);

% 3. Calculate the averaged prediction statistics
Ytrueavg=avgspec(Ytrue,nrep);
residualsavg=avgspec(residuals,nrep);
```

```
PRESSavg=sum(residualsavg.^2);
RMSEavg=sqrt(PRESSavg./length(nrep));
R2avg=1-PRESSavg./sum(msub(Ytrueavg,mean(Ytrueavg)).^2);
```

JK_CI

```
function [Yu, sYu, Yu_jk]=jk_ci(X,Y,Xu,A,nrep)
% function [Yu, sYu, Yu_bs]=jk_ci(X,Y,Xu,A,B,nrep)
%   Jack-knife estimate of PLS prediction uncertainty.
%
%   Inputs:  X, Y    Calibration spectra and concentrations
%            Xu      New spectra with unknown concentrations
%            A, B    PLS rank and number of bootstrap iterations
%   Outputs: Yu, sYu Predicted concentrations and standard deviations
%            Yu_jk   All the jack-knife-predicted concentrations
%
% NB does not currently support mean centring.

% Uses: expind.m, imp_pls_k1.m

n=size(X,1);
nu=size(Xu,1);
Yu_jk=zeros(nu,n);

if length(nrep)>1
    ns=length(nrep); % Number of calibration samples
else
    ns=n/nrep;
end

[W P Q beta]=imp_pls_k1(X,Y,A);
Yu=Xu*beta;

for ii=1:n
    Xnew=X; Ynew=Y;
    if ns~=n
        I=expind(ii,nrep); Xnew(I,:)=[]; Ynew(I)=[];
    else
        Xnew(ii,:)=[]; Ynew(ii)=[];
    end
    [W P Q beta]=imp_pls_k1(Xnew,Ynew,A);
    Yu_jk(:,ii)=Xu*beta;
end

muYu_jk=mean(Yu_jk,2);
R_jk=(muYu_jk*ones(1,n) - Yu_jk).^2;

sYu=sqrt((n-1)/n * sum(R_jk,2)/n);
```

BSOBJ_CI

```
function [Yu, sYu, Yu_bs]=bsobj_ci(X,Y,Xu,A,B,nrep)
% function [Yu, sYu, Yu_bs]=bsobj_ci(X,Y,Xu,A,B,nrep)
%   Use object bootstrapping to estimate PLS prediction uncertainty.
%
%   Inputs:  X, Y    Calibration spectra and concentrations
%            Xu      New spectra with unknown concentrations
%            A, B    PLS rank and number of bootstrap iterations
%            nrep    Number of spectra per sample (vector or scalar)
%   Outputs: Yu, sYu Predicted concentrations and standard deviations
```

```
%             Yu_bs    All the bootstrap-predicted concentrations
%
% NB does not currently support mean centring.

% uses: expind.m, imp_pls_k1.m

if nargin<6; nrep=1; end

n=size(X,1);          % Number of calibration spectra
nu=size(Xu,1);        % Number of new spectra

if length(nrep)>1
    ns=length(nrep); % Number of calibration samples
else
    ns=n/nrep;
end

Yu_bs=zeros(nu,n);

[W P Q beta]=imp_pls_k1(X,Y,A);
Yu=Xu*beta;

for ii=1:B % Pick ns samples with replacement from the calibration set
    I=ceil(rand(ns,1)*ns);
    if ns~=n; I=expind(I,nrep); end
    Xnew=X(I,:); Ynew=Y(I);
    [W P Q beta]=imp_pls_k1(Xnew,Ynew,A);
    Yu_bs(:,ii)=Xu*beta;
end

sYu=std(Yu_bs,0,2);
```

BSRES_CI

```
function [Yu, sYu, Yu_bs]=bsres_ci(X,Y,Xu,A,B)
% function [Yu, sYu, Yu_bs]=bsres_ci(X,Y,Xu,A,B,nrep)
%   Use residual bootstrapping to estimate PLS prediction uncertainty.
%
%   Inputs:  X, Y    Calibration spectra and concentrations
%            Xu      New spectra with unknown concentrations
%            A, B    PLS rank and number of bootstrap iterations
%   Outputs: Yu, sYu Predicted concentrations and standard deviations
%            Yu_bs   All the bootstrap-predicted concentrations
%
% NB does not currently support mean centring.

% uses: imp_pls_k1.m

n=size(X,1);
nu=size(Xu,1);
Yu_bs=zeros(nu,n);

[W P Q beta]=imp_pls_k1(X,Y,A);
Yu=Xu*beta;

% Calibration-set residuals
Yres=(X*beta-Y)/(1-A/n);

for ii=1:B
    I=ceil(rand(n,1)*n);
    Ynew=Y+Yres(I);
```

```
    [W P Q beta]=imp_pls_k1(X,Ynew,A);
    Yu_bs(:,ii)=Xu*beta;
end

sYu=std(Yu_bs,0,2);
```

PREDVAR

```
function [Ynew, varnew, neff]=predvar(Xnew, model, stats, varY, nY, ...
                                      varX, nX, varE);

% function [Ynew, varnew]=predvar(Xnew, model, stats, varY, ...
%                                 nY, varX, nX, varE)
%   Estimate concentrations and error variance via an approximate EIV
%       formula. See Faber et al, Anal. Chem. 70 2972-2982 (1998)
%   Inputs:
%     Xnew = new spectra
%     model, stats = structures from pls.m
%     varY, nY, varX, nX = error variances for reference concentrations
%       and spectra, and their degrees of freedom. Default = 0.
%     varE = variance of the residuals (will be estimated if not given)
%   Outputs:
%     Ynew, varnew = predicted concentrations and their error variances
%     neff = estimate of the number of degrees of freedom for varnew
%       (determined by Satterthwaite's rule).

% Uses: plspred.m, leverage.m,

A=size(model.P,2);                % Rank
[Ncal M]=size(stats.Xcalresid); % Numbers of spectra and wavelengths

[Ynew statsnew]=plspred(Xnew,model);
hnew=leverage(Xnew,model);

b=norm(model.B)^2; % Squared length of regression vector

% Sort out the variances if they are not supplied
if nargin<8
    if nargin<6; varX=0; nX=1;
        if nargin<4; varY=0; nY=1; end; end
    % Estimate the residual variance
    % In the limit of negligible varX and varY, varE reduces to
    % MSEPcal/(N-A)
    if model.meancentre
        varE=Ncal*stats.RMSEPcal.^2/(Ncal-A-1) - varY - b*varX;
    else
        varE=Ncal*stats.RMSEPcal.^2/(Ncal-A) - varY - b*varX;
    end
    if varE<0; varE=0; end
end

if model.meancentre
    varnew=(1/Ncal+hnew).*(varE + varY + b*varX) + varE + b*varX;
else
    varnew=          hnew.*(varE + varY + b*varX) + varE + b*varX;
end

% Calculate the effective degrees of freedom by Satterthwaite's rule.
% See Burdick and Graybill, Confidence Intervals on Variance Components
% Marcel Dekker, 1992 pp 29-30.
V1=(1+hnew)*varE; n1=Ncal-A;
```

```
V2=(1+hnew)*b*varX; n2=nX;
V3=hnew*varY; n3=nY;

neff=floor((V1+V2+V3).^2./(V1.^2./n1+V2.^2./n2+V3.^2./n3));
```

JCR

```
function [xpts,ypts]=jcr(x,y,a,n)
% function [xpts,ypts]=jcr(x,y,a,n)
% Calculates points to plot a joint confidence region.
%   {xpts,ypts} are points on the ellipse defining the joint confidence
%    region for the slope and intercept of the line fitted through
%   {x,y}, at the significance level a (default 0.05 for 95% CI). n is
%   the number of points to generate (default 100).
%
%   Uses: f_inv from the OctaveForge statistics toolbox (an equivalent
%   function is available in the Matlab statistics toolbox).

if nargin<4; n=100;
if nargin<3; a=0.05;
if nargin<2; error('Two arguments are required.'); end; end; end

N=length(x);
S=sum(x); Y=sum(y);
Q=sum(x.^2); L=sum(y.^2); P=sum(x.*y);
D=N*Q-S^2;
mhat=(N*P-S*Y)/D; bhat=(Q*Y-S*P)/D;
ssq=1/(N-2)*(L - Y^2/N - D/N*mhat^2);
F=f_inv(1-a,2,N-2);

% Ellipse is given by (Mandel & Linnig, Anal. Chem. 29 743 (1957)
%   N(b-bhat)^2 + 2S(b-bhat)(m-mhat) + Q(m-mhat)^2=2Fssq
%
% Much easier to plot in polar coordinates.
% See http://mathworld.wolfram.com/Ellipse.html
% Using the general quadratic formula
%   ax^2 + 2bxy + cy^2 + 2dx + 2fy + g = 0 with x=(b-bhat), y=(m-mhat)
% a=N; b=S; c=Q; d=f=0; g=-2Fssq
%
% Using b'=ax1 and a'=ax2 and phi=-phi reproduces Mandel & Linnig's
% Figure 2.

ax1=sqrt(2*(-2*F*ssq*S^2 + 2*N*Q*F*ssq) / ...
    ((S^2-N*Q)*((Q-N)*sqrt(1+4*S^2/(N-Q)^2)-(N+Q))) );
ax2=sqrt(2*(-2*F*ssq*S^2 + 2*N*Q*F*ssq) / ...
    ((S^2-N*Q)*((N-Q)*sqrt(1+4*S^2/(N-Q)^2)-(N+Q))) );
phi=-1/2*acot((Q-N)/(2*S));

theta=linspace(0,2*pi,n);
xpts=ax1*cos(theta)*cos(phi)-sin(phi)*ax2*sin(theta)+bhat;
ypts=ax1*cos(theta)*sin(phi)+cos(phi)*ax2*sin(theta)+mhat;
```

JCR_P

```
function a=jcr_p(x,y,b0,m0,tol,maxits)
% function p=jcr_p(x,y,b0,m0,tol,maxits)
%   Finds the significance level p for which the point (b0,m0) lies on
%   the perimeter of the joint confidence ellipse. See jcr.m.
%
%   Inputs:
```

```
%       x, y = points to fit the line through
%       b0, m0 = point to lie on the ellipse (default 0, 1)
%       tol = tolerance for convergence (default ~1e-8)
%       maxits = maximum number of iterations (default 1000)

if nargin<6; maxits=1000;
if nargin<5; tol=sqrt(eps);
if nargin<4; m0=1;
if nargin<3; b0=0; end; end; end; end

% Golden ratio
ratio=1.61803398874989;

N=length(x);
S=sum(x);
Q=sum(x.^2);
Y=sum(y);
L=sum(y.^2);
P=sum(x.*y);
D=N*Q-S^2;
mhat=(N*P-S*Y)/D;
bhat=(Q*Y-S*P)/D;
ssq=1/(N-2)*(L - Y^2/N - D/N*mhat^2);


% All the points at constant theta lie on a line from the centre
ax1=1/sqrt(((S^2-N*Q)*((Q-N)*sqrt(1+4*S^2/(N-Q)^2)-(N+Q))));
ax2=1/sqrt(((S^2-N*Q)*((N-Q)*sqrt(1+4*S^2/(N-Q)^2)-(N+Q))));
phi=-1/2*acot((Q-N)/(2*S));

theta=atan( ax1/ax2 *((m0-mhat)*cos(phi) - (b0-bhat)*sin(phi)) / ...
                     ((m0-mhat)*sin(phi) + (b0-bhat)*cos(phi)) );

% Need to pick the right quadrant
if theta<0
    if (m0-mhat)>(b0-bhat)*tan(phi)
        theta=theta+pi;
    elseif (m0-mhat)<(b0-bhat)*tan(phi)
        theta=theta+2*pi;
    end
elseif theta>0
    if (m0-mhat)>(b0-bhat)*tan(phi)
        theta=theta;
    elseif (m0-mhat)<(b0-bhat)*tan(phi)
        theta=theta+pi;
    end
else % This test isn't quite right, but is unlikely to matter
    if (m0-mhat)>0
        theta=0;
    else
        theta=pi;
    end
end

% Now need to find the value of p that minimises the distance between
% (b,m) and (b0,m0), using the golden search. To start with, p is
% guaranteed to be bounded by 0 and 1.

q=0; qold=realmax; n=0;
left=tol; right=1-tol;

% "Golden search" for the minimum
```

```
    while abs(q-qold)>tol
        n=n+1;
71      qold=q;
        if n>maxits;
            warning('Maximum number of iterations exceeded!'); break;
        end
        q=(right+left*ratio)/(ratio+1);
76      q2=(right+q*ratio)/(ratio+1);
        [b1,m1]=getpoint(q,N,S,Q,Y,L,P,D,ssq,theta);
        [b2,m2]=getpoint(q2,N,S,Q,Y,L,P,D,ssq,theta);
        d1=(b1+bhat-b0)^2+(m1+mhat-m0)^2;
        d2=(b2+bhat-b0)^2+(m2+mhat-m0)^2;
81      if d2<d1;
            left=q; right=right;
        else
            left=left; right=q2;
        end
86  end

    p=q;

    function [b,m]=getpoint(a,N,S,Q,Y,L,P,D,ssq,theta)
91      F=f_inv(1-a,2,N-2);
        ax1=sqrt(2*(-2*F*ssq*S^2 + 2*N*Q*F*ssq) / ...
            ((S^2-N*Q)*((Q-N)*sqrt(1+4*S^2/(N-Q)^2)-(N+Q))) );
        ax2=sqrt(2*(-2*F*ssq*S^2 + 2*N*Q*F*ssq) / ...
            ((S^2-N*Q)*((N-Q)*sqrt(1+4*S^2/(N-Q)^2)-(N+Q))) );
96      phi=-1/2*acot((Q-N)/(2*S));
        b=ax1*cos(theta)*cos(phi)-sin(phi)*ax2*sin(theta);
        m=ax1*cos(theta)*sin(phi)+cos(phi)*ax2*sin(theta);
```

C.2.2 Optics functions

REFL3

```
 1  function [rptot, rstot, R] = refl3(n0, n1, n2, theta, dol)

    %function [rptot, rstot, R] = refl3(n0, n1, n2, theta, dol)
    %
    %   Calculates the (complex) amplitude reflection coefficients for
 6  %   light incident on a thin film/substrate system.
    %
    %   Inputs:
    %    n0, n1, n2    Refractive indices of the incident, film, and
    %                  substrate materials respectively.  n1 and n2 may be
11  %                  complex (n + ik).
    %    theta         Incidence angle in degrees (at normal incidence,
    %                  theta = 0).
    %    dol           Ratio of the film thickness to the radiation
    %                  wavelength.
16  %
    %   Each argument may be a scalar or an array, but all arrays must be
    %   of common size.
    %
    %    Outputs:
21  %     rp   Fresnel coefficients for p-polarisation (transverse magnetic)
    %     rs   Fresnel coefficients for s-polarisation (transverse electric)
    %     R    Reflectance for unpolarised light (R=0.5(|rp|^2 + |rs|.^2))
    %
    %   References:
26  %     1. Pedrotti, F. L. and Pedrotti, L.S., Introduction to Optics
```

```
%       (Second Edition), Prentice Hall NJ, 1993 pp 422
%    2. Heavens, O. S., Optics of Thin Solid Films, Dover NY, 1991
%       pp 57

% Convert theta to radians
theta=theta * pi/180;

% Cosines of the complex 'propagation angles' in the three media
costheta0=cos(theta);
costheta1=sqrt(1-n0.^2./n1.^2.*sin(theta).^2);
costheta2=sqrt(1-n0.^2./n2.^2.*sin(theta).^2);

% Fresnel reflection coefficients for the two interfaces
rp01=(n0.*costheta1-n1.*costheta0) ./ (n0.*costheta1+n1.*costheta0);
rp12=(n1.*costheta2-n2.*costheta1) ./ (n1.*costheta2+n2.*costheta1);
rs01=(n0.*costheta0-n1.*costheta1) ./ (n0.*costheta0+n1.*costheta1);
rs12=(n1.*costheta1-n2.*costheta2) ./ (n1.*costheta1+n2.*costheta2);

% Phase change on passing through the film
exp2iphi=exp(4*i*pi.*dol.*n1.*costheta1);

% Overall Fresnel coefficient
rstot=(rs01+rs12.*exp2iphi)./(1+rs01.*rs12.*exp2iphi);
rptot=(rp01+rp12.*exp2iphi)./(1+rp01.*rp12.*exp2iphi);

% Calculate R
if nargout>2; R=0.5*(abs(rptot).^2 + abs(rstot).^2); end
```

FIELDS3

```
function F=fields3(n1,n2,n3,h,l,theta,z,k)
% function F=fields3(n1,n2,n3,h,l,theta,z,k)
%   Calculates mean square electric field amplitudes at depth z in a
%   three-phase system (see also HANSEN3).
%     n1, n2, n3 are incident, film, substrate refractive indices
%     h is the film thickness, in the same units as l
%     l is the wavelength, in the same units as h
%     z is the depth below the incident medium/film interface:
%       z=0 is the incident medium/film interface
%       z<0 is in the incident medium
%       0<z<h is in the film itself
%       z>h is in the substrate
%     k is the phase: 1 incident, 2 film, 3 substrate.

% HANSEN3 provides more output than REFL3
X=hansen3(n1,n2,n3,h,l,theta);

theta=theta*pi/180;

pizl=4*pi*z/l;

if k==1 % Incident medium
    Esy = 1 + X.Rs + 2*sqrt(X.Rs)*cos(X.drs-pizl*X.eta1);
    Epx = cos(theta)^2*(1+X.Rp-2*sqrt(X.Rp)*cos(X.drp-pizl*X.eta1));
    Epz = sin(theta)^2*(1+X.Rp+2*sqrt(X.Rp)*cos(X.drp-pizl*X.eta1));
elseif k==2 % Film... bah! No equations...
    Es=(1+X.rs)*cos(X.eta2*pizl/2)+i*X.eta1/X.eta2*(1-X.rs)* ...
                                              sin(X.eta2*pizl/2);
    Esy=abs(Es).^2;
    Epx=cos(theta)*(1-X.rp)*cos(X.eta2*pizl/2) + ...
        i*X.eta2/n2^2*n1*(1+X.rp)*sin(X.eta2*pizl/2);
```

```
32      Epx=abs(Epx).^2;
        Epz=n1*sin(theta)/n2^2*n1*(1+X.rp)*cos(X.eta2*piz1/2); + ...
            n1*sin(theta)/X.eta2*cos(theta)*(1-X.rp)*sin(X.eta2*piz1/2);
        Epz=abs(Epz).^2;
    elseif k==3 % Substrate
37      Esy=abs(X.tEs)^2*exp(-4*pi*imag(X.eta3)*(z-h)/l);
        Epx=abs(X.eta3/n3*X.tEp)^2*exp(-4*pi*imag(X.eta3)*(z-h)/l);
        Epz=abs(n1*sin(theta)/n3*X.tEp)^2*exp(-4*pi*imag(X.eta3)*(z-h)/l);
    else error('Invalid k (must be 1, 2 or 3).')
    end
42
    F.Esy=Esy; F.Epx=Epx; F.Epz=Epz;
```

HANSEN3

```
    function X=opt_hansen_3layer(n1,n2,n3,h,l,theta)
 2  % function X=opt_hansen_3layer(n1,n2,n3,h,l,theta)
    %  Calculates optical properties for a three-layer system
    %   Inputs:
    %      n1  refractive index (real) of the incident medium
    %      n2  refractive index (complex; n+ik) of the film material
 7  %      n3  refractive index (complex) of the substrate
    %      h   film thickness
    %      l   wavelength (in same units as h)
    %      theta incidence angle (in degrees)
    %   Outputs: structure X with fields
12  %      eta1, eta2, eta3 eta_j=N_j cos(theta_j)
    %      beta    (complex) phase change due to the film
    %      rs, tEs, rp, tHp, tEp   Fresnel coefficients
    %      Rs, Rp  reflectance
    %      Ts, Tp  transmittance
17  %      drs, drs    phase changes on reflection/transmission

    theta=theta*pi/180;

    % "Angle-dependent refractive index" terms
22  eta1=n1.*cos(theta);
    eta2=sqrt(n2.^2-n1.^2.*sin(theta).^2);
    eta3=sqrt(n3.^2-n1.^2.*sin(theta).^2);

    % Phase change on passing through the film
27  beta=2*pi*eta2.*h./l;

    % Fresnel coefficients for the two interfaces
    [r1s t1s]=fresnel(eta1,eta2);
    [r2s t2s]=fresnel(eta2,eta3);
32  [r1p t1p]=fresnel(n2.^2.*eta1,n1.^2.*eta2);
    [r2p t2p]=fresnel(n3.^2.*eta2,n2.^2.*eta3);

    % Overall Fresnel coefficients
    [rs tEs]=fresnel3(r1s,r2s,t1s,t2s,beta);
37  [rp tHp]=fresnel3(r1p,r2p,t1p,t2p,beta);
    tEp=n1./n3.*tHp;

    % Reflectance and transmittance
    Rs=abs(rs).^2;
42  Ts=real(eta3)./eta1.*abs(tEs).^2;
    Rp=abs(rp).^2;
    Tp=real(eta3./n3.^2)./(eta1./n1.^2).*abs(tHp).^2;

    % Phase changes for reflection and transmission
```

```
drs=angle(rs); drp=angle(rp);
dts=angle(tEs); dtp=angle(tEp);

% Dump all the interesting variables into a struct
varnames={'eta1', 'eta2', 'eta3', 'beta', 'rs', 'tEs', 'rp', 'tHp', ...
          'tEp', 'Rs', 'Ts', 'Rp', 'Tp', 'drs', 'dts', 'drp', 'dtp'};
for ii=1:length(varnames)
    eval(['X.',varnames{ii},'=',varnames{ii},';']);
end


% Function to calculate single-interface Fresnel coefficients
function [r,t]=fresnel(eta1,eta2)
    r=(eta1-eta2)./(eta1+eta2);
    t=2*eta1./(eta1+eta2);

% Function to calculate overall Fresnel coefficients
function [rtot,ttot]=fresnel3(r1,r2,t1,t2,beta)
    rtot=(r1+r2.*exp(2*i*beta)) ./ ...
         (1+r1.*r2.*exp(2*i*beta));
    ttot=t1.*t2.*exp(i*beta) ./ ...
         (1+r1.*r2.*exp(2*i*beta));
```

BREWSTER

```
function theta=brewster(n1,n2,tol,maxits)
% function theta=brewster(n1,n2)
%   Calculates the Brewster (or pseudo-Brewster) angle between phases
%   with refractive indices n1 and n2. theta is in degrees. n1 and n2
%   must be scalars. n1 should be real but n2 may be complex.
%
%   For real n2, theta=atan(n2/n1).
%   For complex n2, theta is the angle which minimises the reflectance
%   of p-polarised light (calculated by a golden section search with
%   tolerance tol and up to maxits iterations).
%
%   tol defaults to sqrt(eps) and maxits defaults to 1000.

if ~isreal(n1); warning('n1 should be real'); end

if isreal(n1) & isreal(n2)
    theta=atan(n2/n1)*180/pi;
else
    if nargin<4; maxits=1000;
    if nargin<3; tol=sqrt(eps); end; end
    ratio=1.61803398874989;
    q=0; qold=realmax; n=0;
    left=0; right=90;
    % "Golden search" for the minimum
    while abs(q-qold)>tol
        n=n+1;
        qold=q;
        if n>maxits;
            warning('Maximum number of iterations exceeded!'); break;
        end
        q=(right+left*ratio)/(ratio+1);
        q2=(right+q*ratio)/(ratio+1);
        r1=refl(n1,n2,q); R1=abs(r1.^2);
        r2=refl(n1,n2,q2); R2=abs(r2.^2);
        if R2<R1;
            left=q; right=right;
```

```
        else
            left=left; right=q2;
        end
    end
    theta=q;
end

function rp=refl(n1,n2,theta)
    % Calculates the Fresnel reflection coefficient for p polarisation
    costheta1=cos(theta*pi/180);
    costheta2=sqrt(1-n1.^2./n2.^2.*sin(theta*pi/180).^2);
    rp=(n1.*costheta2-n2.*costheta1) ./ (n1.*costheta2+n2.*costheta1);
```

C.2.3 Utility functions

The following three functions (`STRETCHY`, `EXPIND` and `AVGROWS`) are to assist with converting between sample and spectrum indices. These operations are fairly trivial if every sample has the same number of spectra, but become awkward when the numbers of spectra differ from sample to sample. All of these functions refer to a vector `nrep`, which is simply a list of the number of spectra belonging to each sample.

STRETCHY

```
function Yspec=stretchy(Ysam,nrep)
%function Yspec=stretchy(Ysam,nrep)
%   If Ysam is a vector, each element is repeated nrep times to give
%   Yspec.  If Ysam is a matrix, then each row is repeated nrep times.
%   If nrep is a vector with length equal to the number of rows of
%   Ysam, then the iith element (or row) of Ysam is repeated nrep(ii)
%   times.
%
%   e.g. stretchy([1 2 3],3) returns [1 1 1 2 2 2 3 3 3]
%        stretchy([1 2; 3 4],2) returns [1,2; 1,2; 3,4; 3,4]
%        stretchy([1 2 3],[1 2 3]) returns [1 2 2 3 3 3].

[Ysam,iscolvec,ismat]=rowvec(Ysam);
[rows cols]=size(Ysam);

if length(nrep)==1
    if ismat
        Yspec=repmat(Ysam,1,nrep);
        Yspec=reshape(Yspec', cols, rows*nrep)';
    else
        Yspec=repmat(Ysam,nrep,1);
        Yspec=reshape(Yspec,1,[]);
        if iscolvec; Yspec=Yspec'; end
    end
elseif (ismat & length(nrep)==rows) | ...
                                (~ismat & length(nrep)==length(Ysam))
    Yspec=[];
    if ~ismat, Ysam=Ysam'; rows=length(Ysam); end
    for ii=1:rows
        Yspec=[Yspec; repmat(Ysam(ii,:),nrep(ii),1)];
    end
    if (~iscolvec & ~ismat), Yspec=Yspec'; end
else error('Length of nrep doesn''t match size of Ysam')
```

```
end

function [outvec, iscolvec, ismat]=rowvec(invec);
    [rows cols]=size(invec); ismat=0; iscolvec=0;
    if     rows==1; outvec=invec;
    elseif cols==1; outvec=invec'; iscolvec=1;
    else            outvec=invec; ismat=1;
    end
```

EXPIND

```
function specind=expind(samind, nrep)
% function specind=expind(samind, nrep)
%   Convert sample indices samind to spectral indices specind, where
%   there are nrep spectra per sample.  specind is a vector with
%   nrep*length(samind) elements.  If samind is a scalar then specind
%   is a row vector.
%
%   If there are differing numbers of spectra per sample, nrep must be
%   a vector.
%
%   e.g. expind([1,3,7]',3) returns [1;2;3;7;8;9;19;20;21];
%        expind(4,2) returns [7 8];
%        expind([1 2 4], [4 3 2 1]) returns [1 2 3 4 5 6 7 10].

if isempty(samind); specind=[]; return; end

[specind,iscolvec,ismat]=rowvec(samind);
if ismat; error('samind must be a vector.'); end

if length(nrep)==1;
    specind=(specind-1)*nrep;
    specind=repmat(specind, [nrep,1]);
    specind=madd(specind, [1:nrep]');
    specind=reshape(specind, 1,[]);
else
    nrep=rowvec(nrep); specind=[];
    allsamind=[1 cumsum(nrep(1:end-1))+1];
    for ii=1:length(samind)
        specind=[specind, (allsamind(samind(ii))+ ...
                                          [0:nrep(samind(ii))-1])];
    end
end

if iscolvec; specind=specind'; end

function [outvec, iscolvec, ismat]=rowvec(invec);
   [rows cols]=size(invec); ismat=0; iscolvec=0;
   if     rows==1; outvec=invec;
   elseif cols==1; outvec=invec'; iscolvec=1;
   else            outvec=invec; ismat=1;
   end
```

AVGROWS

```
function Xavg=avgrows(X,nrep)
%function Xavg=avgrows(X,nrep)
%   Calculates average spectra.  The first row of Xavg is the mean of
%   the first nrep rows of X.  If nrep is a vector such that
%   sum(nrep) == number of rows of X, then the iith average is the
```

```
%   average of nrep(ii) spectra.
%
%   e.g. avgspec([1.1 1.9; 0.9 2.1; 4.3 6.2; 3.7 5.8],2) returns
%   [1 2; 4 6].

% If nrep is a scalar, convert it to a vector
[rows cols]=size(X);
if length(nrep)==1
    if rem(rows,nrep); error('rows not a multiple of nrep'); end
    nrep=nrep*ones(rows/nrep,1);
else
    if sum(nrep)~=rows;
        error('sum of nrep doesn''t equal rows of X');
    end
    if size(nrep,2)>size(nrep,1); nrep=nrep'; end
end

Xavg=zeros(length(nrep),cols);
samind=[1; cumsum(nrep(1:end-1))+1];

for ii=1:length(nrep)
    Xavg(ii,:)=mean(X(samind(ii):samind(ii)+nrep(ii)-1,:),1);
end
```

SMILEY

```
% Draws a three-dimensional smiley face for Katy
% 1. Set up a grid
n=101;
x=linspace(-1,1,n); y=x;
[X Y]=meshgrid(x,y);

% 2. Polar coordinates are easier for the mouth
R=sqrt(X.^2+Y.^2);
T=atan(X./Y);

% 3. Generate the various parts of the face
Z1=real(sqrt(1-X.^2-Y.^2));                    % head
Z2=real(sqrt(0.05-X.^2-Y.^2));                 % nose
Z3=real(sqrt(0.05-(X-0.4).^2-(Y-0.4).^2));     % left eye
Z4=real(sqrt(0.05-(X+0.4).^2-(Y-0.4).^2));     % right eye
Z5=0.6*exp(-(R-0.7).^2*250).*exp(-T.^2*2);     % mouth
Z5(ceil(n/2):n,:)=0;

% 4. Plot the face and set the viewpoint
Z=Z1+Z3+Z4-Z5;
H=surf(X,Y,Z);
axis tight; grid off;
set(gca,'view',[-16,80]); set(H,'facecolor',[1 1 1]);
```

C.3 Other code listings

C.3.1 OPUS laser wavenumber calibration

The following OPUS macro is modified from the instrument test suite laser wavenumber test macro to use the MCT detector. The macro `IT_LWN_MCT_3mmAP.MTX` sets the source aperture to 3mm, whereas `IT_LWN_MCT_OPENAP.MTX` leaves the source aperture open (the corresponding experiment

files IT_FR_MCT.XPM and IT_FR_MCT_3mmAP.XPM are required). The recommended laser wavenumber is calculated as

$$\mathrm{LWN}_{rec.} = \mathrm{LWN}_{current} \times \bar{\nu}_{ref}/\bar{\nu}_{meas} \tag{C.1}$$

where $\mathrm{LWN}_{current}$ is the wavenumber currently being used by the transform programme: if this is changed, the macro must be updated. The peak used is an H_2O absorption band at 1554.353 cm^{-1}. The macro just measures a high-resolution single channel spectrum with weak apodisation, measures the peak position, and calculates the laser wavenumber required to put the peak in the right place.

```
VARIABLES SECTION

FILE <$ResultFile 1> = ScSm, ScSm/Peak;
*NUMERIC <old_wavenumber> = 15798.000000000000000;
NUMERIC <peak_position> = 0;
NUMERIC <true_peak> = 1554.353;
NUMERIC <recommended_lwn>=0;

PROGRAM SECTION

<$ResultFile 1> = MeasureSample (0, {EXP='IT_FR_MCT.XPM',
                                     XPP='C:\OPUS\Macro\ben'});
PeakPick ([<$ResultFile 1>:ScSm], {NSP=9, PSM=1, WHR=0,
          LXP=1555.000000, FXP=1553.000000, QP8='NO',
          QP9=0.200000, PTR=20.000000, QP4='NO', QP7=0.800000,
          QP6='NO', QP5=80.000000, PPM=3, QPA='OVR', QP0='NO',
          QP3=4, QPC='NO', QPD=3});
<peak_position> = FromReportMatrix ([<$ResultFile 1>:ScSm/Peak],
                                    1, 0, 1, 1);
<recommended_lwn>=<true_peak>/<peak_position> * <old_wavenumber>;
Message ('Reference peak <[,3]true_peak>;
          Measured peak <[,3]peak_position>;
          Recommended laser wavenumber <[,3]recommended_lwn> cm-1',
          ON_SCREEN, NO_TIMEOUT);

PARAMETER SECTION

NSP=9;
PSM=1;
WHR=0;
LXP=1555.000000;
FXP=1553.000000;
QP8=NO;
QP9=0.200000;
PTR=20.000000;
QP4=NO;
QP7=0.800000;
QP6=NO;
QP5=80.000000;
PPM=3;
QP0=NO;
QP3=4;
QPC=NO;
```

```
QPD=3;
```

C.3.2 Python J-CAMP reader

```
# dx2csv.py
#   Script to convert a subset of JCAMP-DX format files to
#   comma-separated variable format.  No error-checking whatsoever is
#   done.  It has been used to convert absorbance spectra saved as
#   JCAMP-DX by OPUS 4.2, GRAMS/AI 7, and Spectral.

import sys,re

if len(sys.argv) < 3:
    print 'Usage: python dx2csv.py dxfile.dx csvfile.csv'
    sys.exit()

# Regular expression to match a substring containing +, -, 0-9, e, E or
# ., which should have to be a number in a .dx file.
numre=re.compile(r'[-\+\d\.eE]+')

# Patterns to look for in the comment section of the file and the
# corresponding variables.
matchlist=['##FIRSTX','##LASTX','##XFACTOR','##YFACTOR','##NPOINTS']
varlist=['firstx=float','lastx=float','xfactor=float','yfactor=float',
         'npoints=int']

dxfile=open(sys.argv[1],"r")
while 1:
    # Check each line for each of the patterns
    s = dxfile.readline()
    if re.match('##XYDATA',s):
        break
    count=0
    for a in matchlist:
        if re.match(a,s):
            # Set the value of the appropriate variable
            execstr=varlist[count]+\
                    '(s[numre.search(s).start():numre.search(s).end()])'
            exec(execstr)
            continue
        count=count+1

# Calculate the x interval
deltax = (lastx-firstx)/(npoints-1)

# Calculate all the x values
X=[]
for n in range(npoints):
    X=X+[firstx+n*deltax]

# Read all the y values from the ##XYDATA section of the file. Each y
# value consists of a + (or space) or a - followed by an integer, and
# there are several strung together on a line.
```

```
    # The first number on a line is an abscissa value, and might not be an
    # integer.
52  yre=re.compile(r'[-\+ ]?[\d\.]+')
    Y=[]
    while 1:
        s=dxfile.readline()
        count=0
57      if s[0]=='#':
            break
        for m in yre.finditer(s):
            if count>0:
                Y=Y+[float(s[m.start():m.end()])*yfactor]
62          count=count+1

    dxfile.close()

    # Write the CSV file
67  csvfile=open(sys.argv[2],'w')
    for n in range(npoints):
        wstr=str(X[n])+','+str(Y[n])+'\n'
        csvfile.write(wstr)
    csvfile.close()
```

References

[1] V. P. Tolstoy, I. V. Chernyshova, and V. A. Skryshevsky. *Handbook of Infrared Spectroscopy of Ultrathin Films*. Wiley, Hoboken, New Jersey, 2003.

[2] United States Food and Drug Administration. *Guide to inspections: Validation of cleaning processes*. 1993.

[3] Andreas O. Zeller. Cleaning validation and residue limits: a contribution to current discussions. *Pharmaceutical Technology*, 17(10):70–80, 1993.

[4] K. M. Jenkins and A. J. Vanderwielen. Cleaning validation: An overall perspective. *Pharmaceutical Technology*, 18(4):60–73, 1994.

[5] Gamal Amer and Praful Deshmane. Ensuring successful validation: The logical steps to efficient cleaning procedures. *Biopharm*, 14(3):26–32, 2001.

[6] Active Pharmaceutical Ingredients Committee. Cleaning validation in active pharmaceutical ingredient manufacturing plants, 1999. Available from http://apic.cefic.org/pub/4CleaningVal9909.pdf (current Aug. 2006).

[7] Gary L. Fourman and Michael V. Mullen. Determining cleaning validation acceptance limits for pharmaceutical manufacturing operations. *Pharmaceutical Technology*, 17(4):54–60, 1993.

[8] Destin A. LeBlanc. Establishing scientifically justified acceptance criteria for cleaning validation of finished drug products. *Pharmaceutical Technology*, 22(10):136–148, 1998.

[9] Destin A. LeBlanc. Establishing scientifically justified acceptance criteria for the cleaning validation of APIs. *Pharmaceutical Technology*, 24(10):160–168, 2000.

[10] Herbert J. Kaiser. Methods for pharmaceutical cleaning validation. In K. L. Mittal, editor, *Surface Contamination and Cleaning*, volume 1, pages 75–84. VSP, 2003.

[11] Destin A. LeBlanc. Rinse sampling for cleaning validation studies. *Pharmaceutical Technology*, 22(5):66–74, 1998.

[12] Michael J. Shifflet and Mark Shapiro. Development of analytical methods to accurately and precisely determine residual active pharmaceutical ingredients and cleaning agents on pharmaceutical surfaces. *American Pharmaceutical Review*, 5(2):35–41, 2002.

[13] Pei Yang, Kim Burson, Debra Feder, and Fraser Macdonald. Method development of swab sampling for cleaning validation. *Pharmaceutical Technology*, pages 84–94, January 2005.

[14] Destin A. LeBlanc. "Visually clean" as a sole acceptance criterion for cleaning validation protocols. *PDA Journal of Pharmaceutical Science & Technology*, 56(1):31–36, 2002.

[15] Richard J. Forsyth, Vincent Van Nostrand, and Gregory P. Martin. Visible residue limit for cleaning validation and its potential application in a pharmaceutical research facility. *Pharmaceutical Technology*, 28(10):58–71, 2004.

[16] Richard J. Forsyth and Vincent Van Nostrand. Application of visible-residue limit for cleaning validation in a pharmaceutical manufacturing facility. *Pharmaceutical Technology*, 29(10):152–161, 2005.

[17] Herbert J. Kaiser and Maria Minowitz. Analyzing cleaning validation samples: What method? *Journal of Validation Technology*, 7(3):226–236, 2001.

[18] Tahseen Mirza, Michael J. Lunn, Frederick J. Keeley, Ron C. George, and John R. Bodenmiller. Cleaning level acceptance criteria and a high pressure liquid chromatography procedure for the assay of meclizine hydrochloride residue in swabs collected from pharmaceutical manufacturing equipment surfaces. *Journal of Pharmaceutical and Biomedical Analysis*, 19:747–756, 1999.

[19] R. Klinkenberg, B. Streel, and A. Ceccato. Development and validation of a liquid chromatographic method for the determination of amlodipine residues on manufacturing equipment surfaces. *Journal of Pharmaceutical and Biomedical Analysis*, 32:345–352, 2003.

[20] James M. Smith. Selecting analytical methods to detect residue from cleaning compounds in validated process systems. *Pharmaceutical Technology*, 17:88–98, 1993.

[21] Destin A. LeBlanc. Sampling, analyzing, and removing surface residues found in pharmaceutical manufacturing equipment. *Microcontamination*, pages 37–40, 1993.

[22] Walter K. Gavlick, Lane A. Ohlemeier, and Herbert J. Kaiser. Analytical strategies for cleaning agent residue determination. *Pharmaceutical Technology*, pages 136–144, 1995.

[23] S. A. Francis and A. H. Ellison. Infrared spectra of monolayers on metal mirrors. *Journal of the Optical Society of America*, 49(2):131–139, 1959.

[24] Robert G. Greenler. Infrared study of adsorbed molecules on metal surfaces by reflection techniques. *Journal of Chemical Physics*, 44(1):310–315, 1966.

[25] Robert G. Greenler. Reflection method for obtaining the infrared spectrum of a thin layer on a metal surface. *Journal of Chemical Physics*, 50(5):1963–1968, 1969.

[26] Robert G. Greenler, Robert R. Rahn, and John P. Schwartz. The effect of index of refraction on the position, shape, and intensity of infrared bands in reflection-absorption spectra. *Journal of Catalysis*, 23:42–48, 1971.

[27] Robert G. Greenler. Design of a reflection-absorption experiment for studying the IR spectrum of molecules adsorbed on a metal surface. *Journal of Vacuum Science and Technology*, 12(6): 1410–1417, 1975.

[28] D. L. Allara, A. Baca, and C. A. Pryde. Distortions of band shapes in external reflection infrared spectra of thin polymer films on metal substrates. *Macromolecules*, 11(6):1215–1220, 1978.

[29] Richard A. Dluhy and Donald G. Cornell. In situ measurement of the infrared spectra of insoluble monolayers at the air-water interface. *Journal of Physical Chemistry*, 89:3195–3197, 1985.

[30] Richard A. Dluhy. Quantitative external reflection infrared spectroscopy analysis of insoluble monolayers spread at the air-water interface. *Journal of Physical Chemistry*, 90:1373–1379, 1986.

[31] Wilford N. Hansen. Electric fields produced by the propagation of plane coherent electromagnetic radiation in a stratified medium. *Journal of the Optical Society of America*, 58(3):380–390, 1968.

[32] Yuichi Ishino and Hatsuo Ishida. Spectral simulation of uniaxially oriented monolayers in the infrared. *Langmuir*, 4:1341–1346, 1988.

[33] Pochi Yeh. Electromagnetic propagation in birefringent layered media. *Journal of the Optical Society of America*, 69(5):742–756, 1979.

[34] Atul N. Parikh and David L. Allara. Quantitative determination of molecular structure in multilayered thin films of biaxial and lower symmetry from photon spectroscopies. I. Reflection infrared vibrational spectroscopy. *Journal of Chemical Physics*, 96(2):927–945, 1992.

[35] Ian J. Hodgkinson and Qi hong Wu. *Birefringent Thin Films and Polarizing Elements*. World Scientific, Singapore, 1997.

[36] O. S. Heavens. *Optical Properties of Thin Solid Films*. Dover, New York, 1991.

[37] Francis Graham-Smith and Terry A. King. *Optics and Photonics: An Introduction*. Wiley, Chichester, 2000.

[38] John E. Bertie. Optical constants. In John M. Chalmers and Peter R. Griffiths, editors, *The Handbook of Vibrational Spectroscopy*, volume 1, pages 88–100. Wiley, Chichester, 2002.

[39] John E. Bertie and C. Dale Keefe. Infrared intensities of liquids XXIV: Optical constants of liquid benzene-h_6 at 25 °C extended to 11.5 cm^{-1} and molar polarizabilities and integrated intensities of benzene-h_6 between 6200 and 11.5 cm^{-1}. *Journal of Molecular Structure*, 695-696: 39–57, 2004.

[40] H. R. Philipp. SiO_2 glass. In Edward D. Palik, editor, *Handbook of Optical Constants of Solids*, pages 749–763. Academic Press, Orlando, Florida, 1985.

[41] Warren J. Smith. *Modern Optical Engineering: The Design of Optical Systems*. Optical and Electro-Optical Engineering Series. McGraw Hill, Inc., New York, 1990.

[42] Sh. A. Furman and A. V. Tikhonravov. *Basics of: Optics of Multilayer Systems*. Editions Frontiéres, Paris, 1992.

[43] H. Angus Macleod. *Thin-Film Optical Filters*. Institute of Physics Publishing, Bristol, 2001.

[44] Wilford N. Hansen. Internal reflection spectroscopy in electrochemistry. In Rolf H. Muller, editor, *Optical Techniques in Electrochemistry*, volume 9 of *Advances in Electrochemistry and Electrochemical Engineering*, pages 1–60. Wiley, New York, 1973.

[45] D. Y. Smith, E. Shiles, and Mitio Inokuti. The optical properties of metallic aluminum. In Edward D. Palik, editor, *Handbook of Optical Constants of Solids*, pages 369–406. Academic Press, Orlando, Florida, 1985.

[46] J. A. Mielczarski. External reflection infrared-spectroscopy at metallic, semiconductor, and nonmetallic substrates. 1. Monolayer films. *Journal of Physical Chemistry*, 97(11):2649–2663, 1993.

[47] Eric W. Weisstein. Brewster's angle. In *Eric Weisstein's World of Physics*. . http://scienceworld.wolfram.com/physics/BrewstersAngle.html (Current Aug. 2006).

[48] Svante Wold. Chemometrics; what do we mean with it, and what do we want from it? *Chemometrics and Intelligent Laboratory Systems*, 30(1):109–115, 1995.

[49] Harald Martens and Tormod Næs. *Multivariate Calibration*. Wiley, Chichester, 1989.

[50] Kenneth R. Beebe, Randy J. Pell, and Mary Beth Seasholtz. *Chemometrics: A Practical Guide*. Wiley, New York, 1998.

[51] Richard R. Kramer. *Chemometric Techniques for Quantitative Analysis*. New York, 1998.

[52] K. Danzer, M. Otto, and L. A. Currie. Guidelines for calibration in analytical chemistry part 2. Multispecies calibration - (IUPAC Technical Report). *Pure and Applied Chemistry*, 76(6): 1215–1225, 2004.

[53] J. H. Kalivas. Multivariate calibration, an overview. *Analytical Letters*, 38(14):2259–2279, 2005.

[54] P. K. Hopke. The evolution of chemometrics. *Analytica Chimica Acta*, 500(1-2):365–377, 2003. Times Cited: 4.

[55] Paul Geladi. Some recent trends in the calibration literature. *Chemometrics and Intelligent Laboratory Systems*, 60:211–224, 2002.

[56] Douglas A. Skoog. *Principles of Instrumental Analysis*. CBS College Publishing, third edition, 1985.

[57] David M. Haaland. Multivariate calibration methods applied to quantitative FT-IR analyses. In John R. Ferraro and K. Krishnan, editors, *Practical Fourier Transform Infrared Spectroscopy*, pages 396–469. Academic Press, San Diego, 1990.

[58] The Mathworks. Matlab 6.5 release 13, 2002.

[59] John W. Eaton (and others). GNU Octave. http://www.gnu.org/software/octave/.

[60] Charles L. Lawson and Richard J. Hanson. *Solving Least Squares Problems*. Prentice-Hall, Englewood Cliffs, New Jersey, 1974.

[61] Gilbert Strang. *Linear Algebra and its Applications*. Harcourt Brace Jovanovich, Orlando, Florida, 1988.

[62] Edmund R. Malinowski. *Factor Analysis in Chemistry*. Wiley, New York, second edition, 1981.

[63] Edmund R. Malinowski. Theory of error in factor analysis. *Analytical Chemistry*, 49(4):606–612, 1977.

[64] Edmund R. Malinowski. Determination of the number of factors and the experimental error in a data matrix. *Analytical Chemistry*, 49(4):612–617, 1977.

[65] Edmund R. Malinowski. Abstract factor analysis of data with multiple sources of error and a modified Faber-Kowalski F-test. *Journal of Chemometrics*, 13:69–81, 1999.

[66] N. R. Draper and H. Smith. *Applied Regression Analysis*. Wiley, New York, second edition, 1981.

[67] Agnar Höskuldsson. PLS regression methods. *Journal of Chemometrics*, 2(3):211–228, 1988.

[68] Peter D. Wentzell and Lorenzo Vega-Montoto. Comparison of principal components regression and partial least squares regression through generic simulations of complex mixtures. *Chemometrics and Intelligent Laboratory Systems*, 65:257–279, 2003.

[69] Takeshi Hasegawa. Principal component regression and partial least squares modeling. In John M. Chalmers and Peter R. Griffiths, editors, *Handbook of Vibrational Spectroscopy*, volume 3, pages 2293–2312. Wiley, Chichester, 2002.

[70] David M. Haaland and Edward V. Thomas. Partial least-squares methods for spectral analyses. 1. Relation to other quantitative calibration methods and the extraction of qualitative information. *Analytical Chemistry*, 60:1193–1202, 1988.

[71] Fredrik Lindgren, Paul Geladi, and Svante Wold. The kernel algorithm for PLS. *Journal of Chemometrics*, 7:45–59, 1993.

[72] Bhupinder S. Dayal and John F. MacGregor. Improved PLS algorithms. *Journal of Chemometrics*, 11:73–85, 1997.

[73] S. de Jong. SIMPLS—an alternative approach to partial least-squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18(3):251–263, 1993.

[74] Nicolaas M. Faber and Bruce R. Kowalski. Propagation of measurement errors for the validation of predictions obtained by principal component regression and partial least squares. *Journal of Chemometrics*, 11:181–238, 1997.

[75] H. Van der Voet. Pseudo-degrees of freedom for complex predictive models: The example of partial least squares. *Journal of Chemometrics*, 13(3-4):195–208, 1999.

[76] M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society Series B—Methodological*, 36(2):111–147, 1974.

[77] Qing-Song Xu and Yi-zeng Liang. Monte Carlo cross validation. *Chemometrics and Intelligent Laboratory Systems*, 56:1–11, 2001.

[78] S. Gourvénec, J. A. Fernández Pierna, D. L. Massart, and D. N. Rutledge. Evaluation of the PoLiSh smoothed regression and the Monte Carlo cross-validation for the determination of the complexity of a PLS model. *Chemometrics and Intelligent Laboratory Systems*, 68:41–51, 2003.

[79] Qing-Song Xu, Yi-Zeng Liang, and Yi-Ping Du. Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration. *Journal of Chemometrics*, 18:112–120, 2004.

[80] Ron Wehrens, Hein Putter, and Lutgarde M. C. Buydens. The bootstrap: a tutorial. *Chemometrics and Intelligent Laboratory Systems*, 54:35–52, 2000.

[81] Nicolaas M. Faber. A closer look at the bias-variance trade-off in multivariate calibration. *Journal of Chemometrics*, 13:185–192, 1999.

[82] Harald A. Martens and Pierre Dardenne. Validation and verification of regression in small data sets. *Chemometrics and Intelligent Laboratory Systems*, 44:99–121, 1998.

[83] Robert L. Green and John H. Kalivas. Graphical diagnostics for regression model determinations with consideration of the bias/variance trade-off. *Chemometrics and Intelligent Laboratory Systems*, 60:173–188, 2002.

[84] Michael C. Denham. Choosing the number of factors in partial least squares regression: estimating and minimizing the mean squared error of prediction. *Journal of Chemometrics*, 14: 351–361, 2000.

[85] Abraham Savitzky and Marcel J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964.

[86] Hannibal H. Madden. Comments on the Savitzky-Golay convolution method for least-squares smoothing and differentiation of digital data. *Analytical Chemistry*, 50(9):1383–1386, 1978.

[87] Peter A. Gorry. General least-squares smoothing and differentiation by the convolution (Savitzky-Golay) method. *Analytical Chemistry*, 62:570–573, 1990.

[88] M. S. Srivastava. *Methods of Multivariate Statistics*. Wiley, New York, 2002.

[89] Walter Lindgren, Jan-Åke Persson, and Svante Wold. Partial least squares method for spectrofluorimetric analysis of mixtures of humic acid and ligninsulfonate. *Analytical Chemistry*, 55:643–648, 1983.

[90] Alejandro C. Olivieri, Nicolaas M. Faber, Joan Ferré, Ricard Boqué, John H. Kalivas, and Howard Mark. Uncertainty estimation and figures of merit for multivariate calibration. *Pure and Applied Chemistry*, 78(3):633–661, 2006.

[91] R. DiFoggio. Examination of some misconceptions about near-infrared analysis. *Applied Spectroscopy*, 49(1):67–75, 1995.

[92] Harald Martens. Modified jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR). *Food Quality and Preference*, 11:5–16, 2000.

[93] J. A. Fernández Pierna, L. Jin, F. Wahl, and D. L. Massart. Estimation of partial least squares regression prediction uncertainty when the reference values carry a sizeable measurement error. *Chemometrics and Intelligent Laboratory Systems*, 65:281–291, 2003.

[94] A. Phatak, P. M. Reilly, and A. Penlidis. An approach to interval estimation in partial least squares regression. *Analytica Chimica Acta*, 277:495–501, 1993.

[95] Michael C. Denham. Prediction intervals in partial least squares. *Journal of Chemometrics*, 11: 39–52, 1997.

[96] Nicolaas M. Faber, David L. Duewer, Steven J. Choquette, Terry L. Green, and Stephen N. Chesler. Characterizing the uncertainty in near-infrared spectroscopy prediction of mixed-oxygenate concentrations in gasoline: Sample-specific prediction intervals. *Analytical Chemistry*, 70:2972–2982, 1998.

[97] R. Boqué, M. S. Larrechi, and F. X. Rius. Multivariate detection limits with fixed probabilities of error. *Chemometrics and Intelligent Laboratory Systems*, 45:397–408, 1999.

[98] Richard K. Burdick and Franklin A. Graybill. *Confidence Intervals on Variance Components*. Statistics: Textbooks and Monographs. Marcel Dekker, New York, 1992.

[99] André Hubaux and Gilbert Vos. Decision and detection limits for linear calibration curves. *Analytical Chemistry*, 42(8):849–855, 1970.

[100] K. Faber and B. R. Kowalski. Improved estimation of the limit of detection in multivariate calibration. *Fresenius Journal of Analytical Chemistry*, 357(7):789–795, 1997.

[101] Ricard Boqué, N. K. M. Faber, and F. X. Rius. Detection limits in classical multivariate calibration models. *Analytica Chimica Acta*, 423(1):41–49, 2000.

[102] R. Boqué and F. X. Rius. Multivariate detection limits estimators. *Chemometrics and Intelligent Laboratory Systems*, 32:11–23, 1996.

[103] Darren T. Andrews, Liguo Chen, Peter D. Wentzell, and David C. Hamilton. Comments on the relationship between principal components analysis and weighted linear regression for bivariate data sets. *Chemometrics and Intelligent Laboratory Systems*, 34:231–244, 1996.

[104] Peter R. Griffiths and James A. de Haseth. *Fourier Transform Infrared Spectroscopy*. Wiley, New York, 1986.

[105] Bruker Optics. Opus version 4.2, 2003.

[106] Michelle L. Hamilton. *Applications of grazing-angle reflection-absorption Fourier-transform infrared spectroscopy to the analysis of surface contamination*. PhD thesis, Department of Chemistry, University of Canterbury, Christchurch, New Zealand, 2006 (pending completion).

[107] J. F. James. *A Student's Guide to Fourier Transforms*. Cambridge University Press, Cambridge, second edition, 2002.

[108] Fredric J. Harris. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proceedings of the IEEE*, 66(1):51–83, 1978.

[109] John E. Bertie. Specification of components, methods and parameters in Fourier transform spectroscopy by Michelson and related interferometers. *Pure and Applied Chemistry*, 70(10): 2039–2045, 1998.

[110] Francis M. Mirabella. Principles, theory and practice of internal reflection spectroscopy. In John M. Chalmers and Peter R. Griffiths, editors, *Handbook of Vibrational Spectroscopy*, volume 2, pages 1091–1102. Wiley, Chichester, 2002.

[111] James Fitzpatrick and John A. Reffner. Macro and micro internal reflection accessories. In John M. Chalmers and Peter R. Griffiths, editors, *Handbook of Vibrational Spectroscopy*, volume 2, pages 1103–1123. Wiley, Chichester, 2002.

[112] Bernhard Lendl and Boris Mizaikoff. Optical fibers for mid-infrared spectrometry. In John M. Chalmers and Peter R. Griffiths, editors, *Handbook of Vibrational Spectroscopy*, volume 2, pages 1541–1550. Wiley, Chichester, 2002.

[113] James A. Harrington. *Infrared Fibers and Their Applications*. SPIE Press, 2004.

[114] Valentina F. Kokorina. *Glasses for Infrared Optics*. CRC Press, Boca Raton, 1996.

[115] Peter J. Melling and Paul H. Shelley. Spectroscopic accessory for examining films and coatings on solid surfaces, 2001. US Patent No. 6,310,348.

[116] George H. Rieke. *Detection of Light: from the Ultraviolet to the Submillimeter*. Cambridge University Press, Cambridge, 1994.

[117] Jean Timmermans. *Physico-chemical Constants of Pure Organic Compounds*. Elsevier, New York, 1950.

[118] Narayan C. Giri. *Introduction to Probability and Statistics. Part I: Probability*. Statistics: Textbooks and Monographs. Marcel Dekker, New York, 1974.

[119] Daniel Blaudez, Thierry Buffeteau, Bernard Desbat, Patrice Fournier, Anna-Marie Ritcey, and Michel Pézolet. Infrared reflection-absorption spectroscopy of thin organic films on nonmetallic substrates: optimal angle of incidence. *Journal of Physical Chemistry B*, 102:99–105, 1998.

[120] Jürgen Kattner and Helmuth Hoffmann. External reflection spectroscopy of thin films on dielectric substrates. In John M. Chalmers and Peter R. Griffiths, editors, *The Handbook of Vibrational Spectroscopy*, volume 2, pages 1009–1027. Wiley, Chichester, 2002.

[121] Gregory T. Merklin and Peter R. Griffiths. Brewster-angle reflection-absorption infrared spectrometry of organic films on metallic substrates. *Journal of Physical Chemistry B*, 101:7408–7413, 1997.

[122] David F. Edwards. Silicon. In Edward D. Palik, editor, *Handbook of Optical Constants of Solids*, pages 547–569. Academic Press, Orlando, 1985.

[123] Marvin J. Weber. *Handbook of Optical Materials*. CRC, Boca Raton, 2003.

[124] Eric W. Weisstein. Inverse trigonometric functions. In *Mathworld—A Wolfram Web Resource*. . http://mathworld.wolfram.com/InverseTrigonometricFunctions.html (Current Aug. 2006).

[125] Norbert Neuroth. Optical properties: Transmission and reflection. In Hans Bach and Norbert Neuroth, editors, *The Properties of Optical Glass*, pages 82–96. Springer-Verlag, Berlin, 1995.

[126] R. K. Iler. Multilayers of colloidal particles. *Journal of Colloid and Interface Science*, 21(6): 569–594, 1966.

[127] G. Decher and J. D. Hong. Buildup of ultrathin multilayer films by a self-assembly process. 1. Consecutive adsorption of anionic and cationic bipolar amphiphiles on charged surfaces. *Makromolekulare Chemie-Macromolecular Symposia*, 46:321–327, 1991.

[128] P. Bertrand, A. Jonas, A. Laschewsky, and R. Legras. Ultrathin polymer coatings by complexation of polyelectrolytes at interfaces: suitable materials, structure and properties. *Macromolecular Rapid Communications*, 21(7):319–348, 2000.

[129] Kristie Lenahan-Cooper. *Electrostatic Self-Assembly of Linear and Nonlinear Optical Thin Films*. PhD thesis, Virginia Polytechnic Institute and State University, 1999.

[130] Ralph K. Iler. *The Chemistry of Silica*. Wiley, New York, 1979.

[131] G. Decher, J. D. Hong, and J. Schmitt. Buildup of ultrathin multilayer films by a self-assembly process. 3. Consecutively alternating adsorption of anionic and cationic polyelectrolytes on charged surfaces. *Thin Solid Films*, 210:831–835, 1992.

[132] L. Kolarik, D. N. Furlong, H. Joy, C. Struijk, and R. Rowe. Building assemblies from high molecular weight polyelectrolytes. *Langmuir*, 15(23):8265–8275, 1999.

[133] Narinder K. Mehta, Javier Goenaga-Polo, Samuel P. Hernandez-Rivera, David Hernandez, Mary A. Thomson, and Peter J. Melling. Development of an in situ spectroscopic method for cleaning valdiation using mid-ir fibre-optics. *Biopharm*, 15(5):36,38–40,42,71, 2002.

[134] Noel Teelucksingh and K. Bal Reddy. Quantification of active pharmaceutical ingredients on metal surfaces using a mir-ir grazing-angle fiber optics probe - an in-situ cleaning verification process. *Spectroscopy*, 20(10):16–20, 2005.

[135] Michelle L. Hamilton, Benjamin B. Perston, Peter W. Harland, Bryce E. Williamson, Mary A. Thomson, and Peter J. Melling. Grazing-angle fiber-optic IRRAS for in situ cleaning validation. *Organic Process Research and Development*, 9:337–343, 2005.

[136] Jean Steinier, Yves Termonia, and Jules Deltour. Comments on smoothing and differentiation of data by simplified least square procedure. *Analytical Chemistry*, 44(11):1906–1909, 1972.

[137] Benjamin B. Perston, Michelle L. Hamilton, Bryce E. Williamson, Peter W. Harland, Mary A. Thomson, and Peter J. Melling. Grazing-angle fiber-optic FT-IRRAS for the *in situ* detection and quantification of multiple active pharmaceutical ingredients on glass. *Analytical Chemistry*. Submitted.

[138] Kiyoshi Yamamoto and Hatsuo Ishida. Optical theory applied to infrared spectroscopy. *Vibrational Spectroscopy*, 8:1–36, 1994.

[139] J. A. Mielczarski and E. Mielczarski. Infrared external reflection spectroscopy of adsorbed monolayers in a region of strong absorption of substrate. *Journal of Physical Chemistry B*, 103: 5852–5859, 1999.

[140] Mary Beth Seasholtz and Bruce R. Kowalski. The effect of mean centering on prediction in multivariate calibration. *Journal of Chemometrics*, 6:103–111, 1992.

[141] Husheng Yang, Peter R. Griffiths, and J. D. Tate. Comparison of partial least squares regression and multi-layer neural networks for quantification of nonlinear systems and application to gas phase fourier transform infrared spectra. *Analytica Chimica Acta*, 489:125–136, 2003.

[142] Michelle L. Hamilton, Benjamin B. Perston, Peter W. Harland, Bryce E. Williamson, Mary A. Thomson, and Peter J. Melling. Fiber-optic IRRAS for trace analysis on surfaces of varying roughness: Sodium dodecyl sulfate on stainless steel. *Applied Spectroscopy*, 2006.

[143] Merck ChemDAT: the Merck Chemical Database (CD-ROM), 2004.

[144] R. G. Snyder, H. L. Strauss, and C. A. Elliger. C–H stretchlng modes and the structure of *n*-alkyl chains. 1. Long, disordered chains. *Journal of Physical Chemistry*, 86:5145–5150, 1982.

[145] R. A. MacPhail, H. L. Strauss, R. G. Snyder, and C. A. Elliger. C–H stretchlng modes and the structure of *n*-alkyl chains. 2. Long, all-trans chains. *Journal of Physical Chemistry*, 88: 334–341, 1984.

[146] Eigenvector Research. PLS Toolbox 3.5, 2006. Website: http://www.eigenvector.com.

[147] Marc D. Porter, Thomas B. Bright, and David L. Allara. Quantitative aspects of infrared external reflection spectroscopy: Polymer/glassy carbon interface. *Analytical Chemistry*, 58:2461–2465, 1986.

[148] Joshua Lehr. *Pharmaceutical Equipment Cleanliness Validation by Grazing Angle Fibre Optic FT-IRRAS*. BSc (Hons) report (unpublished), Department of Chemistry, University of Canterbury, Christchurch, New Zealand, 2005.

[149] Robert N. Feudale, Nathaniel A. Woody, Huwei Tan, Anthony J. Myles, Steven D. Brown, and Joan Ferré. Transfer of multivariate calibration models: a review. *Chemometrics and Intelligent Laboratory Systems*, 64:181–192, 2002.

[150] Indira S. Adhibetty, Joseph A. McGuire, Boonari Wangmaneerat, Thomas M. Niemczyk, and David M. Haaland. Achieving transferable multivariate spectral calibration models: Demonstration with infrared spectra of thin-film dielectrics on silicon. *Analytical Chemistry*, 63:2329–2338, 1991.

[151] X. Capron, B. Walczak, O. E. de Noord, and D. L. Massart. Selection and weighting of samples in multivariate regression model updating. *Chemometrics and Intelligent Laboratory Systems*, 76:205–214, 2005.

[152] Bhupinder S. Dayal and John F. MacGregor. Recursive exponentially weighted PLS and its applications to adaptive control and prediction. *Journal of Process Control*, 7(3):169–179, 1997.

[153] Chris L. Stork and Bruce R. Kowalski. Weighting schemes for updating regression models—a theoretical approach. *Chemometrics and Intelligent Laboratory Systems*, 48:151–166, 1999.

[154] George W. Snedecor and William G. Cochran. *Statistical Methods*. Iowa State University Press, Ames, Iowa, 8th edition, 1989.

[155] John Mandel. *The Statistical Analysis of Experimental Data*. Dover, New York, 1984.

[156] John Mandel and Frederic J. Linnig. Study of accuracy in chemical analysis using linear calibration curves. *Analytical Chemistry*, 29(5):743–749, 1957.

[157] Eric W. Weisstein. Ellipse. In *Mathworld—A Wolfram Web Resource*. . http://mathworld.wolfram.com/Ellipse.html (Current Aug. 2006).

[158] William H. Press, Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. *Numerical Recipes in C*. Cambridge University Press, Cambridge, second edition, 1992.

[159] Thermo Galactic. GRAMS/AI version 7.01, 2002.

[160] Sujit Banerjee and Dongyun Li. Interpreting multicomponent spectra by derivative minimization. *Applied Spectroscopy*, 45(6):1047–1049, 1991.

[161] G. R. Phillips and J. M. Harris. Polynomial filters for data sets with outlying or missing observations: Application to charge-coupled-device-detected raman spectra contaminated by cosmic rays. *Analytical Chemistry*, 62:2351–2357, 1990.

[162] Yukiteru Katsumoto and Yukihiro Ozaki. Practical algorithm for reducing convex spike noises on a spectrum. *Applied Spectroscopy*, 57(3):317–322, 2003.

[163] Robert S. McDonald and Paul A. Wilks. JCAMP-DX: A standard form for exchange of infrared spectra in computer readable form. *Applied Spectroscopy*, 42:151–162, 1988.

[164] Michael A. Friese and Sujit Banerjee. Lignin determination by FT-IR. *Applied Spectroscopy*, 46 (2):246–248, 1992.