NTNU – Trondheim
Norwegian University of
Science and Technology

Comparison of Principal Component Analysis and Spectral Angle Mapping for Identification of Materials in Terahertz Transmission Measurements

Helle Emilia Nystad

Master of Science in Electronics
Submission date: Januar 2015
Supervisor: Johannes Skaar, IET
Co-supervisor: Magnus W. Haakestad, Forsvarets forskningsinstitutt

Norwegian University of Science and Technology
Department of Electronics and Telecommunications

Problem description

The terahertz range of the electromagnetic spectrum consists of frequencies between 0.1 THz and 10 THz, occupying a gap between radar and infrared frequencies. Many materials, such as plastic, paper and clothes, are almost completely transparent in the THz range, while other materials, such as explosives and drugs, have characteristic fingerprints in this frequency range.

This assignment involves

1. Measuring different materials in a transmission setup using Terahertz time-domain spectroscopy.

2. Comparison of two algorithms, called "Principal Component Analysis" (PCA) and "Spectral Angle Mapping" (SAM) for identification of the materials using the measured THz spectra.

Abstract

The terahertz range of the electromagnetic spectrum ranges from 0.1 to 10 THz, and has some unique properties which make it interesting for security applications. The identification of a range of dangerous substances is possible using THz radiation, because many of these materials feature characteristic absorption lines in this regime. Another property is the ability to penetrate common sealing materials, such as paper, plastic and cloth, enabling the possibility for identification of concealed substances.

This thesis compares two methods, namely principal component analysis (PCA) and spectral angle mapping (SAM), for identification of different materials acting as simulants for dangerous substances. PCA is a method which transforms a number of correlated variables into a smaller number of uncorrelated variables, called principal components. The original data is projected on to these, forming a new coordinate system where the original data is expressed in an optimal way, using much fewer dimensions. SAM is a spectral recognition technique, which calculates the dot product between an unknown spectrum, and a reference spectrum, both treated as vectors.

Measurements on samples containing Tartaric acid, Lactose and RDX (an explosive) were carried out using Terahertz time-domain spectroscopy, and the spectral fingerprints were obtained, and used for training each algorithm. Two spectral characteristics were considered: The absorption spectrum itself, and its derivative, both investigated for two different window widths. Four terahertz images for testing the algorithms were acquired, one using no barrier, and three using either paper, plastic or a piece of cloth for covering the samples. Also tested was the ability to recognize a material when its sample properties differ from those used for training the algorithms, by looking at four different Tartaric acid samples. The algorithms were implemented using MATLAB, and compared using ROC curves.

The performance of PCA showed that careful consideration must be taken when choosing the number of principal components, and that the optimal number differs depending on spectral characteristic.

In general, very good results were obtained when appropriate windowing was applied, and the best overall performance resulted from applying the narrower window, both for PCA and SAM.

A true positive rate above 0.9 with a false positive rate of less than 0.2 could be obtained, regardless of barrier, also in the case of Tartaric acid. For PCA, these results were obtained using the absorption spectrum, while for SAM, this was the case regardless of spectral characteristic.

The paper and plastic barriers were not challenging for either algorithm, and using these yielded essentially the same results as using no barrier in most cases. There were some differences in the performance of PCA and SAM, but these were small. The most challenging barrier was the cloth, for which classification using SAM with the absorption spectrum was slightly better than PCA, but the advantage was small.

Sammendrag

Terahertz-frekvensområdet av det elektromagnetiske spekteret spenner fra 0.1 til 10 THz, og har noen unike egenskaper som gjør det attraktivt bl.a.innen sikkerhet. Identifikasjon av en rekke farlige stoffer er mulig ved å bruke THz-stråling, da flere av disse har karakteristiske absorpsjonslinjer, kalt spektrale fingeravtrykk, i THz-frekvensområdet. En annen egenskap er at flere av materialene som ofte brukes for å skjule disse, er transparente i dette frekvensområdet, hvilket gjør det mulig å identifisere stoffer også når de er forsøkt skjulte.

Dette arbeidet sammenligner to metoder, kalt principal component analysis (PCA) og spectral angle mapping (SAM), for identifisering av ulike materialer som blir brukt som simulanter for farlige stoffer. PCA er en metode som transformerer et antall korrelerte variabler til et mindre antall ukorrelerte variabler, kalt "principal components". Ved hjelp av projeksjon av de originale dataene på et visst antall principal components, uttrykkes dataene på en optimal måte, ved hjelp av færre dimensjoner. SAM er en metode der et spekter representeres som en vektor, og der skalarproduktet mellom et referansespekter og et ukjent spekter regnes ut.

Målinger på prøver som inneholdt vinsyre, laktose og sprengstoffet RDX ble utført ved hjelp av Terahertz tidsdomenespektroskopi, for å finne de spektrale fingeravtrykkene, som ble brukt for å lære opp algoritmene. Absorpsjonsspekteret i seg selv, samt dets deriverte, ble vurderte, for to ulike vindusbredder. Fire terahertz-bilder ble brukt til å teste algoritmene: en der prøvene ikke var tildekte, og tre der prøvene var dekket av enten papir, plast eller et tøystykke. Algoritmenes evne til å gjenkjenne et materiale når dets materialegenskaper varierte i forhold til refreranseprøvene ble også testet, ved å se på fire ulike vinsyreprøver. Algoritmene ble implementerte i MATLAB, og ytelsen ble sammenlignet ved hjelp av ROC-kurver.

Resultatene for PCA viste at valget av antallet "principal components" man uttrykker dataene ved hjelp av, må vurderes nøye, da det optimale antallet er avhengig av hvorvidt man bruker absorpsjonsspekteret eller dets deriverte.

Generelt sett var ytelsen for begge algoritmene veldig god når en hensiktsmessig vindusbredde ble brukt, og de beste resultatene ble oppnådd ved bruk av det smalere vinduet, for både PCA og SAM.

Det var mulig å oppnå en "true positive rate" på over 0.9 med en "false positive rate" på mindre enn 0.2 for alle barrierer, også ved klassifisering av vinsyre. For PCA var dette i noen tilfeller avhengig av at man brukte absorpsjonsspekteret, mens for SAM var det mulig både for absorpsjonsspekteret og dets deriverte.

Papir- og plastbarrierene var ikke utfordrende for noen av algoritmene, og resul-

tatene var essiensielt de samme som når ingen barriere ble brukt for de aller fleste tilfellene. Det var noen forskjeller i ytelsen for PCA og SAM, men disse var små. Den mest utfordrende barrieren var tøystykket, og det viste seg at klassifisering ved bruk av SAM med absorpsjonsspekteret gav litt bedre resultater enn bruk av PCA, men fordelen var liten.

Preface

This thesis concludes my Master of Science degree in "Electronics system design and innovation" at the Norwegian University of Science and Technology (NTNU) in Trondheim, and is written in collaboration with the Norwegian Defence Research Establishment (FFI). The experimental work was conducted at FFIs facilities at Kjeller during the first two weeks of September 2014, before I returned to NTNU to work on the analysis. I have truly enjoyed the process, and learned a lot from overcoming the challenges I came across. There are several people without whom this would not have been possible.

First of all, I would like to thank my supervisors, Magnus W. Haakestad and Johannes Skaar, for making it possible to realize this work. A special thanks to Magnus, for the time and patience he has invested in helping and guiding me whenever I got stuck and could not move forward. Throughout my stay at FFI, and through countless emails, he has answered any question thrown his way.

I would also like to thank my fellow students, for trying to help me solve problems often irrelevant to their own studies. Finally, I would like to thank my parents, for supporting me in every way possible throughout my studies, and my sister, for always cheering me on.

Trondheim,
13.01.2014

Helle Emilia Nystad

Table of Contents

List of Tables

List of Figures

Abbreviations

DFT	Discrete Fourier transform
FT	Fourier transform
FFT	Fast Fourier transform
FTS	Fourier transform spectroscopy
IDFT	Inverse discrete Fourier tranform
IR	Infrared
ITU	International Telecommunication Union
NIR	Near-infrared
OPL	Optical path length
PC	Principal component
PCA	Principal component analysis
SAM	Spectral angle mapping
THz	Terahertz
THz-TDS	Terahertz time-domain spectroscopy

1 Introduction

1.1 Terahertz radiation

Terahertz radiation commonly refers to electromagnetic waves propagating with frequencies ranging from 0.1 to 10 THz (3 mm $> \lambda >$ 30 μm). As of today, no standard definition of this spectral band exists [1], but this definition is commonly used in material research, and will define the THz band throughout this thesis.

Being situated between microwaves and infrared radiation (see Figure 1.1), forming a bridge between conventional electronics and photonics, THz radiation shares some properties with each of its neighbouring regions.

THz radiation is non-ionizing, eliminating the damage associated with ionizing

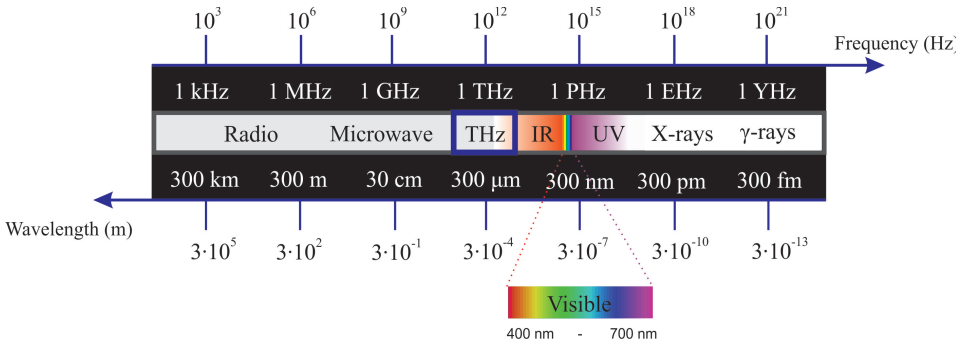


Figure 1.1: The electromagnetic spectrum. The THz band lies between the high-frequency edge of microwaves, and low-frequency edge of infrared.

radiation such as X-rays, yet penetrates a wide variety of non-conduction, nonpolar, organic materials which are opaque in the visible and NIR-regions. In addition, many substances have unique spectral characteristics, known as "spectral finger-

[1] The ITU-designated band ranges from 0.3 to 30 THz (10 μm - 1 mm) [2]. However, this definition intrudes the well established mid-IR band, and is mainly of interest when dealing with ultrabroadband THz pulses [3].

prints", in the THz range. This has been exploited in laboratory demonstrations for material recognition and characterization. Water, being a polar molecule, is one of the strongest absorbers of THz radiation. For many potential applications, this is considered one of the main challenges of THz technology, whether it is vapour in the air or water in biological tissue. Other applications can exploit this property. Distinguishing between materials with varying water content is interesting in e.g. prosess and quality control, as well as biomedical imaging [4].

1.2 Terahertz technology

Technical difficulties in transmitting and detecting THz radiation has limited their use in the past. This lack of technology is known as the "terahertz gap", and arises from the nature of the sources and detectors used in the well-established neighbouring fields, with electronic sources on the low-frequency side, and optical sources on the high-frequency side of the gap. One terahertz corresponds to a photon energy of 4 meV, and common semiconductor materials have a bandgap in the order of one electron volt. On the other side, electronic state transitions of atoms and molecules commonly used in lasers are much higher than THz energies. The need of scaling electronic sources up, or optical sources down to the THz region, also holds for detectors. Not until recently have sufficiently powerful transmitters and sensitive receivers been developed [5], largely due to development of new semiconductor materials, e.g. quantum cascade lasers and high electron mobility transistors [6].

The invention of Terahertz time-domain spectroscopy (THz-TDS) during the 1990s [7] paved the way for new technology in hopes of "closing the gap" and fully exploiting the unique characteristics of THz radiation. Potential applications of THz technology include medical imaging, communication technology, material research and surveillance. Although sources and detectors are available today, THz technology is far from reaching its full potential. It has become a hot topic in research the last decade, and it is expected that THz technology will influence all of these fields in the years to come.

1.3 Motivation and purpose

The combination of non-destructive penetration of common sealing materials, including paper, plastic and textiles, and unique spectral characteristics of many illegal substances, such as explosives and drugs, makes THz radiation highly suitable for security applications. The possibility for detection and recognition of harmful, concealed materials using THz technology is the motivation behind this master's thesis. Possible applications include screening of baggage and mail. Also, because

THz radiation is considered safe for human tissue, combining spectral imaging and identification could be used in scanning humans for weapons or illegal substances by targeting detection to a very specific range of materials, thus avoiding privacy concerns.

The use of robust algorithms is essential in practical applications, and the purpose of this master's thesis is comparing two methods, known as spectral angle mapping (SAM) and principal component analysis (PCA), for material recognition using data obtained by measuring different substances in a transmission THz-TDS setup. SAM is a well-established spectral recognition technique, and the use of PCA in automated, chemical identification using THz spectrometry has been attempted [8] [9], but to the knowledge of the author, no extensive comparison of SAM and PCA has been carried out for THz radiation before.

2 Terahertz time-domain spectroscopy

2.1 Spectral fingerprints

Spectroscopy is the study of interaction of electromagnetic radiation with matter, and relies on the study of absorption lines, which manifest themselves as a decrease in the amplitude of an otherwise continuous spectrum. The spectrum is typically obtained by Fourier transformation of the signal - defined in Section 2.3.1 later in this chapter. Absorption lines are a result of quantum mechanical temporal transitions in atoms or molecules induced by electromagnetic radiation. The frequency of the absorbed radiation is related to the transition energy, according to

$$\Delta E = hf \tag{2.1}$$

where ΔE is the energy difference between the two energy levels involved in the transition, h is Planck's constant and f is the frequency. Atoms and molecules can absorb photons with energies corresponding to the difference between two energy states, and the absorption spectrum is hence characteristic for the material, which is why it is also known as a "spectral fingerprint". A broad range of materials have spectral fingerprints in the THz range, resulting from rotational and vibrational transitions of the molecules. THz spectroscopy systems can be used as a tool for identifying such materials in an otherwise inaccessible part of the electromagnetic spectrum.

2.2 Raw data

Many methods exist for performing spectroscopy on THz pulses, including Fourier-transform spectroscopy (FTS) and THz time-domain spectroscopy (THz-TDS). Contrary to conventional spectroscopy such as FTS, which only measures the spectrum, THz-TDS measures the electric field as a function of time, providing information about both amplitude and phase. The technical details regarding the setup are found in Section 6.1.1 in Chapter 6.

Figure 2.1 shows a plot of a transmission measurement in air and and transmission through a sample containing Tartaric acid. The optical path length (OPL) of the beam is longer when passing through the sample, causing an extra delay, and the pulse is attenuated as a result of absorption and Fresnel reflection.

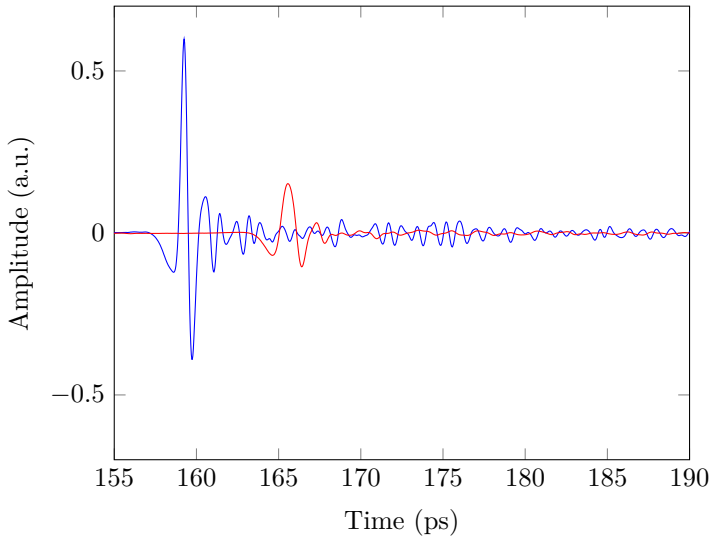


Figure 2.1: THz pulse in air (blue) and a sample containing Tartaric acid (red).

2.3 Signal processing

2.3.1 Fast Fourier transform

The frequency components of a time domain signal can be found by applying the Fourier transform, yielding the spectrum of the signal. It can be applied to both continuous and discrete signals, the latter referred to as the discrete Fourier transform (DFT) for signals of finite length. The fast Fourier transform (FFT), which is applied to the data in this work, is simply an efficient implementation of DFT. Suppose there are N consecutive sampled values in time-domain, $x_n = x(t_n)$, $n = 0, 1, 2, ...N - 1$. The definition of the discrete Fourier transform is [1]

$$X_k \equiv \sum_{n=0}^{N-1} x_n e^{-i2\pi kn/N} \tag{2.2}$$

[1] The sign in the exponent depends on convention. However, regardless of convention, DFT and IDFT always have opposite-sign exponents.

Transforming back to time domain is accomplished using the inverse discrete Fourier transform (IDFT)

$$x_n = \frac{1}{N} \sum_{k=0}^{N-1} X_k e^{i2\pi kn/N} \tag{2.3}$$

An important property of Fourier transforms is that the spectral width is inversely proportional to the temporal width (for so-called transform limited pulses)[10, p. 1124], meaning that a narrow feature in time domain becomes broad in the frequency domain (and vice versa). Another property is that multiplication in the time domain corresponds to convolution in the frequency domain (spectral smoothing). For more information regarding Fourier transforms and their properties, see e.g. [10, p.1122-1128]. More information regarding DFT is provided in e.g. [11].

2.3.2 Windowing

A window function is a mathematical function that is zero-valued outside some chosen interval. The *half width* of a window function, denoted δt, is the time interval from the peak of the window to the time where the window function reaches zero. Many different windows exist, each with advantages and shortcomings, and the choice depends on the application. The details regarding window functions are not provided here, but the effects of applying a window function on the signals in time domain are illustrated by some examples. More information regarding windowing in general can be found in e.g. [12] and [13].

In this thesis, a Blackman-Harris window is applied by multiplying the signals in time domain with the window function. The window function w is given by [13, p.151]

$$w(k) = a_0 + a_1 \cos(\frac{2\pi k}{N-1}) + a_2 \cos(\frac{4\pi k}{N-1}) - a_3 \cos(\frac{6\pi k}{N-1}) \tag{2.4}$$

where $k = 0,1,...,N-1$, and N is the width (in samples) of the discrete-time window function. Several variations of the window exist, depending on the coefficients a. The window used in this work is shown in Figure 2.2, for half widths of $\delta t = 30$ ps and $\delta t = 15$ ps, centered at 50 ps.

A THz pulse in air is plotted together with the Blackman-Harris window in Figure 2.3a. The window needs to be centered on the peak of the pulse. Figure 2.3b shows the same pulse after applying a Blackman-Harris window with half width 30 ps, resulting in zero values of the amplitude outside of the window. The windowing is done for two purposes: Reducing noise, and reducing the effect of absorption caused by water vapour in the air.

Figure 2.4 shows the amplitude spectrum of air, before and after windowing, in the frequency interval 0.3 to 1.5 THz. Note the dips in the spectrum, which arise

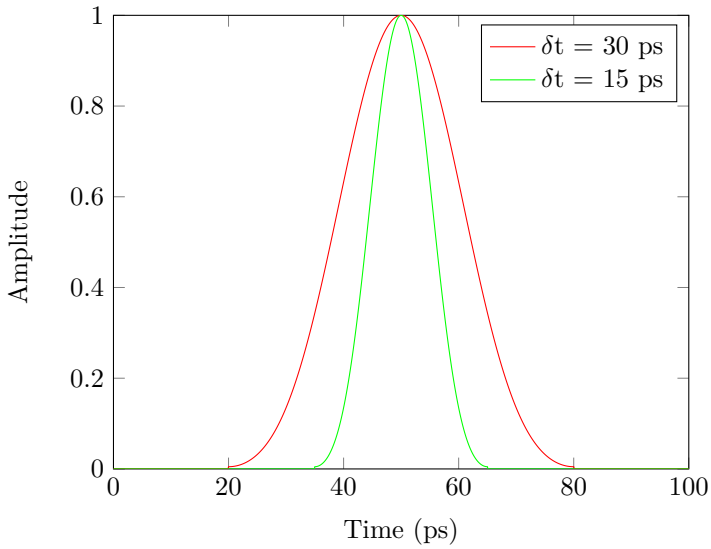
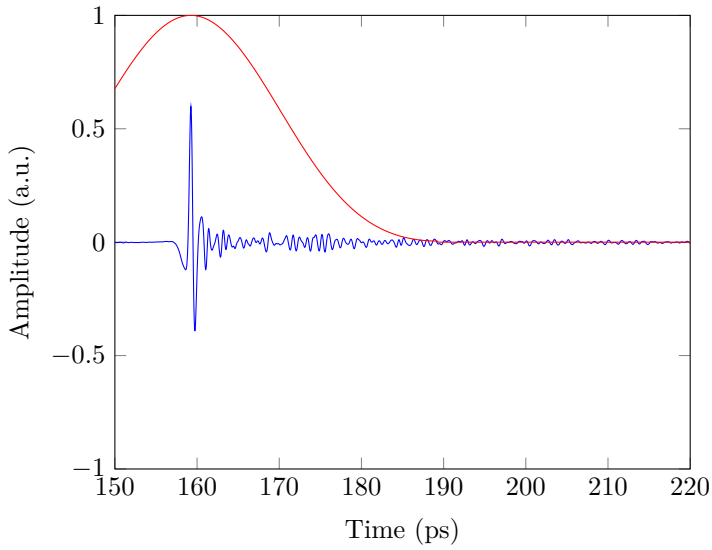
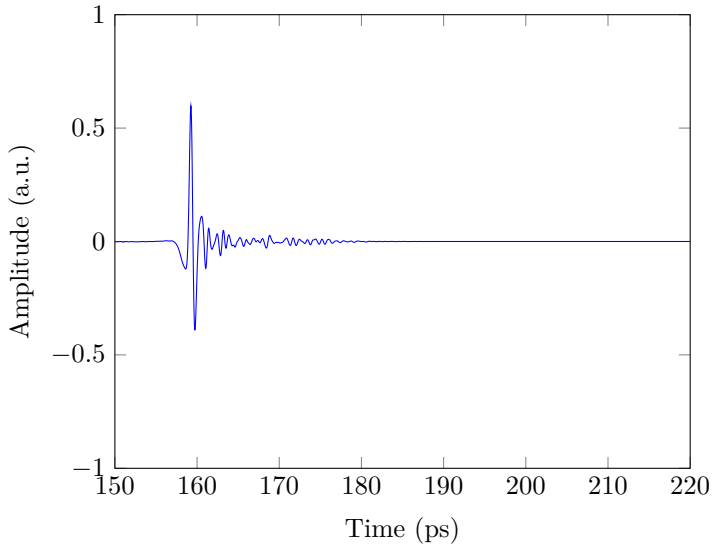


Figure 2.2: A Blackman-Harris window plotted for two different half widths, $\delta t = 30$ ps and $\delta t = 15$ ps.

from absorption by water vapour in the air, before windowing. By applying a window function on the signal in time domain, the absorption lines become less defined. With a narrower window, more smoothing is obtained, but the width of the dips also increases.



(a)



(b)

Figure 2.3: THz pulse in air (blue) (a) plotted together with a Blackman-Harris window (red) with $\delta t = 30$, centered at the signal peak (b) after windowing.

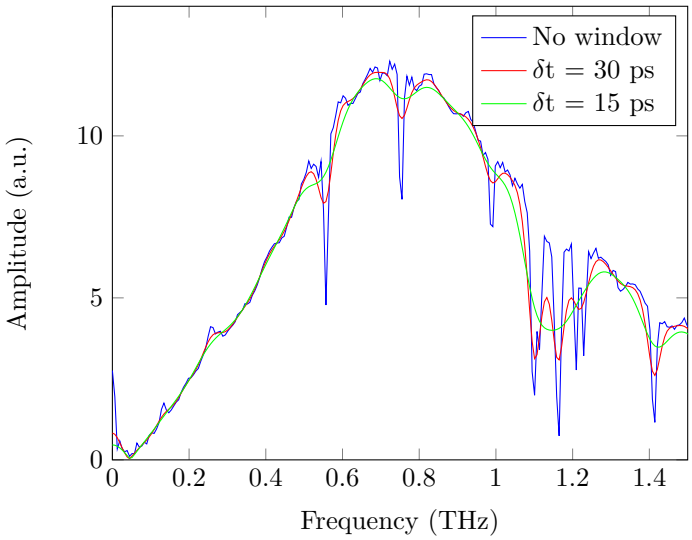


Figure 2.4: Amplitude spectrum of transmission through air, before and after windowing with a Blackman-Harris window, for two different half widths δt.

2.3.3 Transmission and absorption spectra

The detected signal is influenced not only by the sample properties, but also the surroundings (instrumental setup, ambient medium etc.). In addition to carrying out measurements on the samples, it is therefore customary to perform a reference measurement. The appropriate reference depends on the application. In many cases, a measurement in ambient air is used. Division of the sample spectrum, s with a reference spectrum, s_0, both of which are obtained by DFT of the measured signal, will reduce these unwanted effects, yielding the transmission spectrum $T(f)$

$$T(f) = \frac{s(f)}{s_0(f)} = \mathrm{e}^{-\alpha(f)l} \tag{2.5}$$

where $\alpha(f)$ is the attenuation coefficient, which is a product of the absorptivity and concentration, and l is the distance the light travels through the material (path length). Fresnel reflection losses are neglected for simplicity. Absorbance and transmittance are closely related, and the absorption spectrum can be obtained from the transmission spectrum using

$$A(f) = -\alpha(f)l = -\ln[T(f)] \tag{2.6}$$

We see that absorbance and transmittance are inversely related, however, this relationship is not linear, but logarithmic. The absorbance, in turn, is proportional with the attenuation coefficient and the path length.

2.3.4 Signal-to-noise ratio

The signal-to-noise ratio (SNR) is a measure of signal strength relative to noise, and is defined as

$$SNR = \frac{P_S}{P_N}[W] = P_s - P_N[dB] \tag{2.7}$$

where P_S is the power of the signal (meaningful information), and P_N is the power of the background noise (unwanted signal).

3 Principal component analysis

PCA is a widely used statistical method applied in a variety of fields, e.g. natural, medical, behavioural and social sciences. It has been called "one of the most important results from linear algebra" [14], and is a simple, powerful technique for reducing the dimensionality of complex data sets while retaining most of the information [15]. The goal of this chapter is providing an intuitive feel for PCA, as well as familiarizing the reader with the mathematics behind PCA. Before explaining PCA, background mathematics needed for understanding and performing PCA are introduced.

3.1 Background mathematics

The field of statistics revolves around the collection, analysis and presentation of data sets. Data sets are commonly referred to as either *populations* or *samples*. Populations represent entire collections, samples are groups selected from the population. This chapter will revolve around sample data sets, referred to simply as data sets. There are numerous ways of describing a set of data mathematically, including the mean, standard deviation, variance and covariance, which will be discussed below. An introduction to the covariance matrix and its eigenvectors and eigenvalues is also given.

Throughout this thesis, matrices are denoted in upper case bold, vectors in lower case bold and elements in lower case italics, using subscripts.

3.1.1 Mean, variance and covariance

The mean is simply the average of the numbers in the data set along one dimension, and is a measure for central tendency. The equation is:

$$\bar{x} = \frac{1}{n}\sum_{i}^{n} x_i \tag{3.1}$$

where $\bar{x}$ indicates the mean of a data set $\mathbf{x}$, and n is the number of elements in $\mathbf{x}$.

The standard deviation is the average distance from the mean to a data point. In other words, it measures the amount of variation from the average, describing how spread out the data is. It is found using [15]

$$s = \sqrt{\frac{1}{n-1}\sum_{i}^{n}(x_i - \bar{x})^2} \tag{3.2}$$

The standard deviation is always non-negative. A value equal to zero means all the numbers of the set are the same. If the standard deviation is small, the numbers are close to each other and the mean. A high standard deviation indicates the opposite: the numbers are spread out around the mean and each other.

Variance is another way of measuring the spread, closely related to standard deviation. In fact, it is just the standard deviation squared, yielding the formula [15]

$$s^2 = \mathrm{Var}(\mathbf{x}) = \frac{1}{n-1}\sum_{i}^{n}(x_i - \bar{x})^2 \tag{3.3}$$

Mean, standard deviation and variance are ways of describing each dimension of a data set independently. When investigating data with more than one dimension, one is often interested in looking at how two variables change together - their *correlation*. Covariance is a way of doing this. The covariance of the data sets $\mathbf{x}$ and $\mathbf{y}$ is [15]

$$\mathrm{Cov}(\mathbf{x}, \mathbf{y}) = \frac{1}{n-1}\sum_{i}^{n}(x_i - \bar{x})(y_i - \bar{y}) \tag{3.4}$$

The variables inside the brackets indicate which dimensions one is considering. Some properties of covariance include:

$\mathrm{Cov}(\mathbf{x}, \mathbf{y}) = \mathrm{Cov}(\mathbf{y}, \mathbf{x})$, the variables are interchangeable
$\mathrm{Cov}(\mathbf{x}, \mathbf{x}) = \mathrm{Var}(\mathbf{x})$, the covariance between one dimension and itself, is the variance.

The magnitude is not easy to interpret, but the sign tells us whether the variables tend to change together, in which case the covariance is positive, or whether they tend to show opposite behaviour, resulting in a negative covariance. If the covariance is zero, the dimensions are independent of each other, i.e. the variables are uncorrelated.

In a data set with more than two dimensions, the covariance between each pair of dimensions can be calculated. This is most easily accomplished by calculating the *covariance matrix*, which will be introduced in the next section.

3.1.2 Matrix algebra

When dealing with multi dimensional data, it is often useful to organize them in matrices. Let us assume we have a data set $\mathbf{X}$ represented by a m-by-n matrix, where n is the number of samples (observations) and m the number of variables:

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,n} \\ x_{2,1} & x_{2,2} & \dots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} & x_{m,2} & \dots & x_{m,n} \end{pmatrix} = \begin{pmatrix} \mathbf{x_1} \\ \mathbf{x_2} \\ \vdots \\ \mathbf{x_m} \end{pmatrix} \tag{3.5}$$

The *covariance matrix* is a square matrix containing all possible covariance values of a data set, and is obtained using [14]

$$\mathrm{Cov}(\mathbf{X}) = C_X = \frac{1}{n-1}\mathbf{X}\mathbf{X}^{\mathbf{T}} = \frac{1}{n-1}\begin{pmatrix} \mathbf{x_1}\mathbf{x_1}^T & \mathbf{x_1}\mathbf{x_2}^T & \dots & \mathbf{x_1}\mathbf{x_m}^T \\ \mathbf{x_2}\mathbf{x_1}^T & \mathbf{x_2}\mathbf{x_2}^T & \dots & \mathbf{x_2}\mathbf{x_m}^T \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{x_m}\mathbf{x_1}^T & \mathbf{x_m}\mathbf{x_2}^T & \dots & \mathbf{x_m}\mathbf{x_m}^T \end{pmatrix} \tag{3.6}$$

Some important properties of C_X are

C_X is an m-by-m square symmetric matrix
The diagonal terms of C_X are the variances of each variable
The off-diagonal terms of C_X are the covariances between different variables

Eigenvalues and *eigenvectors* are properties of square matrices[1], found by solving the characteristic equation of a square matrix, $\mathbf{A}$. An eigenvector of $\mathbf{A}$ is a non-zero vector $\mathbf{v}$ that, when multiplied with the matrix $\mathbf{A}$, yields the same as when some scalar λ multiplies $\mathbf{v}$:

$$\mathbf{A}\mathbf{v} = \lambda\mathbf{v} \tag{3.7}$$

This is equivalent to

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = 0 \tag{3.8}$$

and has a non-zero solution if and only if

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0, \tag{3.9}$$

which is known as the characteristic equation of $\mathbf{A}$. The left side is a polynomial function of λ, called the characteristic polynomial. Solving this often requires the use of approximate numerical methods. In this thesis, a built-in function in MATLAB is used (see Appendix A).

Eigenvalues and eigenvectors always come in pairs. $\mathbf{v}$ is the eigenvector and λ

[1]Not all square matrices have eigenvalues. This depends on whether the characteristic polynomial has at least one root.

is the eigenvalue corresponding to that eigenvector. What Eq. 3.7 tells us, is that the vectors $\mathbf{v}$ and $\mathbf{Av}$ are parallel. In other words, if the matrix $\mathbf{A}$ does not change the direction of a vector $\mathbf{v}$, only flips it or alters its length, we call the vector $\mathbf{v}$ an eigenvector and the scalar λ an eigenvalue. Another property of eigenvectors is that if $\mathbf{A}$ is symmetrical, the eigenvectors are orthogonal. In addition, if $\mathbf{A}$ is diagonal, the eigenvalues can be used to evaluate the variance of the data set. The eigenvector corresponding to the highest eigenvalue is in the direction of maximum variance, and so on. Both of these are worth noting for later regarding PCA.

3.2 Principal component analysis

PCA is commonly used as the first step for analysing complex, high dimensional data sets [14]. When measuring some phenomenon, one often records more dimensions than needed, because it is hard to know beforehand which variables best describe the dynamics of the system. In general, the variables of raw data are inter-correlated, causing redundancy, and the data is noisy. Spotting patterns in such unwieldy data is hard, sometimes impossible.

Let us look at an example illustrating this. Our original set of data contains a set of absorption spectra, see Figure 3.1. There are 30 different spectra (observations), from three different materials, and for each of the observations, 73 frequencies (variables) are recorded between 0.3 and 1.5 THz. Hence, our data matrix $\mathbf{X}$ is 73-by-30:

$$\mathbf{X} = \begin{pmatrix} \mathbf{x_1} \\ \vdots \\ \mathbf{x_{73}} \end{pmatrix} \tag{3.10}$$

where each row ($\mathbf{x}_i$) contains all measurements for one variable, and each column corresponds to a spectrum. We want to know whether there are relationships between the spectra, but this is hard to tell from our original representation of the data. The question is: Is there a way of re-expressing our set of data in an optimal way? As we shall see, there is, accomplished by using a new *basis*. This is where PCA comes in.

In practice, computing PCA on a data set involves [15] [16]

1) Pre-prosessing the data matrix
2) Calculating the covariance matrix
3) Finding the eigenvalues and eigenvectors of the covariance matrix
4) Representing the data using a new basis, the *principal components*.

We will explain what the principal components are, but first, let us look at the restrictions under which they are computed [16]:

1) The first principal component accounts for the largest amount of variance pos-

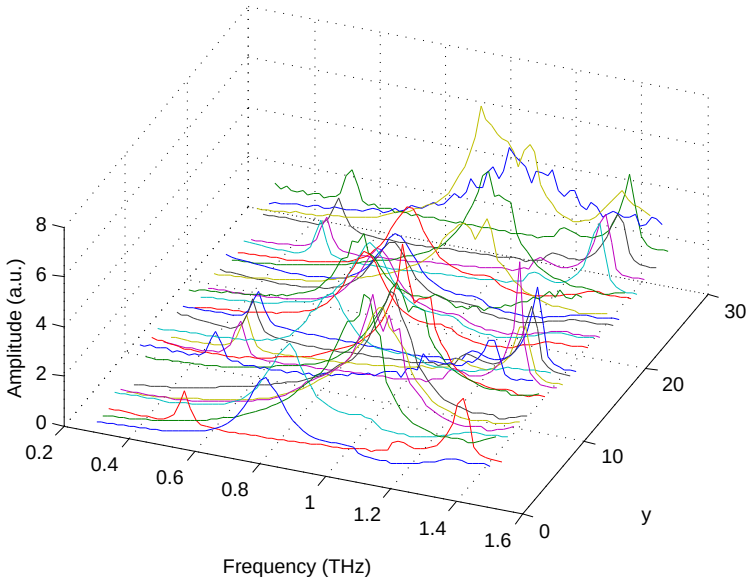


Figure 3.1: Example data for PCA, showing 30 individual absorption spectra, plotted with an offset along the y-axis.

sible
2) The succeeding components are orthogonal to the preceding one, while accounting for as much of the remaining variance as possible
3) The importance of each principal component is based on how large the variance of that component is, i.e. the first principal component is most "principal".

We will explain the steps involved in PCA, and we start out by looking for a linear, algebraic solution to the problem. Then, based on our example data set, we will look at what the data looks like after PCA has been applied.

We are seeking a new basis for expressing our data, i.e. we wish to linearly transform $\mathbf{X}$, into another m-by-m matrix, $\mathbf{Y}$, such that for some m-by-m matrix $\mathbf{P}$

$$\mathbf{Y} = \mathbf{PX} \tag{3.11}$$

This is a change of basis, where the original data $\mathbf{X}$ is being projected on to the columns of $\mathbf{P}$. We must now decide what the best way of re-expressing the data is, and this depends on what features we wish $\mathbf{Y}$ to exhibit. A fundamental assumption of PCA is linearity, and secondly, the variables in the transformed matrix should be uncorrelated [16]. This means that the covariance matrix corresponding to $\mathbf{Y}$, $\mathbf{C}_Y$, should have off-diagonal entries as close to zero as possible. PCA also makes another assumption: Large variances represent important dynamics, while small variances represent noise [16]. Hence, the diagonal entries of $\mathbf{C}_Y$ should be

maximized, i.e. as close to one as possible. In conclusion, we have to choose $\mathbf{P}$ such that $\mathbf{C}_Y$ is a diagonal matrix.

We begin by writing $\mathbf{C}_Y$ in terms of $\mathbf{P}$, using Eq. 3.6 and 3.11:

$$\mathbf{C}_Y = \frac{1}{n-1}\mathbf{Y}\mathbf{Y}^T = \frac{1}{n-1}(\mathbf{P}\mathbf{X})(\mathbf{P}\mathbf{X})^T = \frac{1}{n-1}\mathbf{P}\mathbf{A}\mathbf{P}^T \tag{3.12}$$

where $\mathbf{A} = \mathbf{X}\mathbf{X}^T$, and $\mathbf{A}$ is symmetric[2], and related to the covariance matrix of $\mathbf{X}$. By using the Cayley-Hamilton theorem, which states that every square matrix is diagonalizable by an orthogonal matrix of its eigenvectors [16], we can write

$$\mathbf{A} = \mathbf{E}\mathbf{D}\mathbf{E}^T \tag{3.13}$$

$\mathbf{D}$ is a diagonal matrix. $\mathbf{E}$ is an orthonormal matrix of orthonormal eigenvectors of $\mathbf{A}$. We now decide what the transformation matrix P should be: By choosing the rows of $\mathbf{P}$ to be the eigenvectors of $\mathbf{A}$, we get $\mathbf{P} = \mathbf{E}^T$, and inserting this and 3.13 into 3.12 yields

$$\mathbf{C}_Y = \frac{1}{n-1}\mathbf{E}^T(\mathbf{E}\mathbf{D}\mathbf{E}^T)\mathbf{E} \tag{3.14}$$

For the orthonormal matrix $\mathbf{E}$, we get $\mathbf{E}^T\mathbf{E} = \mathbf{I}$, where $\mathbf{I}$ is the m-by-m identity matrix. We arrive at

$$\mathbf{C}_Y = \frac{1}{n-1}\mathbf{D} \tag{3.15}$$

By choosing $\mathbf{P}$ to be a matrix containing the eigenvectors of the covariance matrix $\mathbf{X}\mathbf{X}^T$, we have diagonalized $\mathbf{C}_Y$, which was the goal. The eigenvectors of the covariance matrix $\mathbf{X}\mathbf{X}^T$ form the new basis (coordinate system) for expressing the data, often in terms of two or three variables, called principal components (PC).

Now we turn our attention back on our data set, $\mathbf{X}$. Based on the assumptions made about the variance, the input to PCA should have a high SNR [16] (defined in Section 2.3.4). Let us assume that this is the case. Secondly, it is important to center the data, using e.g. *mean centering*[3], which involves subtracting the mean from each of the data dimensions (variables), see Eq. 3.1 and 3.10. This is a way for ensuring that the first principal component indeed is in the direction of maximum variance [16]. This is often considered the first step of PCA, although strictly speaking, it is generally done before PCA [17]. After mean-centering and finding the covariance matrix, the eigenvalues (sorted in descending order) and the corresponding eigenvectors, we project the data on to the eigenvectors, using Eq. 3.11, completing all the steps of PCA.

The result of projecting on to the first two prinicipal components (PC1 and PC2) is plotted in Figure 3.2. This representation is called a *score plot*. The scores are the

[2]Proof provided in [16, p. 11].

[3]If the variables are measured with different units, normalizing the data is also customary, but this is not the case here.

coordinates of the observations (spectra) expressed using the principal components. Ideally, scores from the same sample are grouped together in clusters, and from the clustering, we would conclude that the absorption spectra we started out with, come from three different materials, which is indeed the case. This is highlighted using different colors for the scores based on which material they belong to.

We can also plot the relationship between our original variables (frequencies) and the principal components, in a so called *loading plot*, showing how much each of the old variables contribute to the new ones. The higher the loading of a particular frequency is onto a PC, the more it contributes to that PC. This is plotted in Figure 3.3. Each frequency has been assigned a number, and we will give the corresponding frequencies for important loadings. From the loading plot, we look for frequencies that define directions in the score plot. We see that for the green cluster, frequency nr. 66 (1.38 THz) is the strongest contributor. For the red cluster, nr. 31 (0.80 THz), 32 (0.81 THz) and 33 (0.83 THz) stand out. For the blue cluster, frequency nr. 47 (1.01 THz) is dominating along PC1. These are frequencies corresponding to peaks in the absorption spectrum of particular materials (Lactose, RDX and Tartaric acid, respectively).

It is interesting to know how much of the total variance is accounted for by the principal components. This is done using a variance plot, shown in Figure 3.4 for the ten first eigenvectors/principal components. We see that when using two principal components (PC1 and PC2), approximately 85 % of the variance is preserved. When using three, this increases to approximately 95 %. The variance plot can also be used as a tool for deciding how many principal components one should include. It is common to look for sudden changes in the plot, we see one occurring at component two and another at component four. Data along directions with small variances (i.e. with small eigenvalues) can be omitted, reducing the dimensionality without losing too much information. This is what makes PCA useful in many applications. For our example data, the dimensionality can be reduced from 73 to two dimensions, retaining approximately 85 % of the variance.

In summary, PCA converts (possibly) correlated variables into a (smaller) number of uncorrelated variables (principal components), which are linear combinations of the original variables. The goal of PCA is [17]:

1) Extracting the ”most important” information from the data set
2) Reducing the dimensionality by keeping only this information
3) Expressing the information in a way that reveals the underlying structure; and
4) Analysing the structure

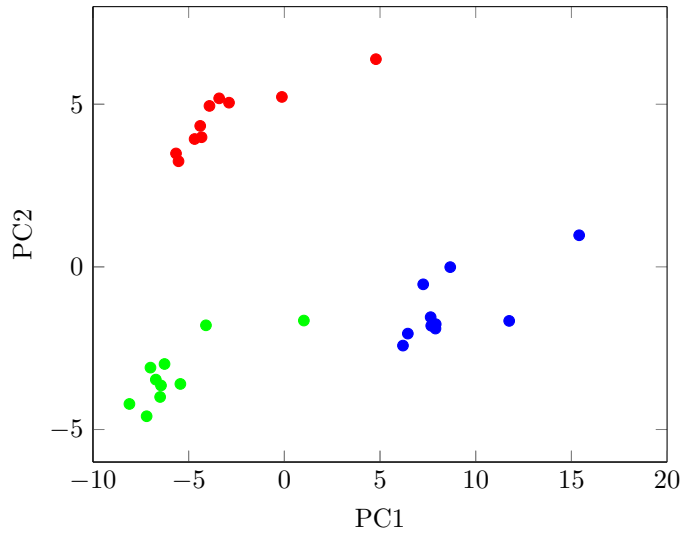


Figure 3.2: PCA example, score plot. Each score has been assigned a different colour based on which material it belongs to.

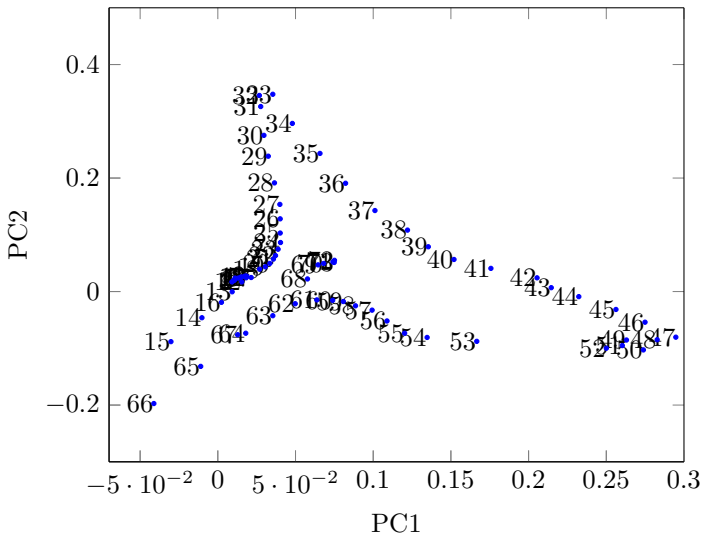


Figure 3.3: PCA example, loading plot. Nr. 66 corresponds to a frequency of 1.38 THz, nr. 31 - 33 range from 0.80-0.83 THz, and nr. 47 corresponds to a frequency of 1.01 THz.

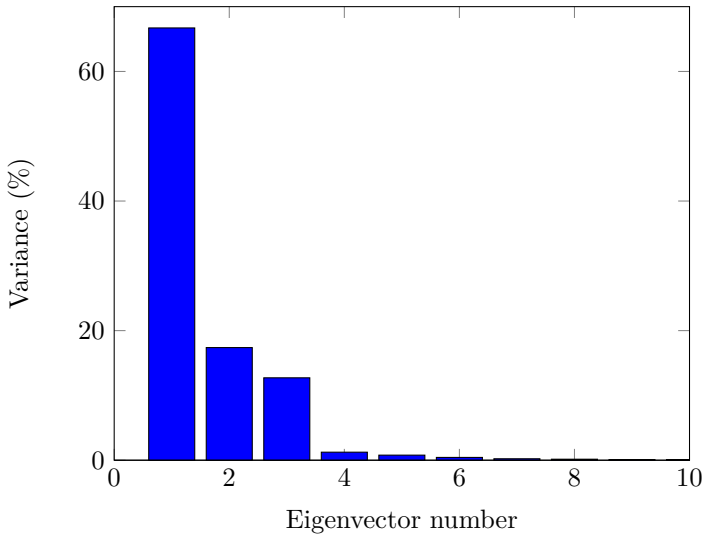


Figure 3.4: PCA example, variance plot.

3.3 Classification

3.3.1 Mahalanobis distance

PCA itself is not a classification method. We assigned different colours to the scores in Figure 3.2, but this was done based on the knowledge of which absorption spectra belonged to which material. PCA can, however, be used as a tool in classification. In order to do so, one needs *training data*. PCA is performed on the training data, and and some test data is projected on to the basis of the training data. This is called *PCA decomposition*. The training data depends on the application. In spectral identification, the training data would be based on some spectral characteristic of different materials (classes), and ideally, sufficient clustering occurs such that observations of different classes are separable when "enough" principal components are used. The number of principal components required increases with the number of classes [18] .

In order to classify each observation, we need a way for deciding which cluster of the training data each observation of the test data belongs to. The cluster is some distribution D, and the observation is a score P in the new coordinate system. It makes sense to decide this based on the distance P has to the center of D. However, the shape of the cluster could vary a lot, and simply measuring the Euclidean distance[4] does not take correlation into consideration. We need a measure for the distance which is unitless and takes into account the correlation of the data. The Mahalabonis distance is such a measure.

[4]The "ordinary" distance between two points given by the Pythagorean formula.

Along each principal component, the Mahalanobis distance is a multi-dimensional measure of how many standard deviations away some point P is from the mean of a distribution D. It is zero at the mean of D, and grows as P moves away from the mean. The Mahalanobis distance is defined as [8]

$$D_M(\mathbf{x}) = \sqrt{(\mathbf{x}-\mu)^T\mathbf{S}^{-1}(\mathbf{x}-\mu)} \tag{3.16}$$

for an observation $\mathbf{x} = (x_1, x_2..., x_m)^T$ from a group of observations with mean $\mu = (\mu_1, \mu_2, ..., \mu_m)^T$ and covariance matrix $\mathbf{S}$. The smaller D_M is, the closer the observation is to the mean of the distribution. By using some predefined threshold, it is possible to classify observations based on how small D_M is.

3.3.2 The curse of dimensionality

When increasing the number of principal components, the separability between different classes also increases. Intuitively, given a fixed number of classes, one would think that in order to improve the classification, one simply has to increase the number of principal components. However, it turns out that increasing the dimensionality only increases the performance of a classifier to a certain point, before a degrade of the performance occurs [19]. This is known as *the curse of dimensionality*, and is a result of *overfitting*. Overfitting occurs because the classifier starts learning exceptions specific to the training data [19]. Also, in highly dimensional spaces, measuring dissimilarities using distance measures, such as the Mahalanobis distance, becomes less effective due to increased sparsity of the training samples [19]. The optimal number of dimensions depends on the classifier and the amount of training data available. If needing to include more dimensions, the amount of training data has to be increased. More information regarding the curse of dimensionality is found in e.g. [19].

4 Spectral Angle Mapping

Similarity measures are widely applied in material recognition. Spectral angle mapping (SAM) is a common, powerful spectral recognition technique, where an unknown spectral characteristic is compared to a reference spectral characteristic. The unknown and the reference are treated as vectors, their dimensionality equal to the number of bands, n. The cosine of the angle between two vectors $\mathbf{a}$ and $\mathbf{b}$ is given by

$$cos(\theta) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|} \tag{4.1}$$

and yields a value between-1 and 1, where ± 1 means parallel vectors (the angle being 0) pointing in the same (+) or opposite (-) directions, and 0 means that the vectors are perpendicular. Considering two spectral vectors $\mathbf{s}$ and $\mathbf{s}_r$, $\mathbf{s}_r$ being the reference spectrum, this is referred to as *spectral correlation* [20], 1 meaning perfectly correlated spectra and 0 means uncorrelated spectra. A value of -1 for the correlation is unphysical, and is not encountered in practice. A visualization of the spectral angle θ is shown in Figure 4.1, for three-dimensional vectors $\mathbf{s}$ and $\mathbf{s}_r$.

By comparing the unknown spectrum to a library of references, each pixel in an unknown image can be assigned to the material it correlates strongest with, given that the correlation is above some predefined threshold.

A commonly used spectral characteristic is the absorbance over a range of frequencies, i.e. the absorption spectrum *A(f)*, defined in Section 2.3.3, but it is also possible to use derivatives, referred to as derivative spectroscopy [21]. In this thesis, the first-order derivative is considered. The first-order derivative, from here on referred to simply as the derivative, is the rate of change of absorbance with respect to frequency, *dA(f)/df*. Figure 4.2a shows a computer-generated example of two absorption spectra, each of them with a single peak. The corresponding derivatives, *dA(f)/df*, are plotted in Figure 4.2a. A peak in the absorption spectrum manifests itself as a curve with a top and a dip when differentiated. It starts and ends at zero, and passes through zero at the peak frequency of the absorption spectrum. As seen in Figure 4.2b differentiation removes baseline shifts, caused by e.g. instrumentation or Fresnel reflections, making spectral recognition easier.

Another reason for using the derivative, is that spectral discrimination can become easier, because small differences between nearly identical absorption spectra are accentuated [21]. However, this accentuation also happens for unwanted effects, e.g. noise and water absorption lines, making the use of a suitable window function essential [22] [21].

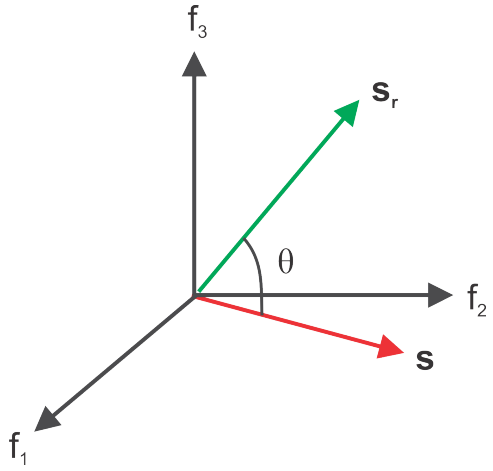
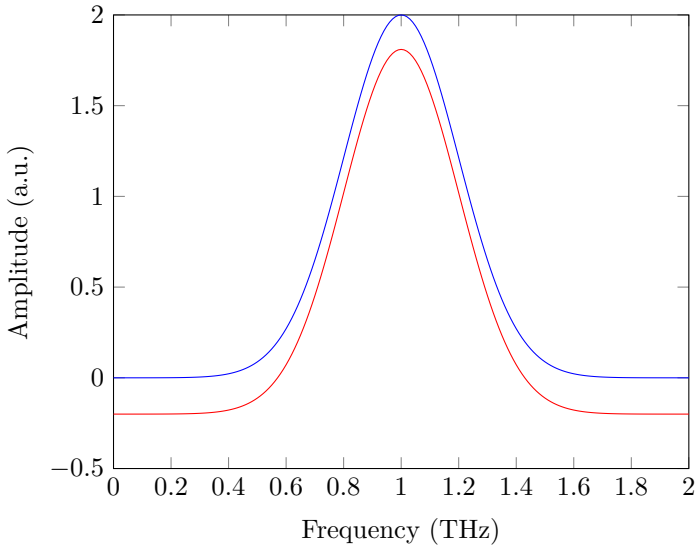
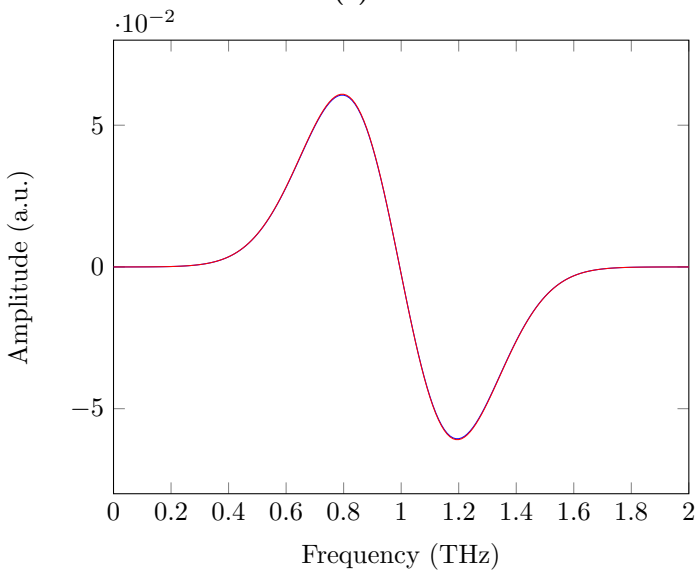


Figure 4.1: Visualization of the spectral angle θ between a spectral vector $\mathbf{s}$ (red) and a reference vector, $\mathbf{s_r}$ (green).



(a)



(b)

Figure 4.2: Computer generated examples of (a) two absorption spectra (b) the derivative of the absorption spectra in (a).

5 Receiver Operating Characteristic

A receiver operating characteristics (ROC) curve is a graphical plot that can be used to evaluate the performance of a classifier. A ROC curve plots the true positive rate (TPR) versus the false positive rate (FPR) at various threshold values.

A classifier that labels an instance of a test set either as positive or negative for a given threshold, has four possible outcomes [23]: A positive instance labelled as positive counts as a *true positive* (TP), which is equivalent with a *hit*. If labelled as negative, it is counted as a *false negative* (FN). A negative instance when labelled as negative, counts as a *true negative* (TN). If labelled as positive, a *false positive* (FP) occurs, equivalent with a *miss*. The ROC curves are generated by counting the fraction of true positives and false positives at various thresholds. For each threshold value, the TPR and FPR are given by

$$TPR = \frac{\sum TP}{P} \tag{5.1}$$

$$FPR = \frac{\sum FP}{N} \tag{5.2}$$

where P is the number of positive instances (total positives) and N the number of negative instances (total negatives). $P + N$ is equal to the total number of instances. Examples of ROC curves for two classifiers are plotted in Figure 5.1. The diagonal line from (0,0) to (1,1) represents the performance of a *random classifier*, for which each instance is randomly classified as either positive or negative. Such a classifier has no information about the class [23]. This line divides the ROC space: Points above the diagonal represent a good classification (better than random), and points below mean that the classification is worse than that of random guessing. A classifier below the diagonal can be said to have useful information, but applies the information incorrectly [23][24]. The upper left corner (1,0) represents perfect classification,with a true positive rate of 1, and a false positive rate of 0. The performance of a classifier can be evaluated by looking at how close the ROC curve lies to this point. In Figure 5.1, both classifiers perform significantly better than a random classifier, but Classifier 1 has a better performance than Classifier 2, because regardless of threshold, the TPR is higher while the FPR is lower.

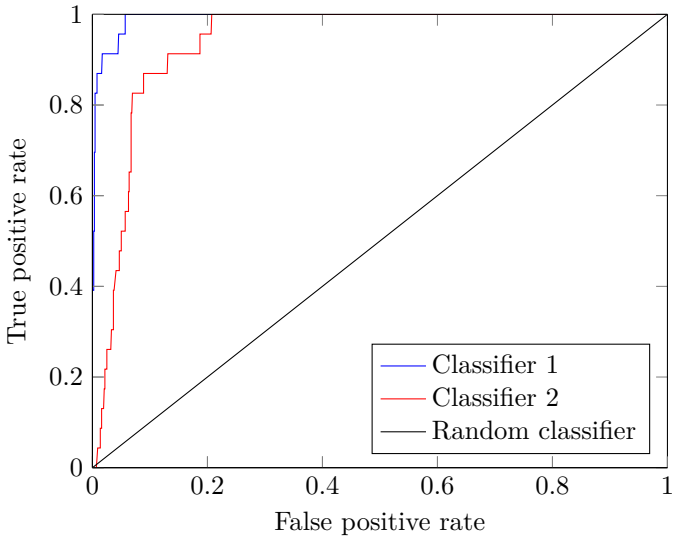


Figure 5.1: ROC curve examples.

6 Experiment

This chapter will describe the experimental setup, including THz-TDS, and a description of the samples and the scientific approach. The data processing in MATLAB is explained, including a description of the training data for PCA and reference spectra for SAM, before explaining how the data was analysed using the algorithms.

Measurements were performed using THz-TDS in transmission on samples containing Tartaric acid, Lactose or RDX, which all have spectral fingerprints in the THz-range. Tartaric acid and Lactose acted as simulants for dangerous substances, and RDX is an explosive widely used in military applications. The spectral characteristics investigated were the absorption spectrum, $A(f)$ and the derivative of the absorption spectrum $dA(f)/dt$, for two different window widths.

The robustness of both PCA and SAM was tested by looking at how well they perform

1) for different materials
2) for different samples containing the same material, i.e. when varying the thickness, concentration and grain size
3) under varying measurement conditions, i.e. covering the samples by a barrier

The performance was investigated and compared using ROC curves. The generation of these is described in Chapter 5.

6.1 Experimental setup

6.1.1 THz time-domain spectroscopy

All measurements were carried out with THz-TDS in transmission, using the setup in Figure 6.1. A brief description of the setup is provided, and more details can be found in [1]. THz pulses are generated by a commercial frequency-doubled, erbium-doped mode-locked femtosecond fiber laser (Toptica photonics), creating 150 fs pulses with average power ∼120 mW and repetition rate ∼90 MHz [1]. The pulses at a wavelength 780 nm then enter a grating stretcher, where they are

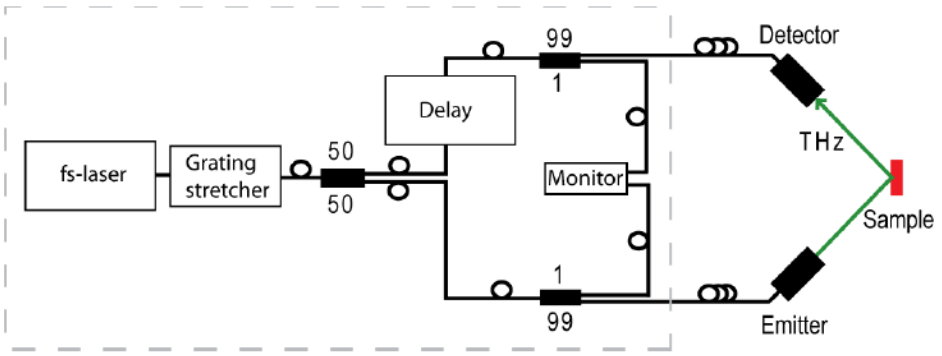


Figure 6.1: THz time-domain spectroscopy setup from [1, p. 2], with permission.

chirped before being coupled into a single-mode fiber. Chirping involves temporally stretching the pulse to a longer duration by a dispersive element (group velocity dispersion), and compensates for the dispersion of the fiber. A 50/50 coupler splits the pulse in two, and directs them to the emitter and detector head, respectively, which are based on photoconductive antennas. The detection pulse is delayed by an optical delay-arm. By sweeping the delay line, the THz pulse amplitude as a function of time is determined.

6.1.2 Samples

Measurements were carried out on samples containing RDX, Lactose and Tartaric acid mixed with Teflon. Teflon acts as a fill material and stabilizes the samples, and is suitable because it absorbs little in the relevant frequency range [1]. To minimize scattering effects, the particle sizes of the materials in the sample should be much less than the THz wavelength. The samples had therefore been ground to a fine powder, before having been mixed with Teflon and compressed into pellets. The weight fraction of each material is given in Table 6.1. The samples had been prepared beforehand. For more details, see [1, p. 2-3] and [20, p.1].

Six samples, listed in Table 6.1, were considered. Four different tartaric acid samples were used, all of them ground, except from Sample 6, which contained unground tartaric acid. The samples were placed in a sample holder taking 3x3 samples, see Figure 6.3. The bottom row was used for reference purposes, and are pure Teflon (bottom left), air (bottom middle) and a metal plate (bottom right).

6.1.3 Scientific approach

The sample holder was placed on a xy-stage which was scanned through the beam, capturing the THz waveform for each stage position (pixel). The xy-stage had a

Table 6.1: Samples and sample properties.

Sample No.	Material	Weight percent	Thickness	Condition
1	Tartaric acid	10 %	4 mm	Ground
2	Lactose	10 %	4 mm	Ground
3	RDX	10 %	4 mm	Ground
4	Tartaric acid	5 %	4 mm	Ground
5	Tartaric acid	10 %	1 mm	Ground
6	Tartaric acid	5 %	4 mm	Unground

Table 6.2: Measurement parameters and conditions.

No.	Cover	nx	ny	Scan time (h)	Temp. (°C)	Rel. hum. (%)
1	-	64	64	80	22.1	46.5
2	-	30	30	17.5	23.2	48.6
3	Paper	30	30	17.5	22.6	52.5
4	Plastic	30	30	17.5	22.2	52.2
5	Cloth	30	30	17.5	22.1	44.4

travel range of 150x150 mm, and the number of positions in each of the directions is given by n_x and n_y, respectively. The step size in each direction is thus given by $150/nx$ mm and $150/ny$ mm. The total scan range for the measurements was $\sim$ 70 ps, with a velocity of 1 ps/s, such that the measurement duration at each transverse position was 70 s. Before imaging, the temperature and relative humidity were measured, and a reference measurement in air was performed. The noise was also measured by covering the beam path with a metal plate. The amplitude spectrum of a measurement in air and with a blocked beam is shown in Figure 6.2, between 0 and 4 THz. We see that between approximately 0.1 and 2 THz, the reproducibility of the signal is good. The THz bandwidth is about 3 THz, and the peak SNR is approximately 60 dB[1].

Five THz images were acquired: two with no coverage of the samples, and three where the samples were covered by a barrier. The barriers used were paper, plastic and cloth. A standard A4 sheet of paper was used. The plastic was 1 mm thick, while the cloth was $\sim$3 mm thick. Table 6.2 lists the measurement parameters and conditions (temperature, relative humidity) for each of the THz images. An image without a barrier (Image 1) was used as a basis for training data for PCA, and for reference spectra for SAM. Only pixels from the top row samples were used, i.e. Sample 1, 2 and 3, all with a weight fraction of 10 % and a thickness of 4 mm. These will be referred to as *training samples* in the context of PCA, and

[1]This corresponds to an amplitude SNR of 1000.

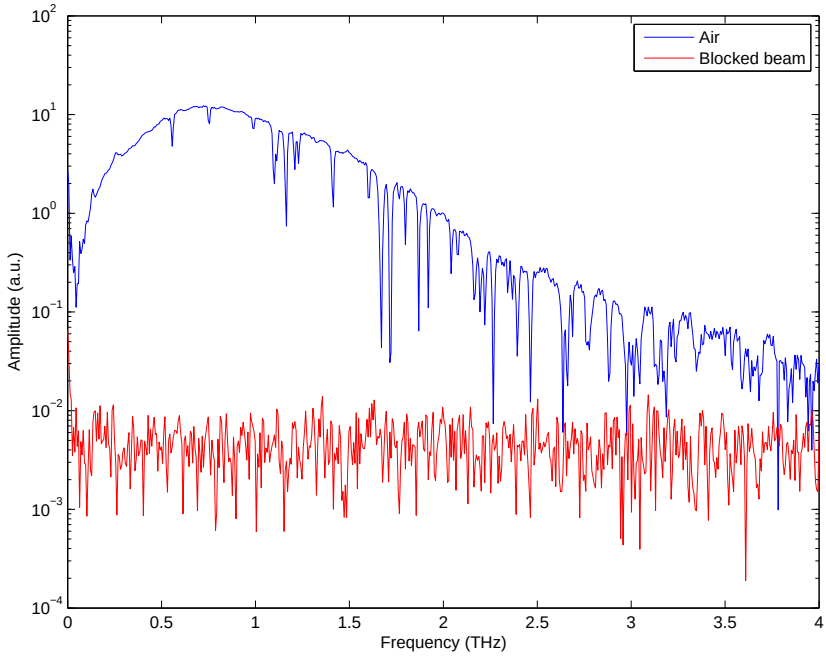


Figure 6.2: Establishment of noise floor, showing the amplitude spectrum of transmission through air and a measurement with blocked beam. The absorption lines seen stem from absorption by water vapour in the air.

Figure 6.3: Sample holder with samples, numbered from 1-6, according to Table 6.2.

reference samples regarding SAM. The other THz images were used for testing the algorithms, and will be referred to as *unknown images*, or test data. The samples of the unknown images are occasionally referred to using the sample numbers from Table 6.1.

6.2 MATLAB

MATLAB R2013 was used for processing the data. The MATLAB code is given in Appendix A.

The measurements described in the previous sections contain a signal for each pixel in the THz image, which is the electric field as a function of time. A Blackman-Harris window, centered on the signal peak, was multiplied with the signal. Two different half widths were used, δt =15 ps and δt = 30 ps. The DFT of the windowed signal was taken, yielding the THz spectrum. Only the part of the spectrum spanning from 0.3 to 1.5 THz was used, due to a high SNR in this region (see Figure 6.2). 85 samples of the signal exist in this frequency interval. The spectrum in air was calculated equivalently, and by dividing the spectrum of the signal with the spectrum in air, the transmission spectrum was obtained. The absorption spectrum was then obtained using Eq. 2.6, and its derivative by differentiation with respect to f. These were in turn used used for training data/reference spectra in the case of Image 1, and as input for both algorithms for Images 2-5. 26 pixels of Lactose, 23 pixels of RDX and a total of 98 pixels of Tartaric acid exist in Images 2-5.

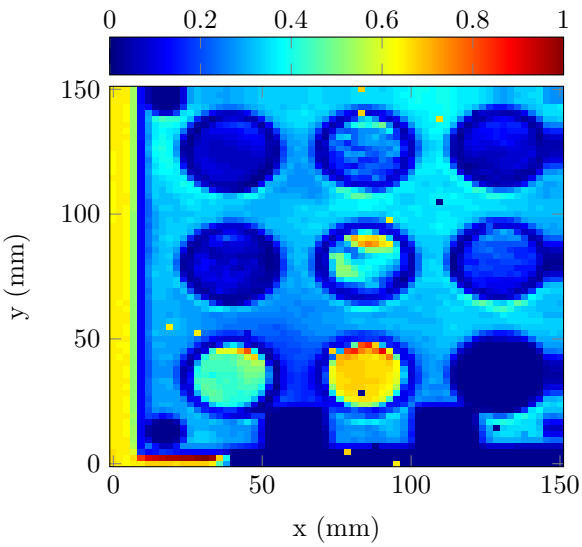


Figure 6.4: Spectral energy density of Image 1 between 0.3 and 1.5 THz.

Only pixels belonging to a sample (not the surroundings, i.e. sample holder and air) were used as training data for PCA. A THz image of Image 1 is shown in Figure 6.4, acquired by considering the spectral energy density of each pixel in the relevant frequency range.

The pixels belonging to a training sample (Sample 1, 2 and 3) were chosen automatically based on the delay of the signal compared to the reference measurement in air, see Figure 6.5. Because the sample holder is thicker than the samples, the delay through the sample holder, in the order of 30 ps, is noticeably bigger than through the samples, where it is ∼6 ps for the training samples (top row).

The pixels were then sorted according to their placement in the THz image, grouping pixels belonging to each of the training samples together. The spectral characteristics were calculated, before performing PCA on all of the training data together.[2] PCA decomposition was then applied on each of the two spectral characteristics of the unknown images.

For SAM, the spectra of the unknown images were normalized before being elementwise multiplied with the normalized reference sample spectra, see Eq. 4.1.

[2]It is possible to perform PCA decomposition on data for each training sample separately, referred to as base class [18]. In this manner, each training sample would have its own PC space, and the unknown image would be decomposed on to each of these, and tested against each base class separately. This approach is mainly of interest when having a significant number of different classes (materials) [18].

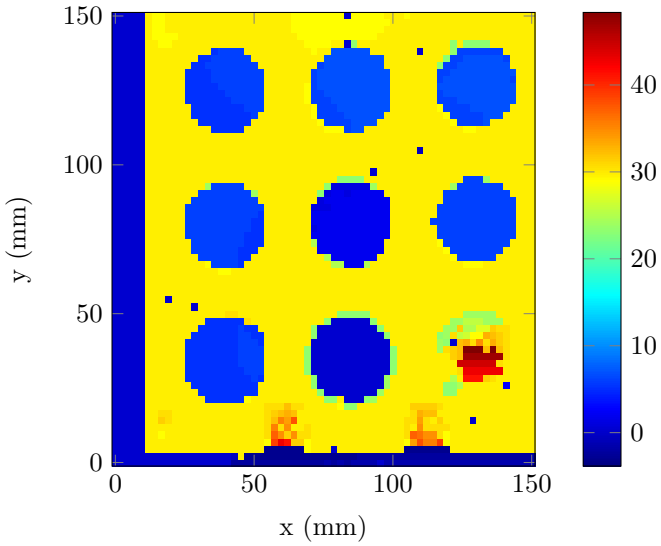


Figure 6.5: Pulse delay is ps of Image 1 compared to air.

7 Results and discussion

7.1 Spectral angle mapping

7.1.1 Reference spectra

The absorption spectra of Lactose, Tartaric acid and RDX (Sample 1, 2 and 3, respectively) are plotted in Figure 7.1, for two window widths $\delta t = 30$ ps and $\delta t = 15$ ps.

For Tartaric acid, a "double top" occurs, which is not characteristic for the material. Typically, Tartaric acid only has a single absorption peak in this frequency range, occurring at 1.1 THz [1]. By comparison of the absorption spectrum with the spectrum of transmission through air in Chapter 2, Figure 2.4, it is evident that two water absorption lines are present between 1.1 and 1.2 THz, and based on this, the "extra" peak for Tartaric acid probably occurs due to absorption by water vapour in the air.

For RDX, an absorption peak is observed at $\sim$0.8 THz. Another, less pronounced peak is present at $\sim$1.05 THz. Lactose has two characteristic peaks, the first occuring at $\sim$0.51 THz and the second at $\sim$1.4 THz. Both RDX and Lactose show similar characteristics to those obtained by others, see [1].

Figure 7.2 shows the derivative of the absorption spectra (from Figure 7.1) of Tartaric acid, Lactose and RDX . Lactose and RDX have been plotted with an offset. Again note the somewhat atypical characteristics for Tartaric acid: Typically, the curve would resemble those of Lactose and RDX more. However, with the absorption spectrum of Tartaric acid having a double peak in this case, an extra top and dip occurs using the derivative.

It is observed that noise is not an issue for either windows, even the wider one with $\delta t = 30$ ps, but using a narrower window smooths out the absorption spectra (and as a result, also their derivative) more. This is most clear in the case of Tartaric acid. The water absorption line is less pronounced, but is still observed both using the absorption spectrum and its derivative. The trade-off using a narrower

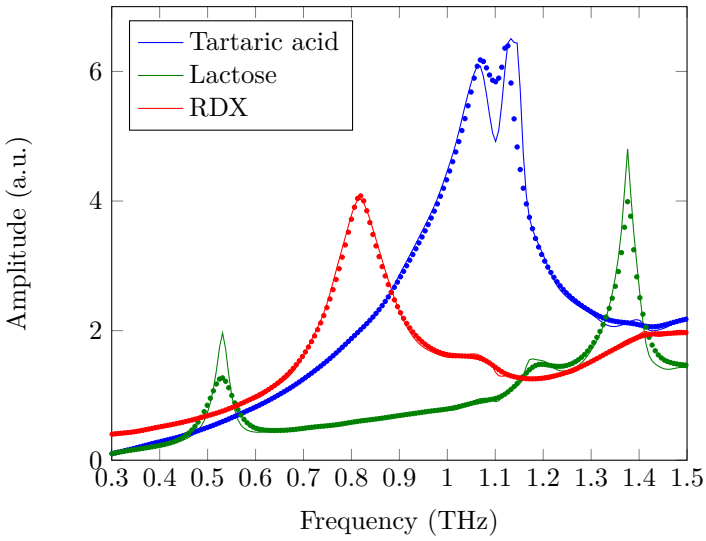
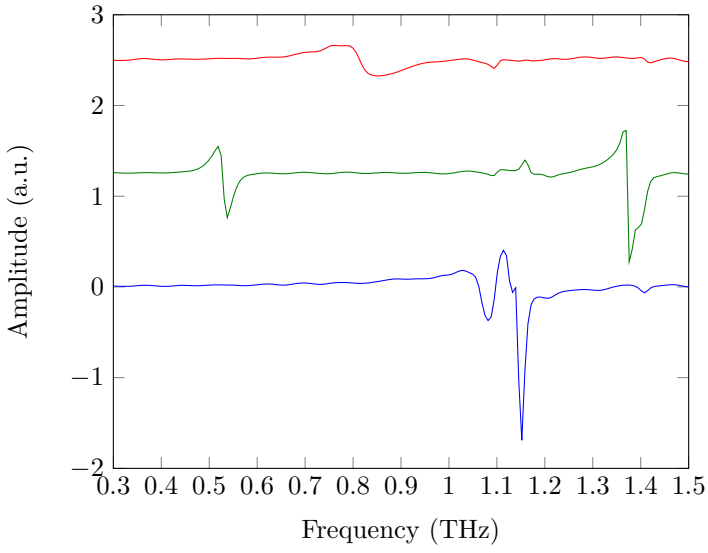
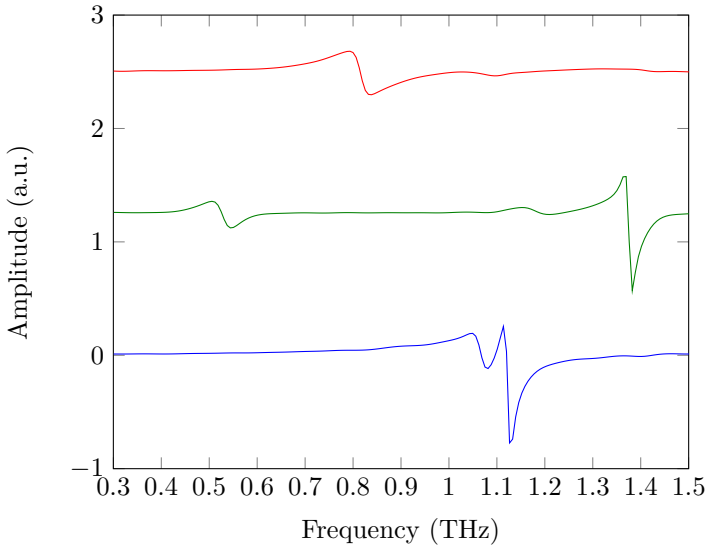


Figure 7.1: Absorption spectra of Tartaric acid, Lactose and RDX, after applying a Blackman-Harris window with half width 30 ps (line) and 15 ps (dots) on the data in time domain.

window is generally a lower amplitude and wider peak, which most easily seen from the absorption spectrum of Lactose. Overall, it is observed that the optimal width of the window will depend on the material.



(a)



(b)

Figure 7.3: *dA(f)/df* of Tartaric acid (blue), Lactose (green, plotted with an offset along the y-axis) and RDX (red, plotted with an offset along the y-axis), after applying a Blackman-Harris window with half width (a) 30 ps (b) 15 ps on the data in time domain.

7.2 Spectral correlation images

The results of applying SAM on the unknown images are presented here, where the correlation of each pixel's spectral characteristic with the reference spectral characteristic is represented in a spectral correlation image. In this context, "the surroundings" refer to pixels in the image which do not belong to the material under consideration.

7.2.1 No barrier (Image 2)

Figure 7.4 shows the spectral correlation of Image 2 with Tartaric acid, comparing the use of *A(f)* and *dA(f)/df* and window widths of $\delta t = 30$ ps and $\delta t = 15$ ps.

If *A(f)* is used, both window widths yield essentially the same results. All Tartaric acid samples are identified, evident by higher correlation values than those of the surroundings. It is also evident that the correlation is highest for sample 1 and 4, and somewhat lower for sample 5 and 6 (see Figure 6.3). This makes sense considering the sample properties: Sample 1 is the same sample as the reference sample, making spectral recognition easy (both 10 %, 4 mm and ground). Sample 4 has the same thickness and grain size as the reference sample, but a lower concentration (5 %). Theoretically, the absorbance is linear with concentration (Eq. 2.6), hence the effect of reducing the concentration is simply a reduction of the amplitude at each wavelength, which does not affect spectral recognition, because the spectral vectors are normalized (see Eq. 4.1). In practice, however, it has been established that the relationship is non-linear to some degree, due to measurement uncertainties if the attenuation reaches the noise floor [1, p. 7], but this effect is not of great importance here.

Using *dA(f)/df* yields better results with $\delta t = 15$ ps compared to $\delta t = 30$ ps. This can be explained by a smoother curve and less defined water absorption line. The spectral correlation is also higher using $\delta t = 15$ ps for all Tartaric acid samples. Especially the thin sample (Sample 3) and the unground sample (Sample 4) benefit from a narrower window.

Overall, the unground sample (Sample 6) is the most challenging sample to recognize for SAM, due to scattering effects which make the peak less pronounced [1]. Using a narrow window in combination with *dA(f)/df* seems to be a good combination for reducing these effects, and also yields the best contrast of all the four cases.

For Lactose (Figure 7.5) and RDX (Figure 7.6) the use of *A(f)* and *dA(f)/df* are compared for $\delta t = 30$ ps. Using $\delta t = 15$ ps yields similar results, which are shown in Appendix B, Figure B.1 and B.2, respectively. For both materials, using both spectral characteristics, the spectral correlation is higher for the correct sample compared to the surroundings, but the use of *dA(f)/df* results in a higher contrast.

The results for using a paper barrier (Image 3) and plastic barrier (Image 4) yield similar results as presented above for all materials, and can be found in Appendix B, Figures B.3 - B.5 (paper barrier) and Figures B.6 - B.8 (plastic barrier). Both are ”easy” barriers, meaning they absorb little and do not have spectral fingerprints in the relevant frequency range, which explains the similarity.

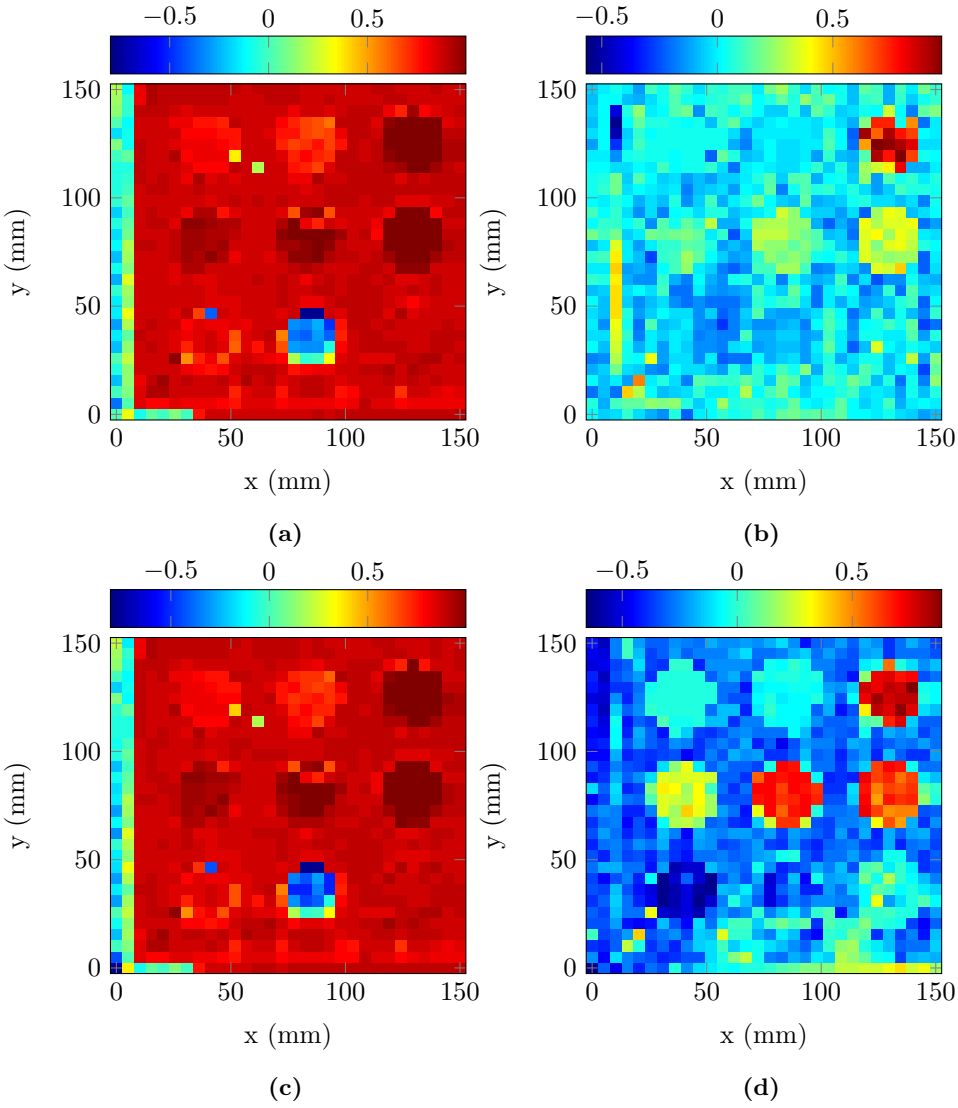


Figure 7.4: Spectral correlation of Tartaric acid using (a) $A(f)$ and $\delta t = 30$ ps (b) $dA(f)/df$ and $\delta t = 30$ ps (c) $A(f)$ and $\delta t = 15$ ps (d) $dA(f)/df$ and $\delta t = 15$ ps. No barrier.

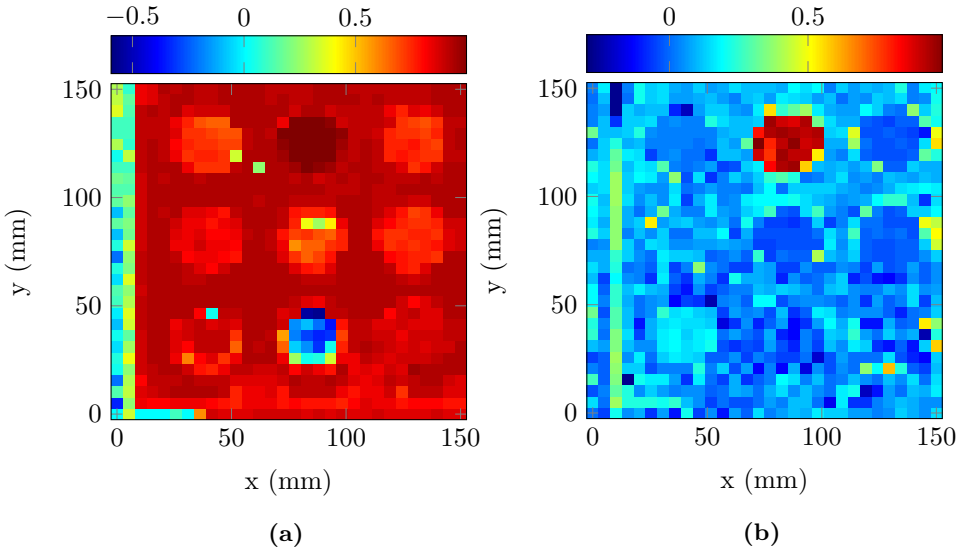


Figure 7.5: Spectral correlation of Lactose using $\delta t = 30$ ps and (a) *A(f)* (b) *dA(f)/df*. No barrier.

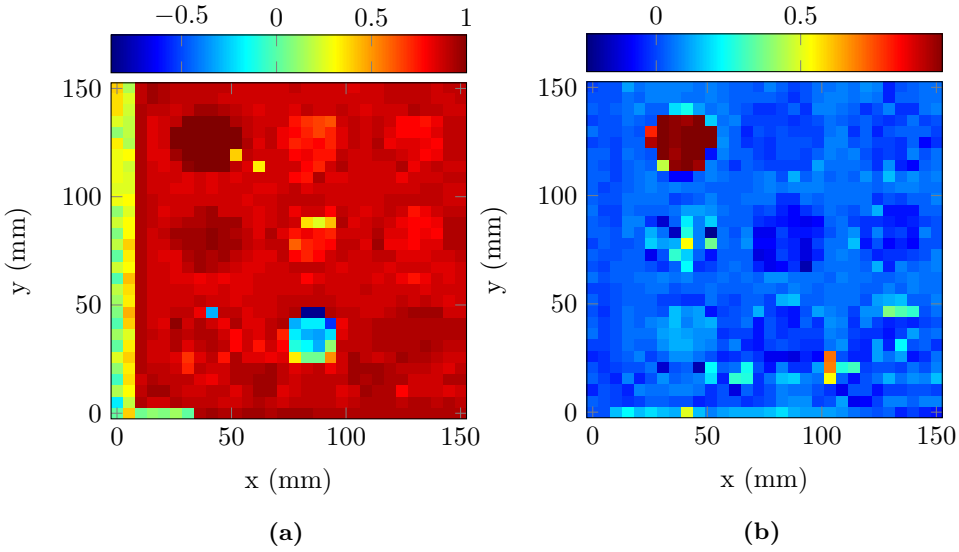


Figure 7.6: Spectral correlation of RDX using $\delta t = 30$ ps and (a) *A(f)* (b) *dA(f)/df*. No barrier.

7.2.2 Cloth barrier (Image 5)

Figure 7.7 shows the results of applying SAM to identify Tartaric acid when the samples are covered with a piece of cloth. Using *A(f)* still yields good results for both window widths. When using *dA(f)/df* and the wider window, the performance is noticeably reduced, especially for Sample 5 and 6, which are hardly distinguishable from the surroundings. Using the narrower window has good effect, but the unground sample is still a challenge, because some of the correlation values are comparable to those of other samples.

The results for Lactose and RDX are shown in Appendix B, Figures B.9 and B.10, respectively. The same tendencies regarding choice of spectral characteristic and window width as when using no barrier are observed, but with somewhat reduced correlation values and a smaller contrast, especially for Lactose. The reduced correlation using cloth compared to no barrier or the other two barriers is expected. The cloth barrier is more challenging because it is thicker than the paper and plastic barriers used, causing more attenuation and an extra delay. However, the spectral correlation images indicate that SAM still works very well, and for RDX, the images are essentially identical.

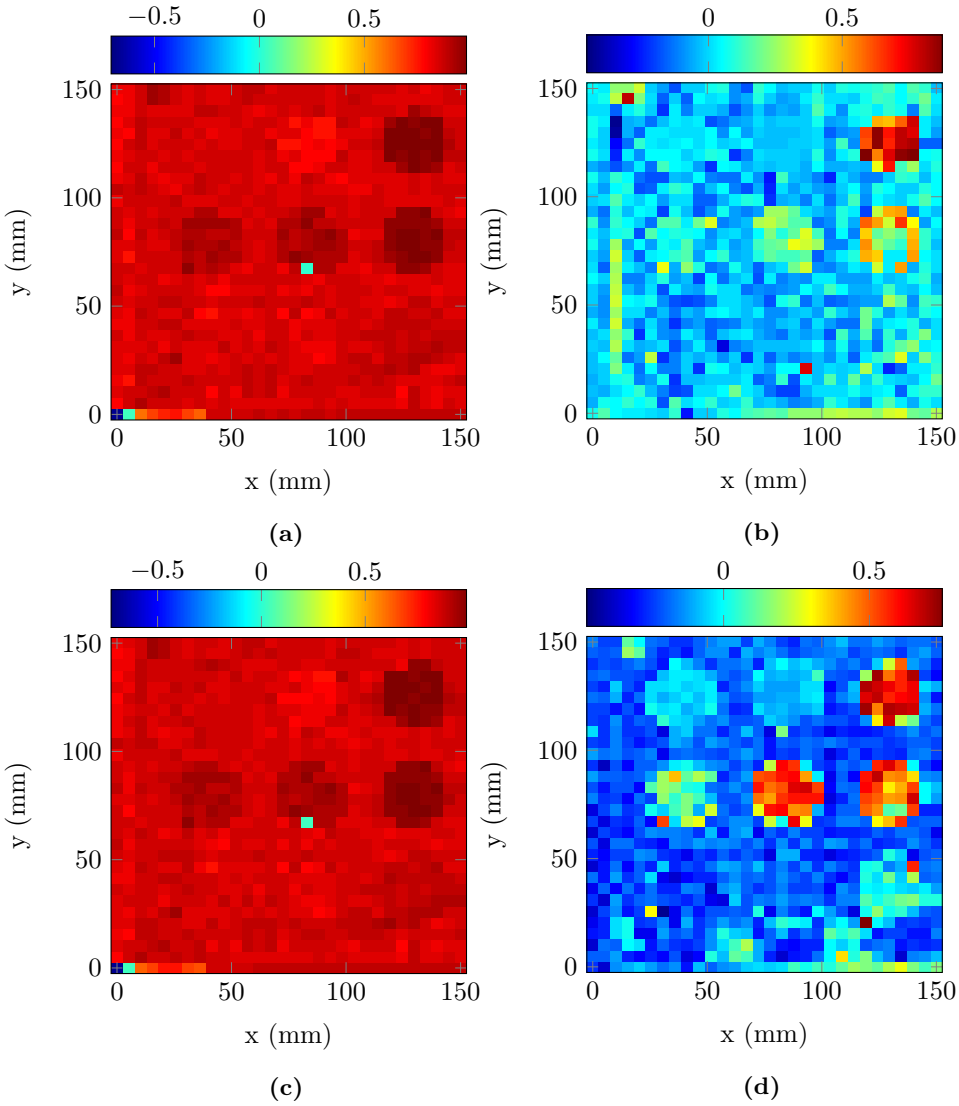


Figure 7.7: Spectral correlation of Tartaric acid with cloth barrier using (a) $A(f)$ and $\delta t = 30$ ps (b) $dA(f)/df$ and $\delta t = 30$ ps (c) $A(f)$ and $\delta t = 15$ ps (d) $dA(f)/df$ and $\delta t = 15$ ps.

7.3 ROC curves (SAM)

7.3.1 No barrier (Image 2)

Figure 7.8 shows ROC curves for Tartaric acid when no barrier is used. As expected from the corresponding spectral correlation images in Figure 7.4, essentially the same performance is obtained using *A(f)* with both window widths, and when using *dA(f)/df*, the narrower window is better, for reasons discussed in Section 7.2.1. We see that compared to using the absorption spectrum, we can get a true positive rate (TPR) of close to 1 without increasing the false positive rate (FPR) considerably, compared to when the absorption spectrum is used.

For Lactose, the ROC curves are plotted in Figure 7.9 for both spectral characteristics and window widths. We see that in all four cases, the performance is very good, with the curve being close to the upper left corner. In this case, the best performance is obtained using the absorption spectrum in combination with the narrower window, which yields a nearly ideal ROC curve, i.e. a nearly perfect classification. This is a little surprising considering Figure 7.2, which shows that the effect of narrowing the window was not really effective in terms of smoothing out the absorption spectrum. However, one must keep in mind that we only see the reference absorption spectrum. The unknown image contains many pixels for Lactose, and each spectrum will vary somewhat, causing some variation in the correlation values, especially when using the derivative of the absorption spectrum. It was also not recorded at the same time as the reference image, and as can be seen in Table 6.2, the relative humidity was slightly higher when recording Image 1. The second peak of Lactose at $\sim$1.4 THz is prone to being affected by water absorption, which is evident from Figure 7.2 and Figure 2.4. The latter also shows that using a window with $\delta t = 15$ ps smooths out the water absorption line at $\sim$1.4 THz nicely.

The ROC curves for RDX are plotted in Figure 7.10 for both spectral characteristics and window widths. We see that regardless of choice of spectral characteristic and window width, all four curves are nearly ideal. If we look at *A(f)* and *dA(f)/df* in Figure 7.1 and 7.2, we see that RDX has nice and smooth curves for all cases, and changing the window has little effect. However, as already mentioned in the case of Lactose, only looking at a single spectrum is not sufficient for explaining the behaviour. A number of spectra from RDX were therefore investigated, and it was observed that 1) The spectra were very similar to the reference spectrum, which is expected because the RDX sample was identical to the sample from which the reference spectrum was obtained. 2) There was little variation from pixel to pixel, which is a result of a high SNR in the relevant frequency range, as well as a homogeneous sample. These two things are also evident in all the four spectral correlation images (Figure 7.6 and B.2), where we see that all RDX correlation values are high and similar to each other, regardless of barrier.

The ROC curves for identification of the materials using a paper and plas-

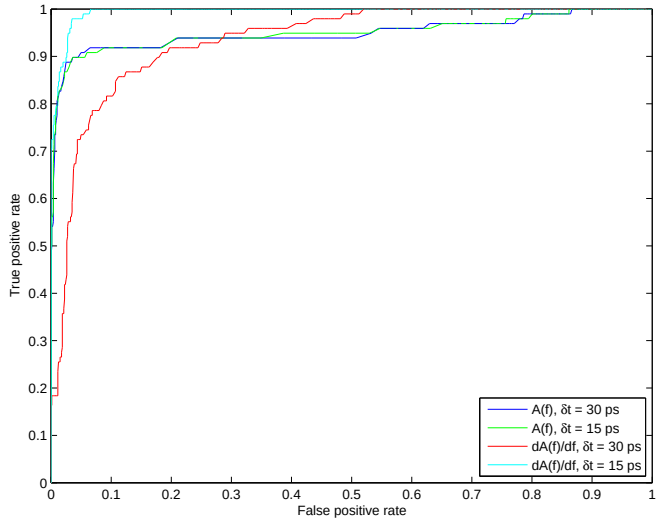


Figure 7.8

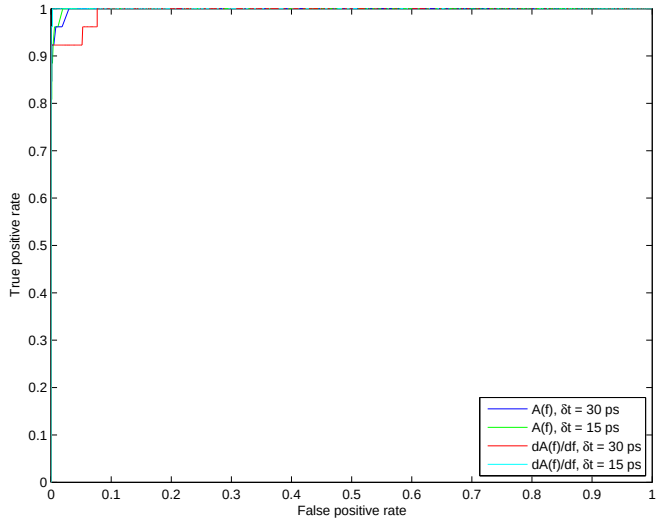


Figure 7.9: ROC curves for identification of Lactose using SAM, for the two spectral characteristics and window widths. No barrier.

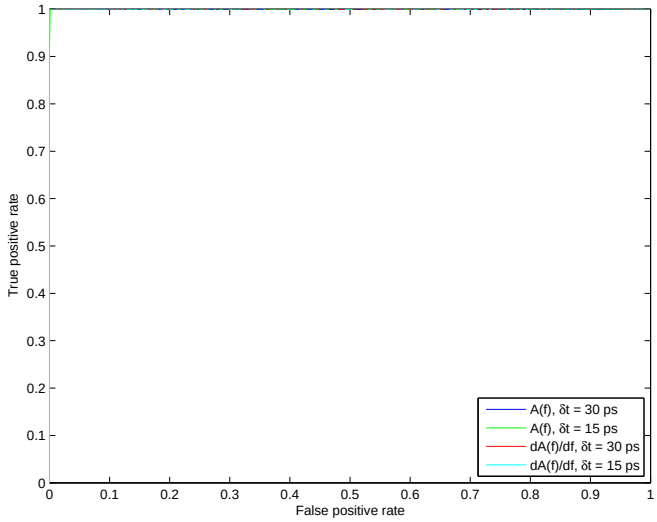


Figure 7.10: ROC curves for identification of RDX using SAM, for the two spectral characteristics and window widths. No barrier.

tic barrier are given in Appendix C.1 and C.2, respectively. For RDX, nearly ideal curves are still obtained, while for Tartaric acid and Lactose, there are some minor differences. In brief, in some cases, introducing a barrier leads to a better performance compared to when no barrier is used. One would expect that regardless of spectral characteristic and window width, the performance would be slightly reduced with a barrier. Taking into consideration that a single sheet of paper and a thin layer of plastic are quite "easy" barriers, the increased performance could be coincidental, because the resolution of the THz images is relatively small. This is further discussed in Appendix C.1.

7.3.2 Cloth barrier (Image 5)

The ROC curves for identification of Tartaric acid when the samples are covered by a piece of cloth, are shown in Figure 7.11. We observe a decrease in performance compared to using no barrier, when dA(f)/df is used, which is expected based on the spectral correlation images (Figure 7.4 and Figure 7.7). The performance is still quite good, especially using the narrower window. Using *A(f)* , an increase in the performance is observed, mainly due to a reduced number of false positives, i.e. caused by fewer pixels of the surroundings (e.g. the sample holder) being identified as Tartaric acid. The sample holder is thick compared to the samples, but still some of the radiation is transmitted, and the cloth will lead to an extra attenuation of the signals through it, which might be beneficial, even though it also attenuates the signals through the sample.

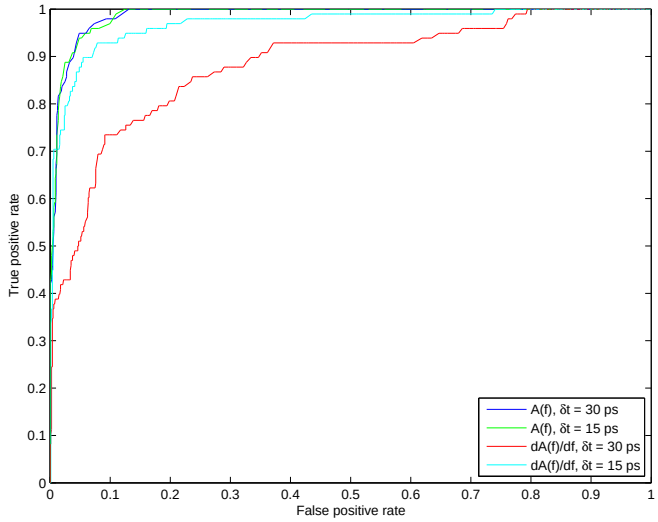


Figure 7.11: ROC curves for identification of Tartaric acid using SAM, for the two spectral characteristics and window widths. Cloth barrier.

Figure 7.12 shows the ROC curves for identification of Lactose with the cloth barrier. For all cases except using $A(f)$ and $\delta t = 15$ ps, a decrease in performance is observed, mainly due to an increased number of false positives. The performance is still very good considering that the the true positive rate is above 0.9 for all cases without having a noticeable increase in the false positive rate.

The ROC curves for RDX are found in Appendix C, Figure C.7. Even with the cloth barrier, there is no noticeable decrease in the performance compared to when no barrier was used. This is a little surprising, but when looking at the spectral correlation image (Figure B.10), we see that the sample is clearly identified in all four cases, even though the contrast is slightly reduced compared to when no barrier is used (Figure 7.6).

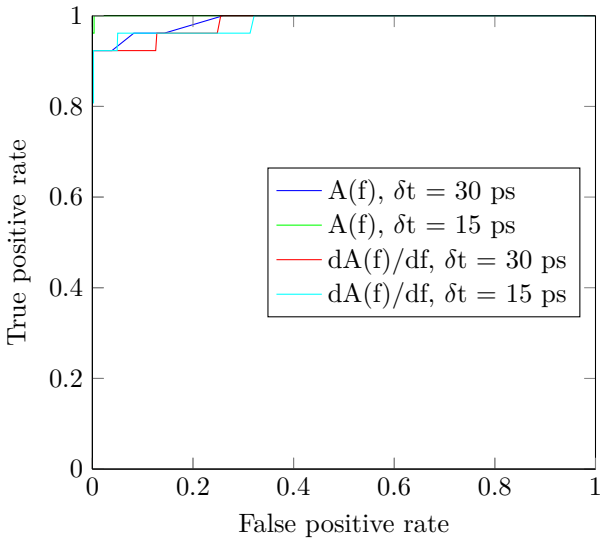


Figure 7.12: ROC curves for identification of Lactose using SAM, for the two spectral characteristics and window widths. Cloth barrier.

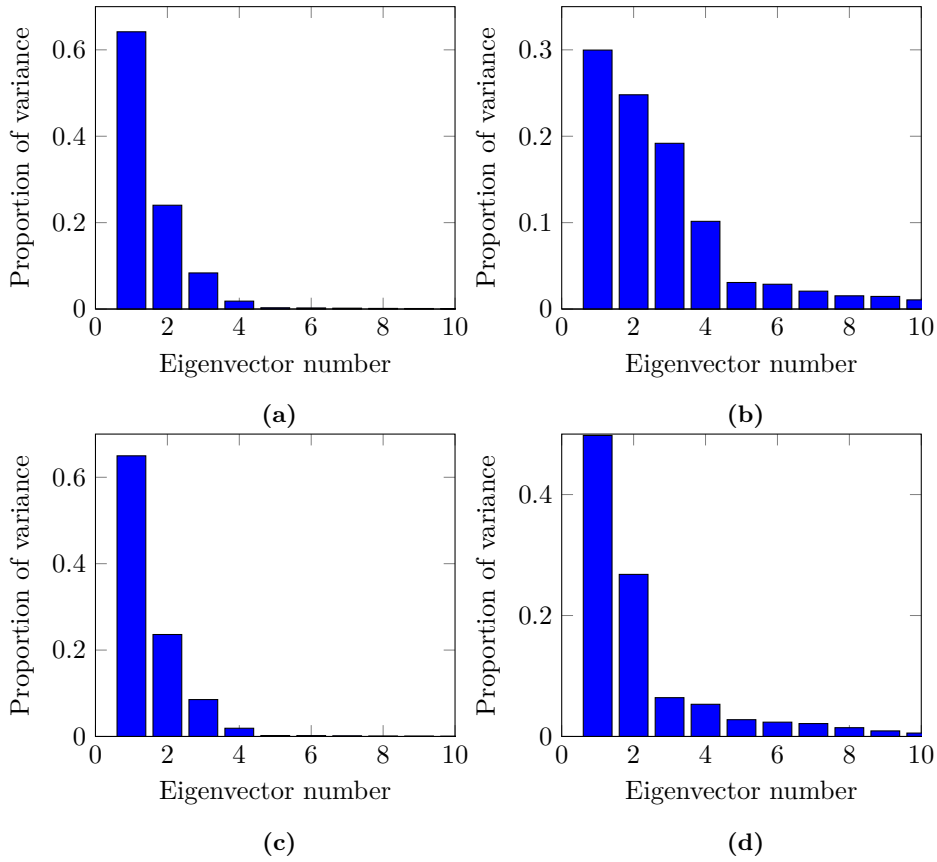


Figure 7.13: Fraction of total variance accounted for by each of the ten first principal components using (a) A(f) and $\delta t = 30$ ps (b) dA(f)/df and $\delta t = 30$ ps (c) A(f) and $\delta t = 15$ ps (d) dA(f)/df and $\delta t = 15$ ps.

7.4 Principal component analysis

7.4.1 Training data

Recall from Chapter 3 that a score plot is a projection of the data onto the principal components (the new subspace/coordinate system), and shows relations among observations.

The training data is an important part of the classification based on PCA, and should therefore be optimized. The variance plot is used as a tool for deciding how many principal components should be included, and in this case it is also used in the process of determining which training data should be used (in combination with other plots). Figure 7.13 shows how much of the total variance is accounted for by projection of the data onto each of the ten first principal components for

the potential training data. We see that when using the absorption spectrum *A(f)* (Figure 7.13a and 7.13c), the results are quite similar for both windows. Approximately 65 % of the total variance is preserved when projecting the data on to the first principal component alone. By taking into account the first three principal components, yielding a reduction from 85 to three dimensions, roughly 90 % of the variance is accounted for.

When using *dA(f)/df* (Figure 7.13b and 7.13d), the choice of window is more important. More principal components are required when using $\delta t = 30$ ps in order to represent the same amount of variance as when using $\delta t = 15$ ps. The latter accounts for ∼85 % of the variance when projecting on to the three first principal components, the corresponding amount being ∼75 % for $\delta t = 30$ ps. Using the absorption spectrum in combination with either window explains more of the variance with fewer principal components than the derivative.

Figure 7.14 shows the training data projected on to the two first principal components (PC1 and PC2). As already indicated by the variance plots (Figure 7.13), the choice of δt is not significant when using *A(f)* (Figure 7.14a and 7.14c). Regardless of the width of the window, two principal components are sufficient for separating the samples from each other. It is evident that projection onto PC1 would be sufficient for separating Tartaric acid and Lactose from each other, because the clusters mainly lie on either side of PC1 = 0. PC2 alone would separate Lactose from RDX, the clusters separated by PC2 = 0.

dA(f)/dt (Figure 7.14b and 7.14d), on the other hand, is clearly affected by the choice of window, as already indicated by the variance plots. Even though clustering is observed in both cases, using a window of 15 ps seems to separate the data better, and contains no outliers.

7.4.2 Score plots

The results of applying PCA decomposition on the unknown images are presented using score plots. Each material has been assigned a colour, but this is not the result of the classification, only a representation of how the points belonging to different materials are distributed after being decomposed onto the new coordinate system based on the training data. The yellow scores belong to the surroundings.

Figure 7.15 shows the score plot of Image 2. We see that when using *A(f)*, similar results are obtained using both window widths. Clustering is present, but the scores of the surroundings overlap the material clusters somewhat. These overlapping scores are mainly from the sample holder, while the scores not interfering with the materials' clusters (PC1 > 20) are air from the left side of the THz image, and from the air reference in the middle of the bottom row (see Figure 6.4)

When using *dA(f)/df* and $\delta t = 30$ ps, the scores of the surroundings completely overlap the scores of RDX. Using $\delta t =$ 15 ps seems a little better, but we see

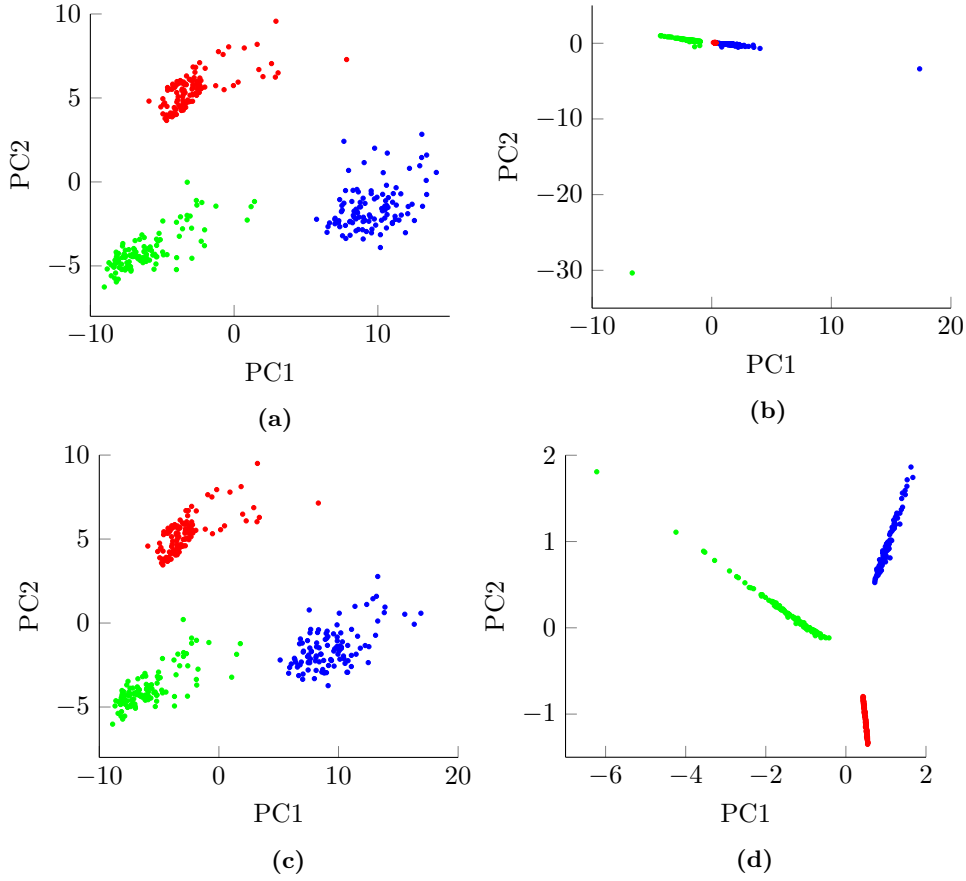


Figure 7.14: Score plot, showing PC1 vs. PC2 for Tartaric acid (blue), Lactose (green) and RDX (red) using (a) *A(f)* and $\delta t = 30$ ps (b) *dA(f)/df* and $\delta t = 30$ ps (c) *A(f)* and $\delta t = 15$ ps (d) *dA(f)/df* and $\delta t = 15$ ps.

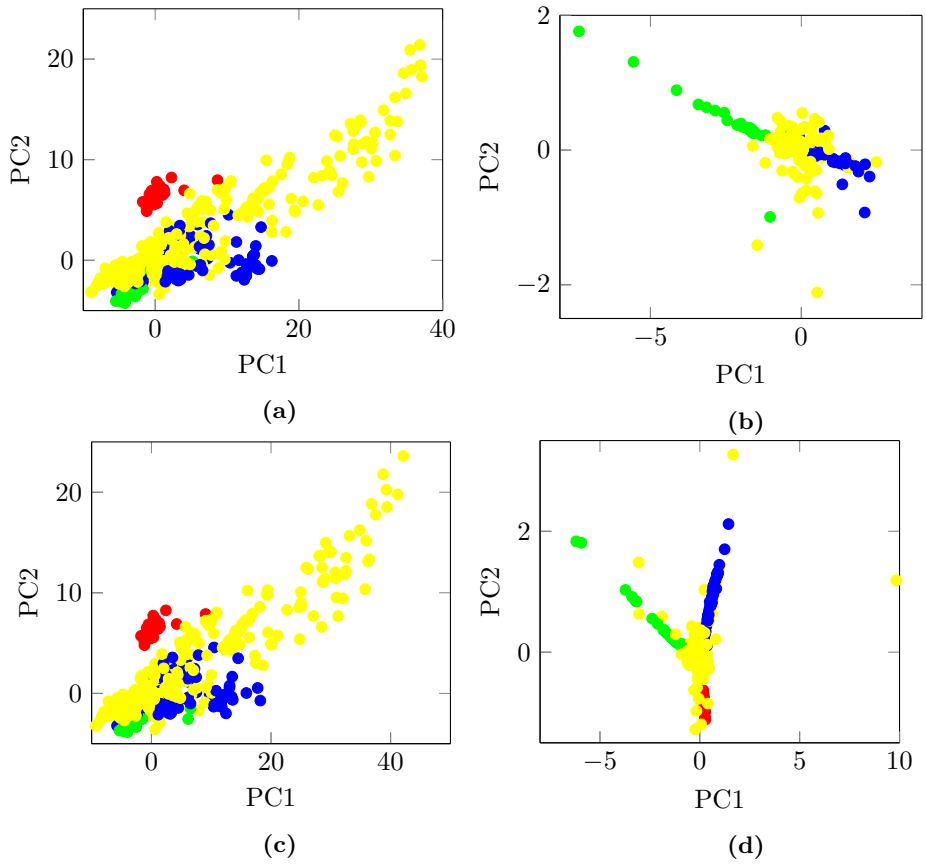


Figure 7.15: Score plot of image without barrier, showing PC1 vs. PC2 for Tartaric acid (blue), Lactose (green) and RDX (red) using (a) *A(f)* and $\delta t = 30$ ps (b) *dA(f)/df* and $\delta t = 30$ ps (c) *A(f)* and $\delta t = 15$ ps (d) *dA(f)/df* and $\delta t = 15$ ps.

that all scores belonging to the surroundings are clustered together in the middle of the materials' clusters. It is however hard to say, based on the score plots alone, which combination will yield the best results, because one also has to consider the correlation, which is done using the Mahalanobis distance.

The results without the scores of the surroundings are plotted in Appendix D, Figure D.1, making it easier to see the distribution of the material clusters. The results of using a paper and plastic barrier are essentially the same, and therefore not included. The results when introducing the cloth barrier, also plotted omitting the scores of the surroundings, can be found in Appendix D, Figure D.2. For *dA(f)/df*, when covering the samples with the cloth, it is observed that the points spread out more compared to the line-like distribution we get when no barrier, a paper barrier or a plastic barrier is used. A spread is also observed for the other

spectral characteristics and window widths, however, this is not as prominent.

7.5 ROC curves (PCA)

7.5.1 Mahalanobis distance

Different Mahalanobis distances (and hence, number of principal components/dimensions) were investigated for Image 2, shown in Figures 7.16 and 7.17, where we see the ROC curves for identification of Tartaric acid.

There are a few interesting things to notice. Theoretically, one would expect the performance to increase when increasing the number of dimensions of the Mahalanobis distance, as is the case when using the absorption spectrum and a window half width of 30 ps (Figure 7.16a). However, we see for the other cases (Figure 7.16b - 7.17b) that this is not necessarily the case. This could partly be explained by the curse of dimensionality (see Section 3.3.2). However, according to the curse of dimensionality, if the performance decreases after adding a dimension, it will not increase again if one keeps adding dimensions, which happens in the case of *A(f)* when using $\delta t = 15$ ps, shown in Figure 7.16b. Up to five dimensions, the performance increases with increasing dimensionality, but for seven dimensions, a decrease in the performance is observed. Then again, it increases when using ten dimensions. Recall that in PCA, we have made the *assumption* that the largest variances represent the most discriminative information, but this need not always be the case. Hence, a less principal component could in fact be more important, and when including it, an increase in performance can occur.

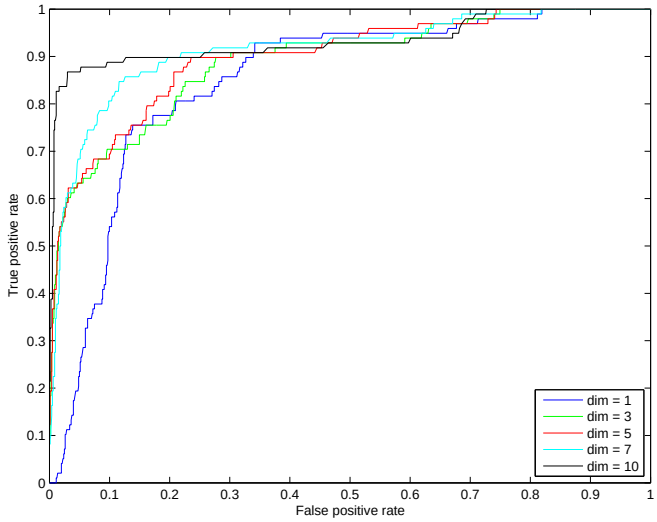
In Figure 7.17b, showing the scores when using *dA(f)/df* and $\delta t = 15$ ps, we see that the ROC curve using three dimensions of the Mahalanobis distance mostly lies below that of a random classifier. There is no obvious reason as to why this happens, but investigation of the Mahalanobis distances to the surroundings revealed that for some reason, when using three dimensions, the Mahalanobis distance to air and the sample holder in the image was in the same range as the distances for Sample 5 and 6 (thin and unground sample). This did not happen when reducing or increasing the dimensionality. It is likely that the scores for these two samples are the ones that overlap with the scores of the surroundings, see Figure 7.14c, and that in three dimensions, this overlap becomes significant. Another observation is that the best performance is obtained using one dimension, and that increasing the dimensionality only makes the classification less accurate, likely caused by the curse of dimensionality.

Further, the combination yielding the best performance over all, i.e. *A(f)* and $\delta t = 15$ ps, was investigated for Image 5, to see how the introduction of a barrier affects the optimal number of dimensions. The ROC curves are plotted in Figure 7.18. We see that in this case, three dimensions results in the best performance. This decrease in the optimal number of dimensions is related to the reduced density of the scores associated with introduction of the cloth barrier (see Figure D.1 and D.2).

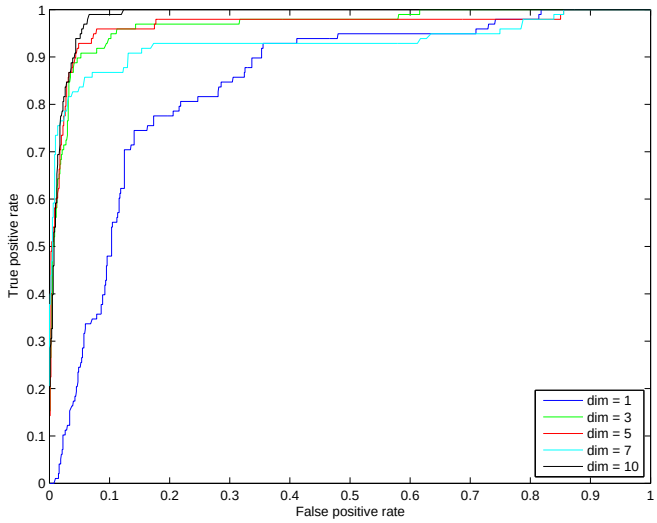
Different dimensions for the Mahalanobis distance were also investigated for Lactose and RDX. Some variation among the different combinations of spectral characteristic and window width was observed, but overall, three dimensions resulted in a good performance.

It is desirable to have the same dimensionality of the Mahalanobis distance for all three materials, since the amount of training data is approximately the same. Using a different dimensionality depending on the sample would be impractical in real world applications, where the sample one is investigating really is unknown, and one does not have the luxury of choosing how many dimensions one wants to include depending on the sample in question. 10 dimensions, which resulted in the best performance for Tartaric acid without a barrier, was investigated also for Lactose and RDX, but this was a poor choice, and was therefore not an option. Three dimensions, however was a good choice also for Tartaric acid. In addition, the optimal number of dimensions was three for Tartaric acid when the cloth barrier was used. Based on these considerations, three dimensions is used when looking at $A(f)$.

Although using three dimensions worked well (and better than one) for Lactose and RDX also when using $dA(f)/df$, it was a bad choice for Tartaric acid for $\delta t = 15$ ps. Therefore, one dimension of the Mahalanobis distance will be used when looking at $dA(f)/df$. However, we must also consider that the ROC curve for Tartaric acid resulted in a performance worse than that of a random classifier, for no obvious reason, which is already discussed. There is a possibility that this would not happen for other data, and therefore, in some cases for RDX and Lactose, the results of using three dimensions will be included in Appendix E.

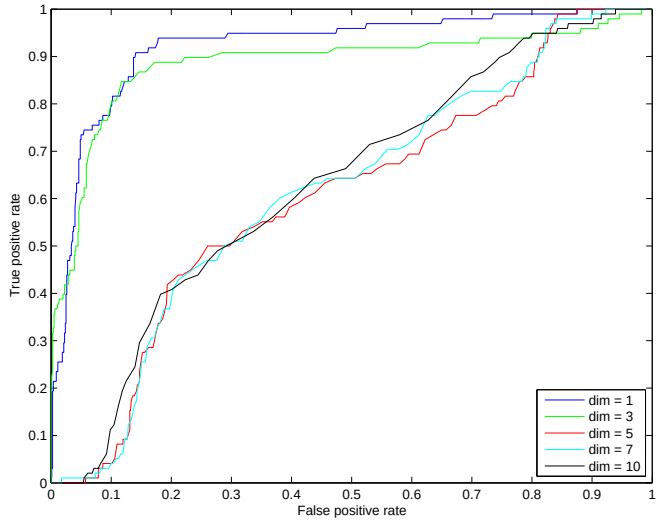


(a)

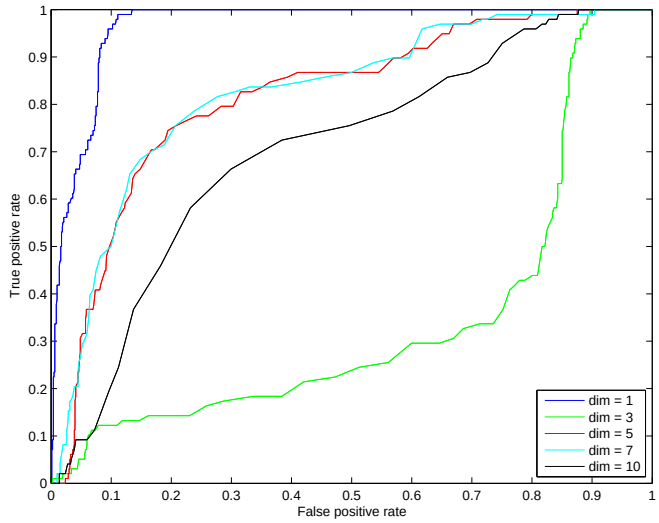


(b)

Figure 7.16: ROC curves for identification of Tartaric acid for different dimensions of the Mahalanobis distance using $A(f)$ and (a) $\delta t = 30$ ps (b) $\delta t = 15$ ps. No barrier.



(a)



(b)

Figure 7.17: ROC curves for identification of Tartaric acid for different dimensions of the Mahalanobis distance using *dA(f)/df* and (a) $\delta t = 30$ ps (b) $\delta t = 15$ ps. No barrier.

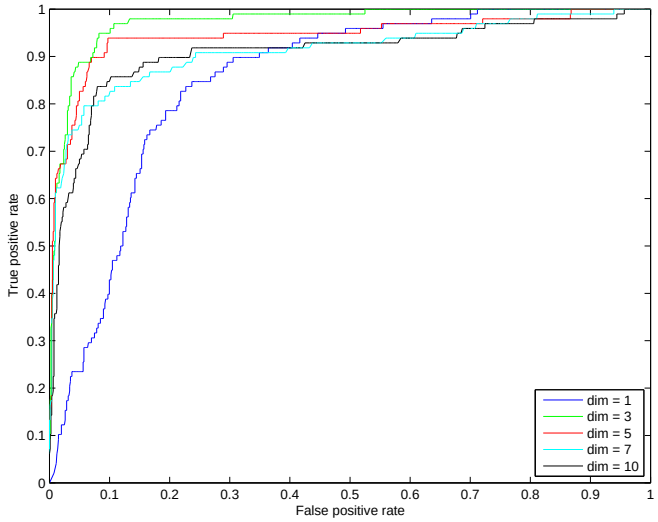


Figure 7.18: ROC curves for identification of Tartaric acid for different dimensions of the Mahalanobis distance using *A(f)* and $\delta t = 15$ ps. Cloth barrier.

7.5.2 No barrier (Image 2)

The ROC curves for Tartaric acid are plotted in Figure 7.19, showing both spectral characteristics and window widths for Image 2. Overall, a good performance is obtained, regardless of choice of spectral characteristic and window. We see that the best performance is obtained using the narrower window, both for the absorption spectrum and its derivative. As evident from Figures 7.15a and 7.15c, there are no major differences in the distribution of the scores for the two window widths when using *A(f)*. We see, however, that the scores of Tartaric acid are somewhat more collected when using $\delta t = 15$ ps. From comparison of the variance plots in Figures 7.13a and 7.13c, we also see that slightly more variance is accounted for using three principal components in the case of $\delta t = 15$ ps compared to $\delta t = 30$ ps. Both of these could explain why the performance is better using the narrower of the two windows.

Figure 7.20 shows the ROC curves for Lactose. We see that using *A(f)* is slightly better than using *dA(f)/df*, and as for Tartaric acid, that $\delta t = 15$ ps is better than $\delta t = 30$ ps. Lactose, contrary to Tartaric acid, would benefit from increasing the number of dimensions for the Mahalanobis distance to three when using *dA(f)/df*, and these results are shown in Appendix E, Figure E.1. We observe that in this case, using $\delta t = 15$ ps, the performance is almost as good as when using *A(f)* and the same window, and better than for *A(f)* and $\delta t = 30$ ps.

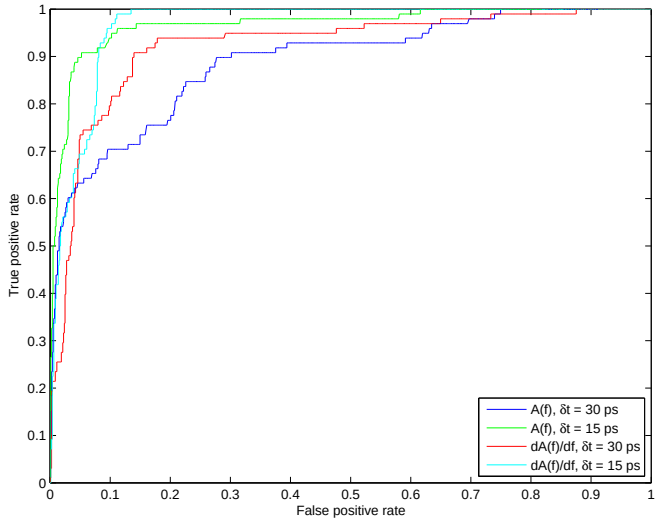


Figure 7.19: ROC curves for identification of Tartaric acid using PCA, for the two spectral characteristics and window widths. No barrier.

The ROC curves for RDX are plotted in Figure 7.21. The best performance is obtained using the absorption spectrum, which yields similar results for both window widths. However, the results using the derivative are surprisingly good when considering that only one dimension is used, especially using $\delta t = 30$ ps, because as was evident from the score plot, projection on to a single dimension did not separate the scores of RDX from those of the surroundings (Figure 7.14b). However, the scores are plotted in Euclidean space, while the classification also takes into account the correlation. Although it looks like RDX is indistinguishable from the surroundings, it is evident from the ROC curve that the scores of RDX can in fact be distinguished from the rest.

The results of using three dimensions for the derivative are shown in Appendix E, Figure E.2, and we see that RDX benefits from increasing the dimensionality, especially when the wider window is used, which results in a nearly ideal ROC curve, the performance being similar to when the absorption spectrum is used. Because of the distribution of the scores, being clustered together very tightly, almost in a point-like manner, increasing the dimensionality leads to better separability, without having the concern of a big decrease in the density of the samples.

The ROC curves using a paper and plastic barrier are given in Appendix E, Figures E.3 - E.5 and Figures E.6 - E.8, for all three materials. For Lactose and RDX, the results are again plotted both using one dimension and three dimensions

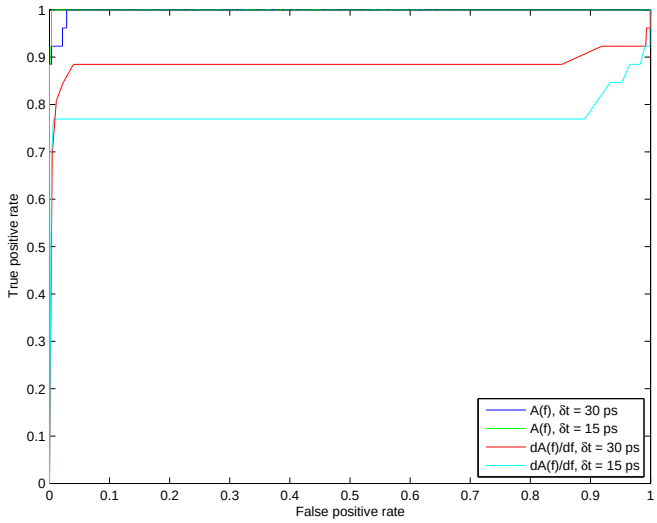


Figure 7.20: ROC curves for identification of Lactose using PCA, for the two spectral characteristics and window widths. No barrier.

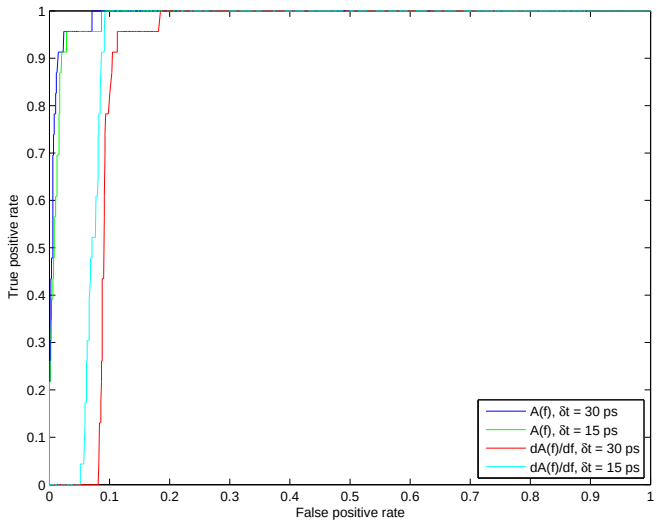


Figure 7.21: ROC curves for identification of Lactose using PCA, for the two spectral characteristics and window widths. No barrier.

of the Mahalanobis distance for *dA(f)/df*. Essentially the same results are obtained with both barriers as without, again observing that the results overall are good. As was the case for SAM, for some combinations of spectral characteristic and window width, a slight increase in the performance is observed compared to when no barrier is used, possibly due to reasons already discussed for SAM. Considering that this happens for both algorithms, it could also be caused by differences in the environment at the time that the measurements were performed, e.g. temperature and relative humidity.

7.5.3 Cloth barrier (Image 5)

Figure 7.22 shows the ROC curves for Tartaric acid when the cloth barrier is introduced. There are no major differences compared to when no barrier is used (Figure 7.19), and we see that all four combinations of spectral characteristic and window width yield good results. The best performance is obtained using *A(f)* in combination with $\delta t = 15$ ps.

The results for Lactose are plotted in Figure 7.23. We see an increase in the performance compared to using no barrier for *dA(f)/df*, possibly due to reasons discussed already, while a slight decrease is observed when using *A(f)*. Overall, the performance is very good, and the best performance is obtained using *A(f)* and $\delta t = 15$ ps, which was also the case for Tartaric acid.

The ROC curves for RDX are shown in 7.24. We see a decrease in the performance for all four combinations of spectral characteristic and window width compared to using no barrier. When using the absorption spectrum, only a slight decrease is observed, and the best performance is obtained using the wider window, but the advantage over the narrower one is small. For *dA(f)/df*, the decrease in performance is more obvious, and using the wider window results in a relatively bad performance of the classifier. As discussed in Section 7.4.2, introducing the cloth barrier resulted in a spread of the scores compared to when no barrier was used, which increases the Mahalanobis distance (Eq. 3.16) and makes the classification more challenging.

The ROC curves when increasing the number of dimensions to three (for *dA(f)/df*) are given in Appendix E, Figure E.10, showing that this is beneficial when using $\delta t = 30$ ps, making the performance comparable to when using the absorption spectrum, as was the case for the other images. However, it leads to a decrease in performance when using $\delta t = 15$ ps, possibly due to the curse of dimensionality, only becoming an issue using the cloth barrier because it leads to a decrease in the density of the scores, which was not the case for the paper and plastic barrier. As discussed, a spreading of the scores is also evident using the wider window, but being that there is a big difference between the two cases, especially for RDX, it is not surprising that the optimal number of dimensions differs.

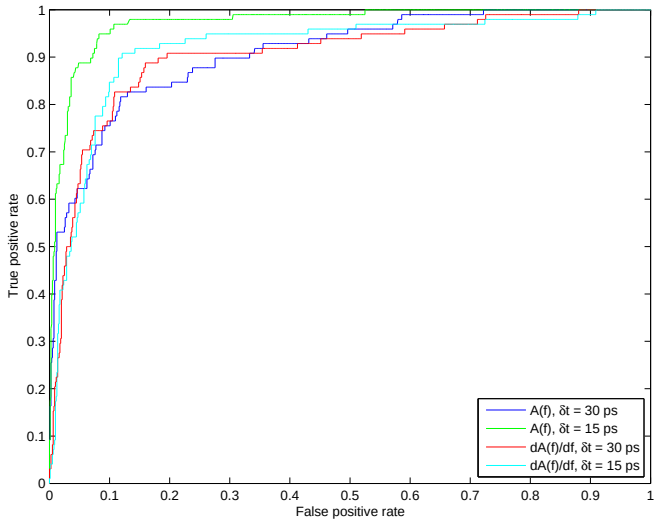


Figure 7.22: ROC curves for identification of Tartaric acid using PCA, for the two spectral characteristics and window widths. Cloth barrier.

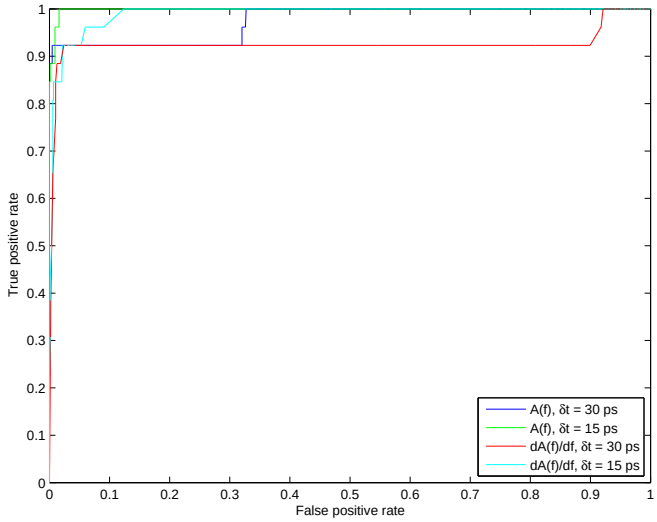


Figure 7.23: ROC curves for identification of Lactose using PCA, for the two spectral characteristics and window widths. Cloth barrier.

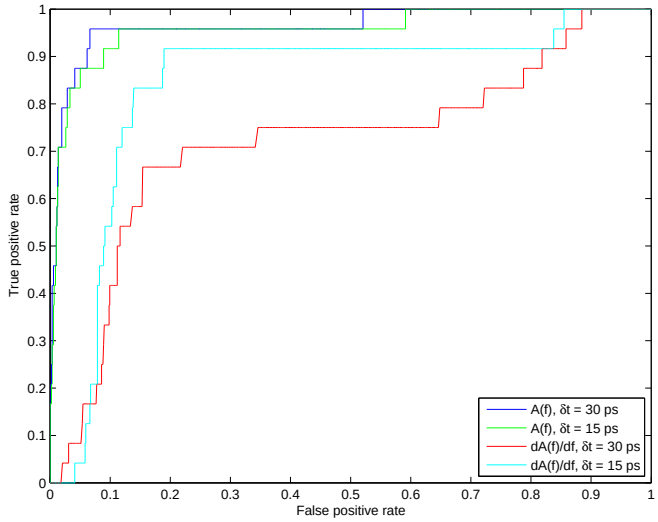


Figure 7.24: ROC curves for identification of RDC using PCA, for the two spectral characteristics and window widths. Cloth barrier.

7.6 Comparison of PCA and SAM

Some ROC curves showing the performance of PCA (using three dimensions for *A(f)* and one for *dA(f)/df*) and SAM in the same plot are presented here. The previous two sections showed that the performance varies somewhat for the different combinations of spectral characteristic and window width, but only one window width, $\delta t = 15$ ps is considered, which for both algorithms resulted in a better overall performance.

Figures 7.25 - 7.27 show the comparison for Tartaric acid, Lactose and RDX, respectively, in air and for the cloth barrier. For the paper and plastic barrier, we observed in the previous sections that the results were similar to using no barrier.

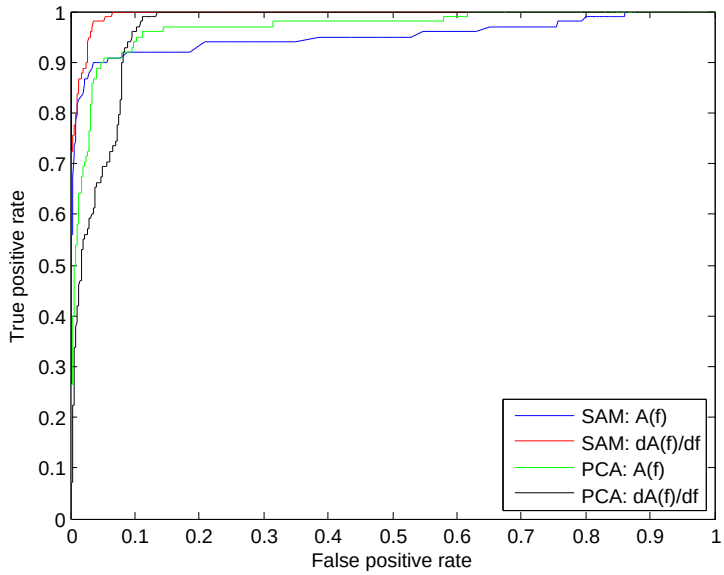
For the most general case, Tartaric acid, we observe that SAM is better than PCA when no barrier is used (Figure 7.25a), if using the derivative of the absorption spectrum (red curve). For PCA, the best result is obtained using the absorption spectrum. However, this is only true if a very small false positive rate is acceptable. For a false positive rate of approximately 0.15, all Tartaric acid pixels are identified using both algorithms, and the derivative becomes the better choice for PCA.

With the cloth barrier (Figure 7.25b), we observe that the best results for both PCA and SAM are obtained using the absorption spectrum, and that the perfor-

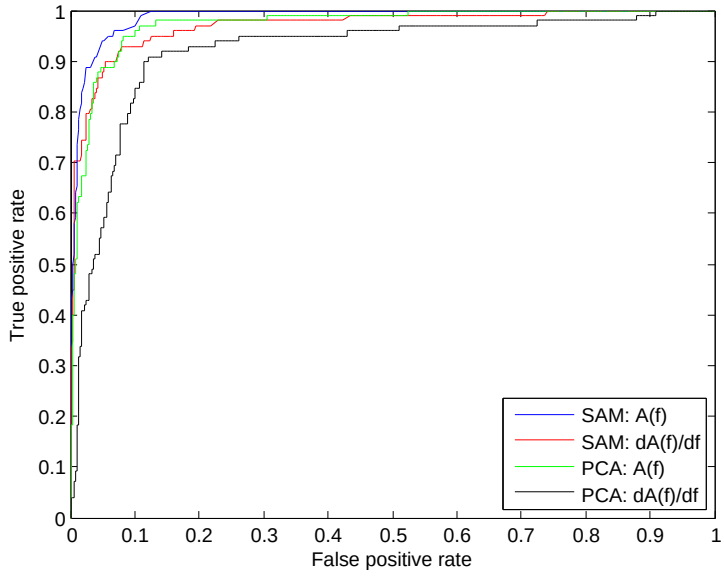
mance is slightly better for SAM. If, however, using the derivative for SAM, which was better without a barrier, PCA and SAM yield essentially the same results, given that for PCA, the absorption spectrum is used.

For Lactose, using no barrier (Figure 7.26a), we see that both PCA and SAM result in nearly ideal ROC curves, the exception being *dA(f)/df* for PCA. The cloth barrier hardly affects the performance of either algorithm when the absorption spectrum is used, seen in Figure 7.27b. Regarding the use of *dA(f)/df* in this case, the performance is also very good and similar for both algorithms. SAM is slightly better if accepting a very small false positive rate (below approximately 0.05), before the results are essentially same. At a false positive rate just above 0.1, PCA becomes the better choice.

Using SAM, the ROC curves for RDX are nearly ideal regardless of spectral characteristic, window and barrier, for reasons discussed in Section 7.3.1. For PCA, the results are also very promising both with and without a barrier. The absorption spectrum is better in either case, and the performance is comparable to that of SAM when the absorption spectrum is used.

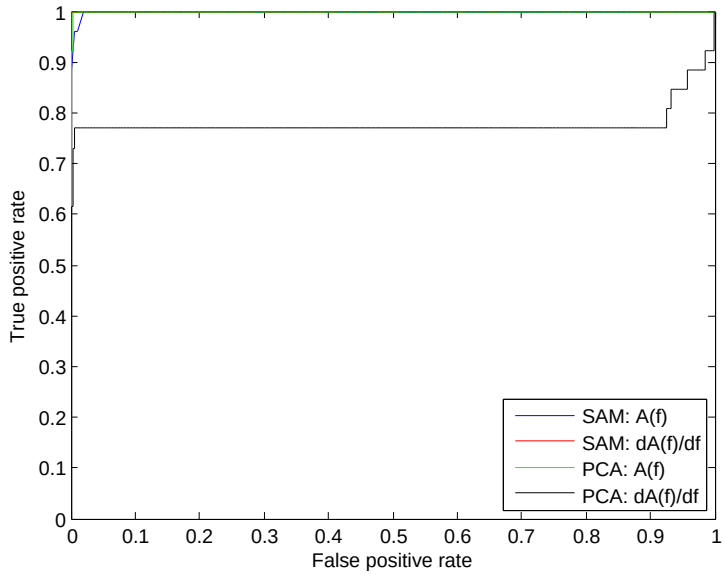


(a)

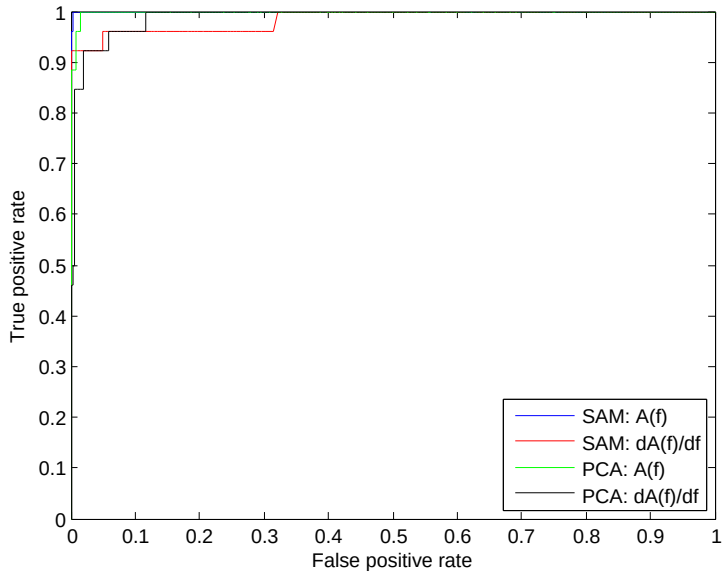


(b)

Figure 7.25: ROC curves for Tartaric acid comparing the use of PCA and SAM, for $\delta t = 15$ ps and both spectral characteristics. (a) No barrier (b) Cloth barrier

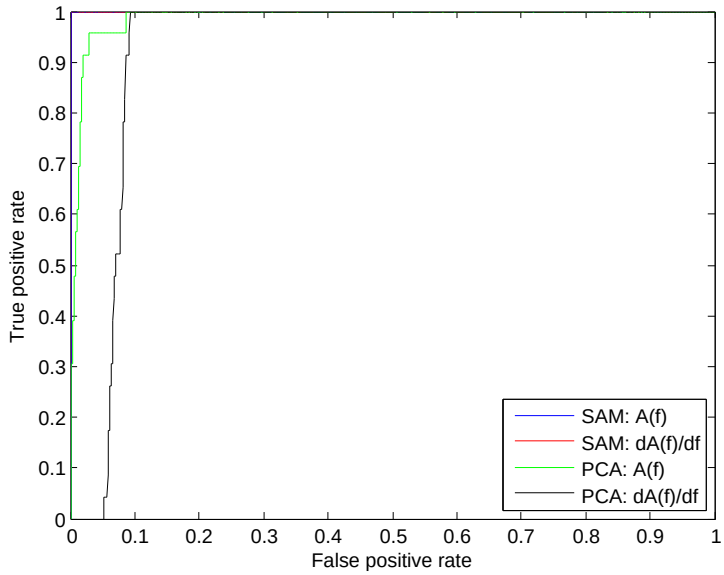


(a)

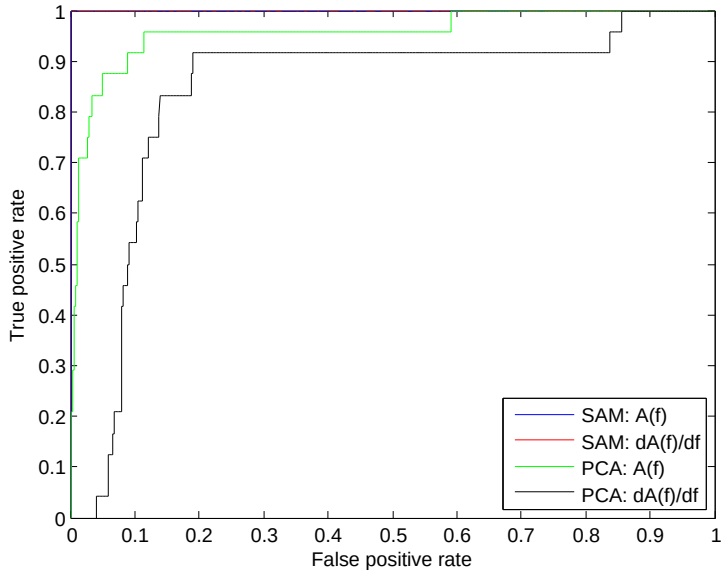


(b)

Figure 7.26: ROC curves for Lactose comparing the use of PCA and SAM, for $\delta t = 15ps$ and both spectral characteristics. (a) No barrier (b) Cloth barrier.



(a)



(b)

Figure 7.27: ROC curves for RDX comparing the use of PCA and SAM, for $\delta t = 15$ ps and both spectral characteristics. (a) No barrier (b) Cloth barrier.

8 Conclusion

We have seen that using THz-TDS in transmission, the spectral fingerprints of samples containing Tartaric acid, Lactose and RDX are obtained. Four THz images for testing the performance of PCA and SAM have been acquired, and the performance evaluated and compared using ROC curves. Both the absorption spectrum and its derivative have been investigated, using two different window widths, $\delta t = 30$ ps and $\delta t = 15$ ps.

Dimensionality

Before evaluating PCA, how the number of principal components affects the results of the classification was investigated, and it was observed that increasing the dimensionality can lead to a degrade of the performance. We found that the optimal number of dimensions depends on the sparsity of the material's score distribution, which in turn depends on the spectral characteristic and barrier used.

In general, the tendency is that when using the derivative of the absorption spectrum, the optimal number of dimensions is lower than for the absorption spectrum, because the density of each material's cluster is higher. We also saw that when introducing a cloth barrier, the optimal number of dimensions decreased in most cases, because of the increased sparsity of the score distributions associated with the barrier. Whether this happens, however, depends on the nature of the distribution before and after introducing the barrier. It was found that for a point-like distribution of the training data (i.e. RDX when using the derivative and $\delta t = 30$ ps), even though introducing the cloth barrier indeed lead to more sparsity, increasing the number of dimensions (to three) was beneficial. In conclusion, increased sparsity is not necessarily a bad thing, if the added dimension contributes to better separability.

Window width and spectral characteristic

It has been established that a window width of $\delta t = 30$ ps smooths the absorption spectra sufficiently. Additional smoothing using $\delta t = 15$ ps does not negatively affect the performance, but little is gained in the case of SAM. However, the nar-

rower window proved beneficial for the derivative, because of additional reduction of the effects of absorption by water vapour in the air.

Identification using PCA was generally most successful for the absorption spectrum and the narrower window. For this combination, the variance accounted for by the three first principal components was largest. The distribution of the scores of each material was also somewhat denser than for the absorption spectrum and the wider window.

Comparison of PCA and SAM

All three materials are clearly identified using both algorithms, even when covered by plastic, paper or cloth. Classification with a true positive rate above 0.9, while still retaining a false positive rate below 0.2, is possible for both algorithms, regardless of barrier, given the right choice of spectral characteristic and window width, which in the case of PCA is the absorption spectrum combined with the narrower window. For SAM, the results are inconclusive. What was clearly indicated, however, was that the narrower window was a better choice when using the derivative.

The results are very promising, but they have to be seen in context with the amount and type of test data used. The number of positives for Lactose and RDX is low (26 and 23, respectively), and minor differences in the ROC curves of PCA and SAM can therefore be coincidental. In addition, the samples are identical to the training/reference samples, and drawing conclusions for more general cases is not possible. The results for Tartaric acid are more general, because several Tartaric acid samples with varying sample properties are considered, with the number of positives hence being about four times bigger than for RDX and Lactose. The ROC curves for identification of Tartaric acid revealed that both the absorption spectrum and its derivative are good choices for SAM, while PCA works better when the absorption spectrum is used. The results are similar, SAM being slightly better when the derivative is used for the case of no barrier, in which case PCA is slightly better for the cloth barrier. If instead using the absorption spectrum for SAM, it is slightly better than PCA in the case of the cloth barrier, while for no barrier, PCA is slightly better.

SAM has the advantage of needing only one reference measurement for each reference sample. However, for cases where the spectral characteristics of the pixels in some unknown THz image vary more compared to the reference measurement, e.g. due to more challenging barriers, this could be a disadvantage. PCA requires several measurements of the same training sample for the training data, but this also means that it could prove more robust to bigger changes in the measurement conditions, because the amount of training data can be increased.

In stead of reducing the performance, introduction of the barriers in some cases

leads to a slightly better performance for both SAM and PCA. This does not mean that a barrier makes identification easier in general. It merely suggests that the barriers used were not very challenging, and because of the low resolution of the THz images, this increase is probably coincidental. However, the fact that the performance was not significantly degraded when introducing the barriers, showed that both PCA and SAM are robust to changes in the measurement conditions for the barriers used. Both algorithms also proved robust to varying sample properties.

Future work

Increasing the performance of both SAM and PCA could be achieved by removal of water absorption lines using signal processing. In addition, for PCA, optimizing the training data, by increasing the amount and removing outliers, could increase the performance. In increasing the amount of training data, more principal components can be used before the curse of dimensionality strikes, which could be beneficial because more of the variance in the data would be accounted for. The number of dimensions required for separating a larger number of classes (materials) would also be interesting to investigate regarding PCA, and whether or not using the derivative in this context would prove beneficial.

Bibliography

[1] M. W. Haakestad and A. D. Van Rheenen, "Thz time-domain detection of explosive simulants using fiber-coupled emitter and detector antennas," *RTO*, 2011.

[2] ITU, "Nomenclature of the frequency and wavelength bands used in telecommunications." `https://www.itu.int/dms_pubrec/itu-r/rec/v/R-REC-V.431-7-200005-I!!PDF-E.pdf`. Accessed: 2014-10-09.

[3] Y.-S. Lee, *Principles of Terahertz Science and Technology*. Springer Science and Business Media, 2009.

[4] E. R. Mueller, "Terahertz radiation: Applications and sources," *The Industrial Physicist*, 2003.

[5] G. Kniffin, "Metamaterial devices for the terahertz band," pp. 1–9, 2009.

[6] S. L. Dexheimer, *Terahertz Spectroscopy: Principles and Applications*. Taylor & Francis Group, 2008.

[7] "The terahertz laboratories." `http://www.thz.soton.ac.uk/sample-page-2/`. Accessed: 2014-10-09.

[8] X.-C. Zhang and D. Brigada, "Chemical identification with information-weighted terahertz sepctrometry," *IEEE*, vol. 2, no. 1, pp. 107–112, 2012.

[9] H. Zhong, A. Redo-Sanchez, and X.-C. Zhang, "Identification and classification of chemicals using terahertz reflective spectroscopic focal-plane imaging system," *Opt. Express*, vol. 14, no. 20, pp. 9130–9141, 2006.

[10] B. E. A. Saleh and M. C. Teich, *Fundamentals of Photonics*. John Wiley & Sons, 2007.

[11] W. H. Press, W. T. Teukolsky, Saul A. andVetterling, B. P. Flannery, and M. Metcalf, *Numerical Recipes in FORTRAN 77: The art of Scientific Computing*. Cambridge University Press, 1992.

[12] LDS-group, "Understanding fft windows," 2008.

[13] K. M. M. Prabhu, "Window finctions and their applications in signal processing," 2014.

[14] M. Richardson, "Principal component analysis," pp. 1–19, 2009.

[15] L. I. Smith, "A tutorial on principal components analysis," pp. 2–22, 2002.

[16] J. Shlens, "A tutorial on principal component analysis," pp. 1–13, 2005.

[17] H. Abdi and L. J. Williams, "Principal component analysis," *John Wiley and Sons, inc. WIREs Comp stat*, vol. 2, no. 4, pp. 433–456, 2010.

[18] T. Krastev, "Brief theory of classifion." `http://classifion.sicyon.com/Theory/index.htm`. Accessed: 2014-11-09.

[19] V. Spruyt, "The curse of dimensionality in classification," 2014.

[20] M. W. Haakestad and A. D. Van Rheenen, "Terahertz imaging spectroscopy - towards robust identification of concealed dangerous substances," *FFI*, 2014.

[21] T. C. O'Haver, "A pragmatic introduction to signal processing with applications in scientific measurement," *University of Maryland at College Park*, 2015. Available from: http://terpconnect.umd.edu/ toh/spectrum/IntroToSignalProcessing.pdf.

[22] A. J. Owen, "Uses of derivative spectroscopy: Application note," *Aligent Technologies*, pp. 1–8. Available from: `http://www.chem.agilent.com/Library/applications/59633940.pdf`.

[23] J. B. Niklas and W. C. Low, "Roc-supervised principal component analysis in connection with the diagnosis of diseases," *Am. J. Transl. Res.*, vol. 3, no. 2, pp. 180–196, 2011.

[24] P. A. Flach and S. Wu, "Repairing concavities in roc curves," *In: Proc. 2003 UK Workshop on Computational Intelligence. University of Bristol*, 2003.

Appendix A: MATLAB code

A.1 Training data and reference spectra

Training.m

```
%%%% Training data for PCA, reference data for SAM %%%%

%% Training samples %%
% The training samples are from a higher resolution THz-image with no
    %
% barrier
    %

% Measurement parameters
xstart = 0;
xstop = 150;
nx = 64;

ystart = 0;
ystop = 150;
ny = 64;

xpos = round(linspace(xstart, xstop, nx));
ypos = round(linspace(ystart, ystop, ny));

% Windowing parameters
v = 0;
windowing = true;           %set to true if window is appplied
hw = 30;                    %change window half width

if windowing == false
    hw = 0;
end

% Frequency range

f1 = 0.3;
f2 = 1.51;              % = 1.51 (not 1.50) for practical purposes
    making
                        % actual range 0.3-1.5

% Other parameters %
```

```
mmode = 'diff';          % set to 'none' if looking at absorbance-spectra.
                         % set to 'diff' if looking at the differentiation of
                         % absorbance spectra
dim = 10;                % number of dimensions included in PCA
numOfSamples = 3;        % number of samples for SAM

%% Reference (air) %%
fn = ['THz_x_0_mm_y_0_mm__V_u.txt'];
[x_a, t_a] = readThz(fn);
[~, i_a] = max(x_a);
if windowing == true
w_a = blha(t_a-t_a(i_a), hw, hw);          %Blackman-Harris window
x_a = x_a.*w_a;
end

[f_a, s_a] = ampl2spec2([t_a, x_a], v);    %Finds spectrum of signal
[i1, i2] = closest_value2(f1, f2, f_a);
s_a = s_a(i1:i2);
f_a = f_a(i1:i2);

%% Training data, PCA %%
c1 = 0;
c2 = 0;
c3 = 0;

T1 = zeros(1,length(s_a)-1);
L1 = zeros(1, length(s_a)-1);
R1 = zeros(1, length(s_a)-1);
m1a = zeros(nx, ny);
for k = 1:nx
    for l = 1:ny
        fn = ['THz_x_' num2str(xpos(k)) '_mm_y_' num2str(ypos(l)) '_mm__V_u' '.txt'];
        if exist(fn, 'file')
            temp = sprintf('%s %s %s', 'Reading ', fn, '\n');
            fprintf(temp)
            [x, t] = readThz(fn);
            [~, i_p] = max(x);

            m1a(l, k) = energy_interval([t, x], 0.3, 1.5); %THz image
            delay = round(t(i_p) - t_a(i_a));

            if delay < 8 && delay > 0
                if windowing == true
                w = blha(t - t(i_p), hw, hw);
                x = x.*w;
                end
                [f, s] = ampl2spec2([t, x], v);
                [i1, i2] = closest_value2(f1, f2, f);
                f = f(i1:i2);
                s = s(i1:i2);
                trans = abs(s./s_a);
                AS = -log(trans(1:length(trans)-1));  %A(f)
                dAS =-diff(trans)./trans(1:length(trans)-1); %dA(f)/df

```

```
89                  %% Grouping of samples %%
90                  if mmode == 'none'
91                      [T1, L1, R1, c1, c2, c3] = findSample(xpos(k),
                            ypos(l), AS, c1, c2, c3, T1, L1, R1);
92                  elseif mmode == 'diff'
93                      [T1, L1, R1, c1, c2, c3] = findSample(xpos(k),
                            ypos(l), dAS, c1, c2, c3, T1, L1, R1);
94                  end
95              end
96          else
97              fprintf('Error: file does not exist \n')
98          end
99
100     end
101 end
102
103 mat = [T1;L1;R1];
104
105 % Returns mean-adjusted matrix, eigenvectors with highest eigenvalues
106 % (number given by dim) and variance
107 [matAdjust, e, Var_pc] = PCanalysis(mat, dim);
108
109 %% Scores %%
110 S = e'*matAdjust';
111 S1 = S(1:3, 1:c1);
112 S2 = S(1:3, c1+1: c1+c2);
113 S3 = S(1:3, c1+c2+1:c1+c2+c3);
114
115 %% SAM %%
116
117 load('SAM_reference') %Loads reference spectra for SAM
118
119  for i = 1:numOfSamples
120         [px, py] = closest_value(xd(i), yd(i), xpos, ypos);
121         fn = ['THz_x_' num2str(px) '_mm_y_' num2str(py) '_mm__V_u' '.
                txt'];
122
123         if exist(fn, 'file')
124             temp = sprintf('%s %s %s', 'Reading ', fn, '\n');
125             fprintf(temp)
126             [ x, t ] = readThz( fn );
127             [~, i_p] = max(x);
128             if windowing == true
129             w = blha(t-t(i_p), hw, hw);
130             x = x.*w;
131             end
132             [f, s] = ampl2spec2([t, x],v);   %spectrum
133             [i1, i2] = closest_value2(f1, f2, f);
134             s = s(i1:i2);
135             f = f(i1:i2);
136             f_sam = f;
137             trans = abs(s)./abs(s_a);
138             Trans(i,:) = trans(1:length(trans)-1);
139             AS = -log(trans(1:length(trans)-1));
140             dAS = -diff(trans)./trans(1:length(trans)-1);
141             if mmode == 'none'
142                 Ref_sam(i,:) = AS;
```

```
            elseif mmode == 'diff'
                Ref_sam(i,:) = dAS;
            end

        else
            fprintf('Error: file does not exist \n')
        end
end
```

A.2 Analysis

Analysis.m

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%%%% Analysis of unknown image, PCA and SAM %%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Measurement parameters
xstart = 0;
xstop = 150;
nx = 30;

ystart = 0;
ystop = 150;
ny = 30;

xpos = round(linspace(xstart, xstop, nx));
ypos = round(linspace(ystart, ystop, ny));

numOfPixels = nx*ny;

%% Unknown THz image %%
b = 0;
mat2 = zeros(numOfPixels, length(f_a)-1);
delay_mat = zeros(nx, ny);
delmat = zeros(nx, ny);
Label1 = zeros(numOfPixels, 1);
Label2 = zeros(numOfPixels, 1);

%  Reference spectra %
K1 = Ref_sam(1,:);
K2 = Ref_sam(2,:);
K3 = Ref_sam(3,:);

%% Normalization of reference data and data %%
K1 = K1/sqrt(sum(K1.^2));
K2 = K2/sqrt(sum(K2.^2));
K3 = K3/sqrt(sum(K3.^2));

%% SAM %%
corrmat1 = zeros(ny,nx);
corrmat2 = zeros(ny,nx);
corrmat3 = zeros(ny,nx);
```

```
41
42 for k = 1:nx
43     for l = 1:ny
44         b = b+1;
45         fn = ['THz_x_' num2str(xpos(k)) '_mm_y_' num2str(ypos(l)) '
               _mm__V_u' '.txt'];
46         if exist(fn, 'file')
47             temp = sprintf('%s %s %s', 'Reading ', fn, '\n');
48             fprintf(temp)
49             [x, t] = readThz(fn);
50             [~, i_p] = max(x);
51             delay = round(t(i_p) - t_a(i_a));
52
53             w = blha(t - t(i_p), hw, hw);
54             x = x.*w;
55
56             delmat((ny+1-l),k) = delay;
57
58             [f, s] = ampl2spec2([t, x], v);
59             [i1, i2] = closest_value2(f1, f2, f);
60             f = f(i1:i2);
61             s = s(i1:i2);
62             trans = abs(s./s_a);
63             AS = -log(trans(1:length(trans)-1));
64             dAS =-diff(trans)./trans(1:length(trans)-1);
65
66             %% Classification SAM: correlation between each pixel and
                   ref.%
67             if mmode == 'none'
68                 mat2(b,:) = AS;
69                 corrmat1(l,k) = correlation(K1, AS'/sqrt(sum(AS.^2)));
70                 corrmat2(l,k) = correlation(K2, AS'/sqrt(sum(AS.^2)));
71                 corrmat3(l,k) = correlation(K3, AS'/sqrt(sum(AS.^2)));
72
73             elseif mmode == 'diff'
74                 mat2(b,:) = dAS;
75                 corrmat1(l,k) = correlation(K1, dAS'/sqrt(sum(dAS.^2))
                       );
76                 corrmat2(l,k) = correlation(K2, dAS'/sqrt(sum(dAS.^2))
                       );
77                 corrmat3(l,k) = correlation(K3, dAS'/sqrt(sum(dAS.^2))
                       );
78             end
79             %% Labels for ROC %%
80             if delay < 8 && delay > 0
81                 [LAB1, LAB2] = findTrue( xpos, ypos, k, l);
82                 delay_mat((ny+1-l),k) = LAB1;
83                 Label1(b) = LAB1;
84                 Label2(b) = LAB2;
85             else
86                 delay_mat((ny+1-l),k) = 0;
87                 Label1(b) = 0;
88                 Label2(b) = 0;
89             end
90         else
91             fprintf('Error: file does not exist \n')
92         end
```

```
    end
end

%% Transformation: PCA %%
mu2 = mean(mat2);
matAdjust2 = bsxfun(@minus, mat2, mu2);
finalData2 = e'*matAdjust2';

%% Mahalanobis distance: Classification PCA %%

X1 = S(1:mahal_dim,1:c1)';
X2 = S(1:mahal_dim, c1+1:c1+c2)';
X3 = S(1:mahal_dim, c1+c2+1:c1+c2+c3)';
Y = finalData2';

D1 = mahal(Y(:,1:mahal_dim),X1);
D2 = mahal(Y(:,1:mahal_dim), X2);
D3 = mahal(Y(:,1:mahal_dim), X3);
D = [D1, D2, D3];
```

A.3 Functions

```
function [ matAdjust, e, Var_pc ] = PCanalysis(data, dim)

mu = mean(data);
matAdjust = bsxfun(@minus, data, mu);
covMat = cov(matAdjust);
[eigVec, eigValMat] = eig(covMat);
eigVal = diag(eigValMat);
[eigValSort, i_sort] = sort(abs(eigVal),1,'descend');
[ e ] = biggestValue(eigVec, dim, i_sort);
Var_pc = bsxfun(@rdivide, eigValSort, sum(eigVal));

end
```

Appendix B: Spectral correlation images

B.1 No barrier (Image 2)

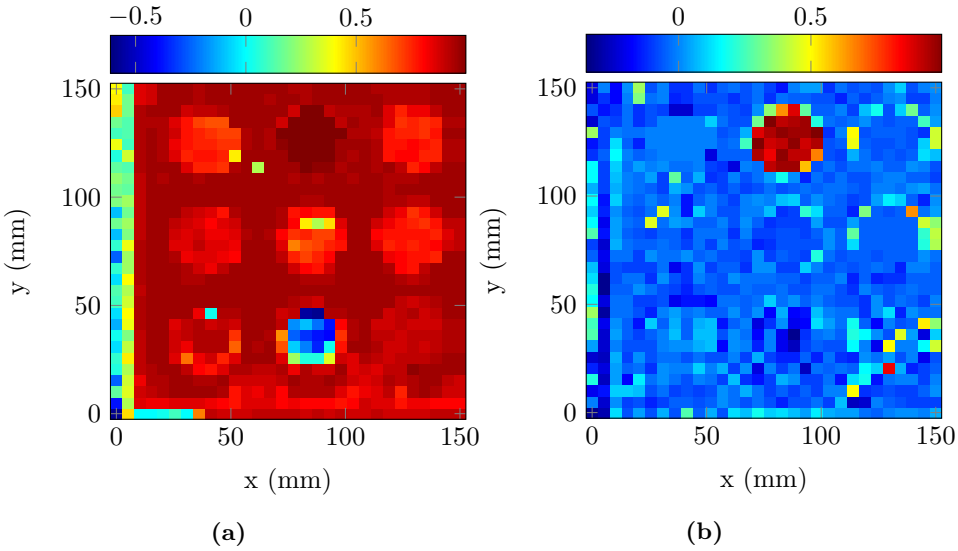


Figure B.1: Spectral correlation of Lactose using $\delta t = 15$ ps and (a) *A(f)* (b) *dA(f)/df*.

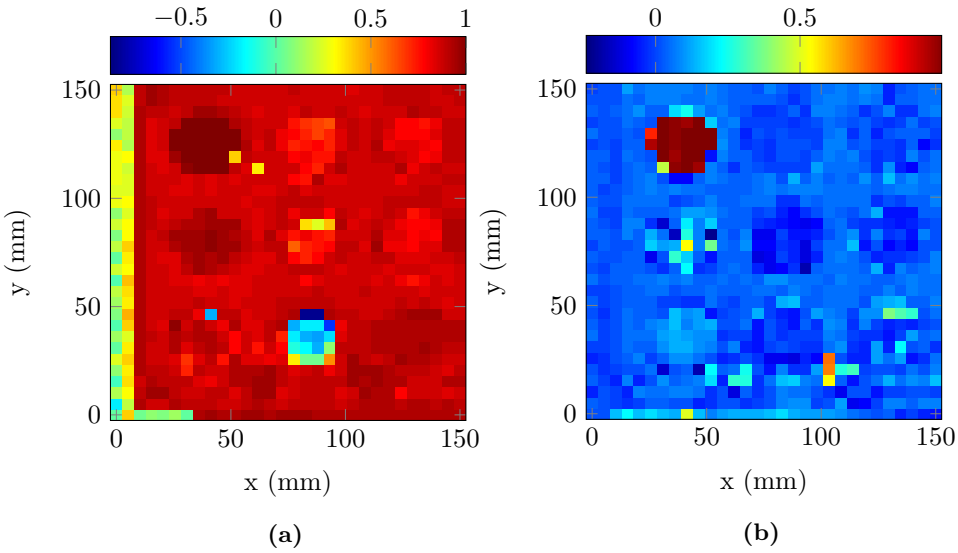


Figure B.2: Spectral correlation of RDX using $\delta t = 15$ ps and (a) *A(f)* (b) *dA(f)/df*

B.2 Paper barrier (Image 3)

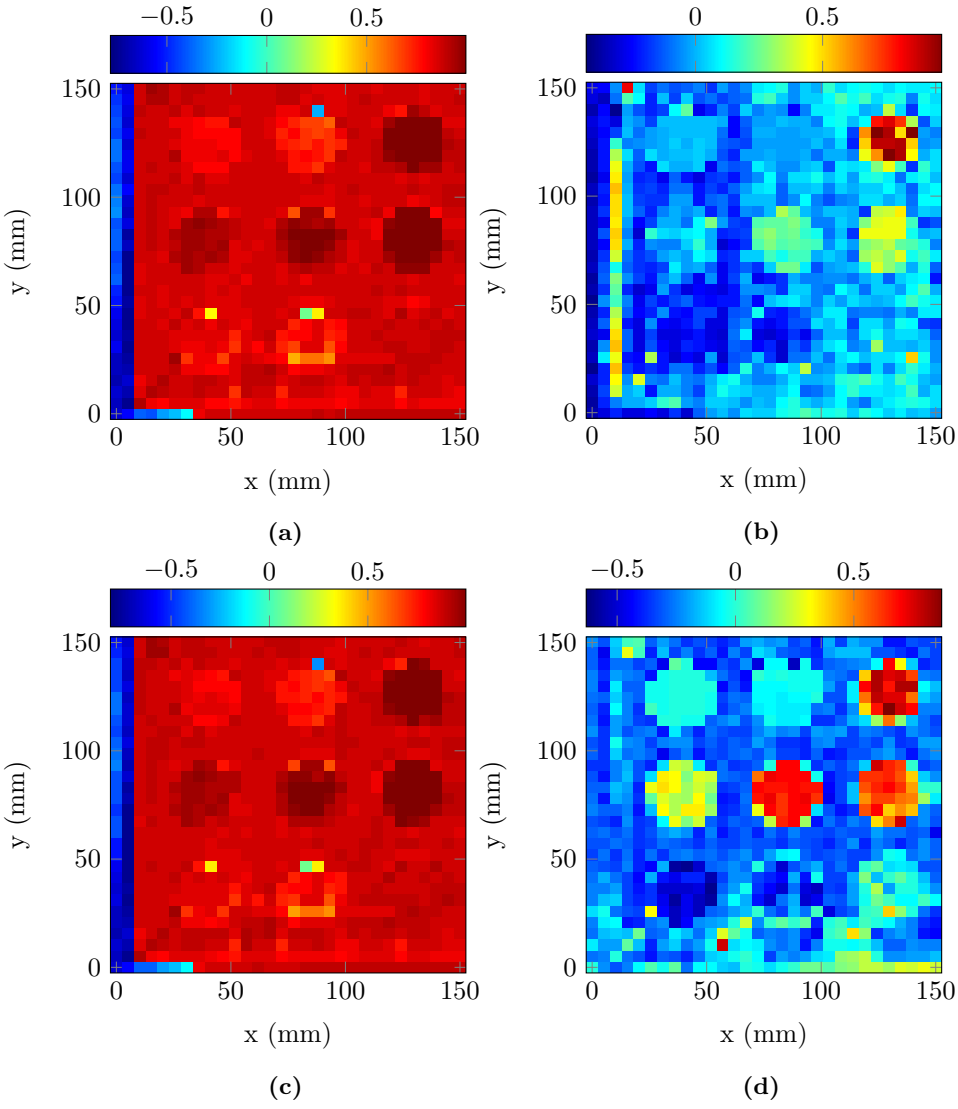


Figure B.3: Spectral correlation of Tartaric acid with paper barrier, using (a) *A(f)* and $\delta t = 30$ ps (b) *dA(f)/df* and $\delta t = 30$ ps (c) *A(f)* and $\delta t = 15$ ps (d) *dA(f)/df* and $\delta t = 15$ ps

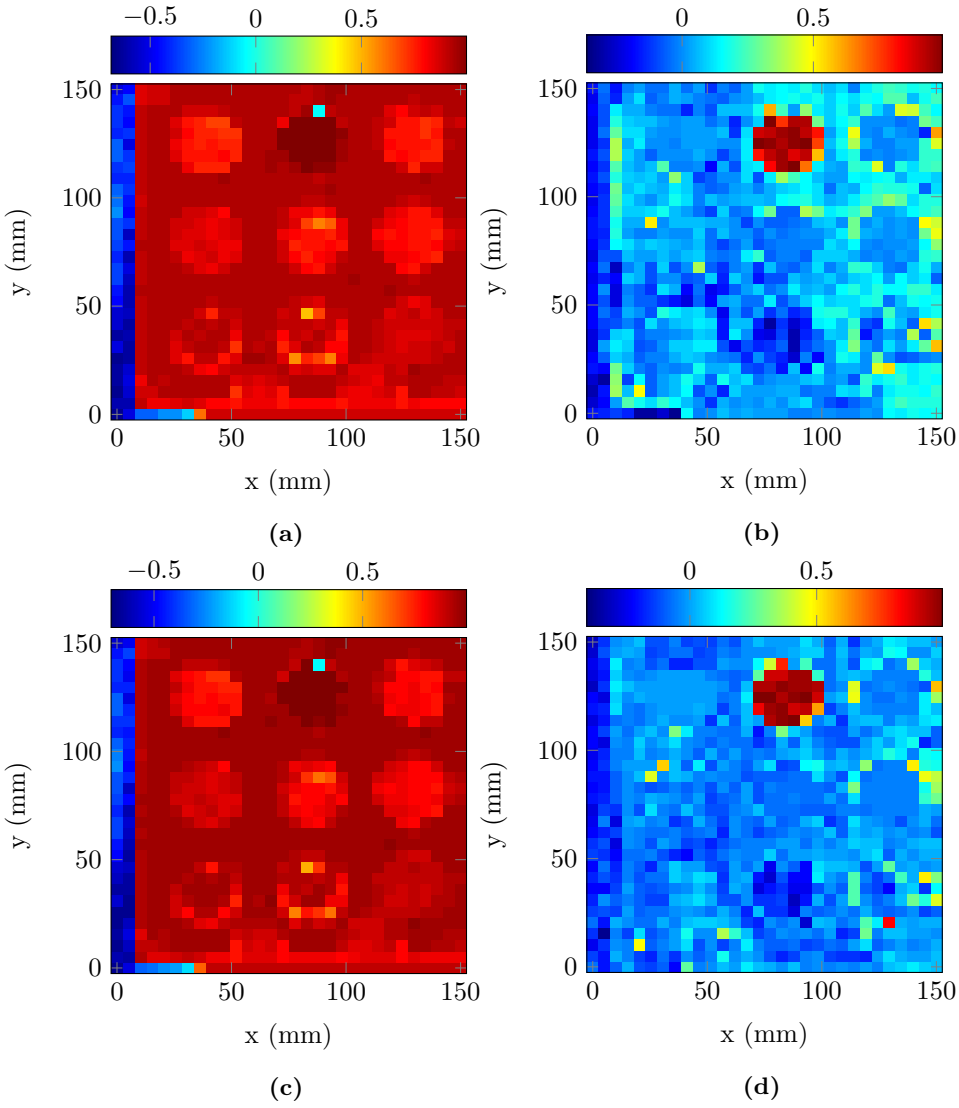


Figure B.4: Spectral correlation of Lactose with paper barrier, using (a) $A(f)$ and $\delta t = 30$ ps (b) $dA(f)/df$ and $\delta t = 30$ ps (c) $A(f)$ and $\delta t = 15$ ps (d) $dA(f)/df$ and $\delta t = 15$ ps

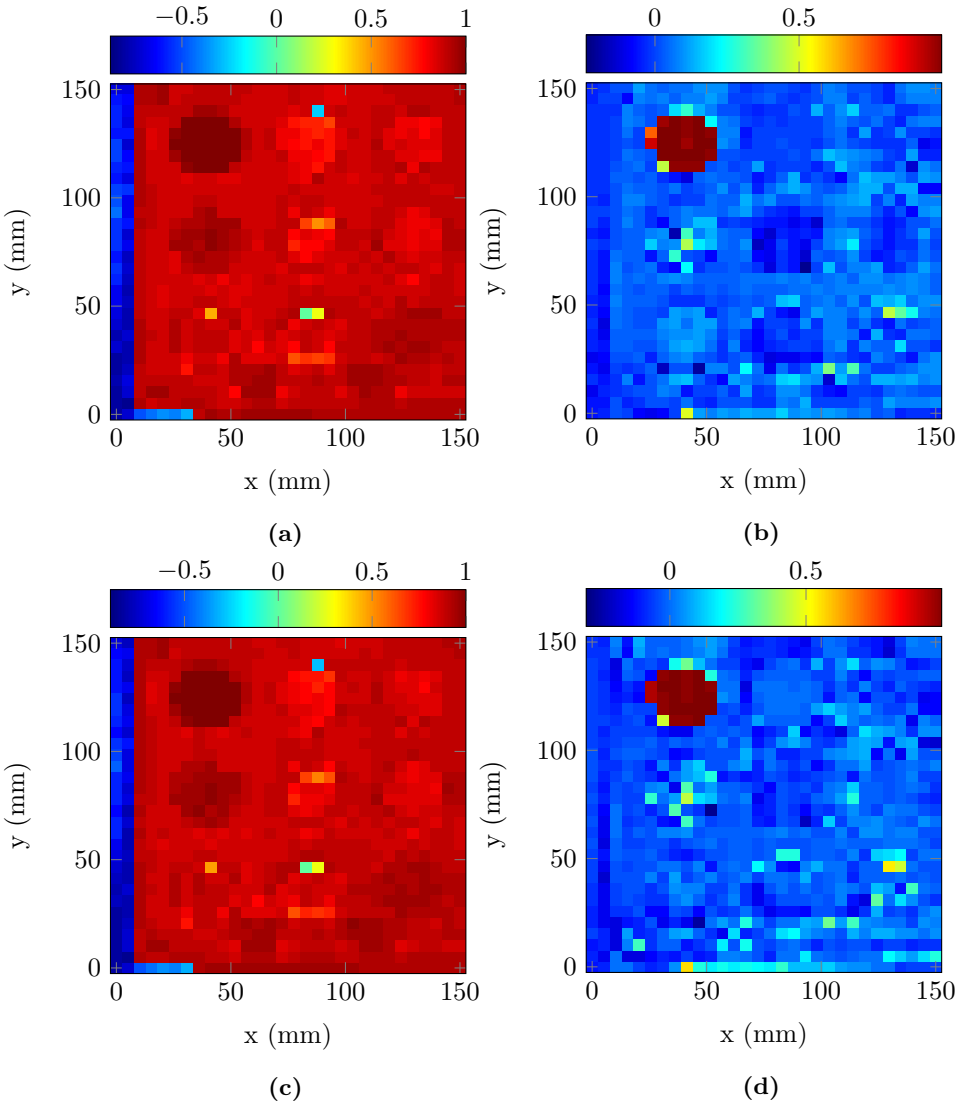


Figure B.5: Spectral correlation of RDX with paper barrier, using (a) $A(f)$ and $\delta t = 30$ ps (b) $dA(f)/df$ and $\delta t = 30$ ps (c) $A(f)$ and $\delta t = 15$ ps (d) $dA(f)/df$ and $\delta t = 15$ ps

B.3 Plastic barrier (Image 4)

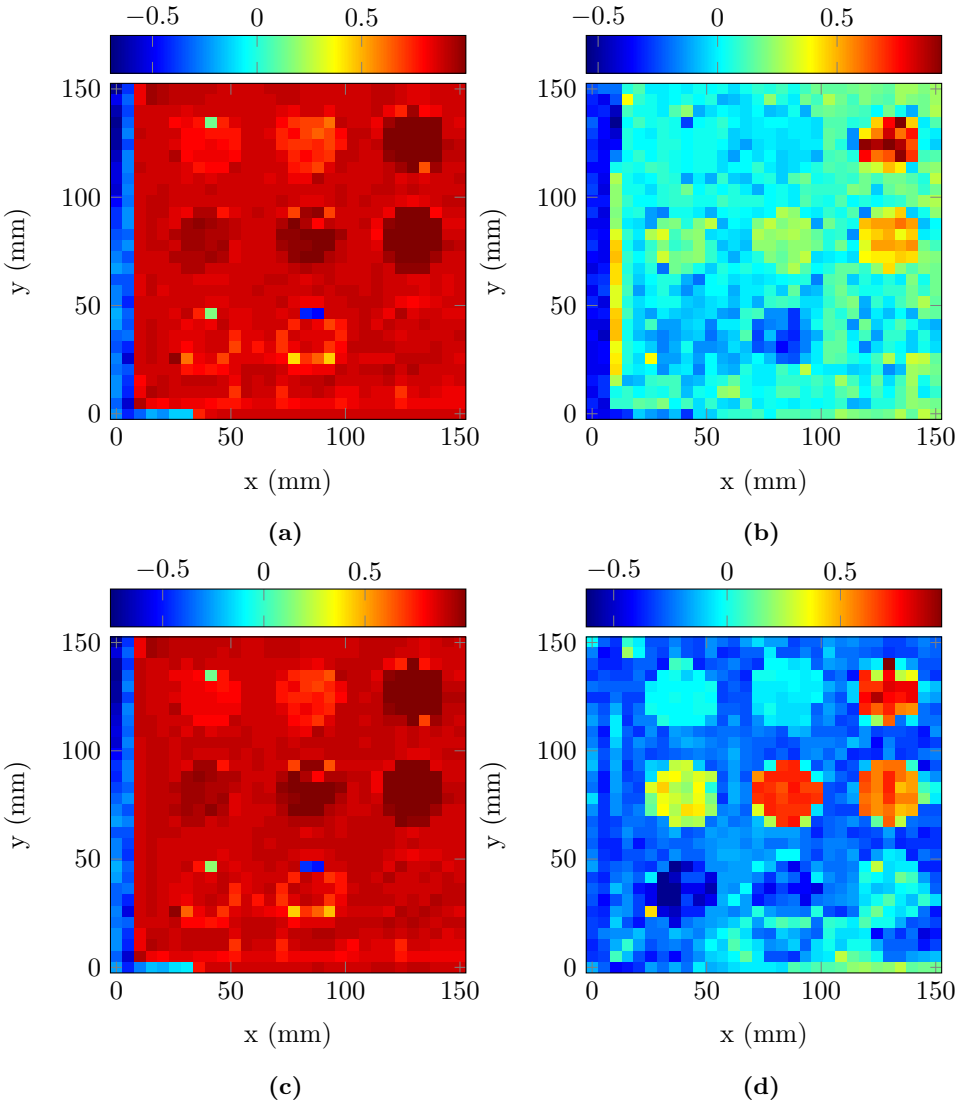


Figure B.6: Spectral correlation of Tartaric acid with plastic barrier, using (a) $A(f)$ and $\delta t = 30$ ps (b) $dA(f)/df$ and $\delta t = 30$ ps (c) $A(f)$ and $\delta t = 15$ ps (d) $dA(f)/df$ and $\delta t = 15$ ps

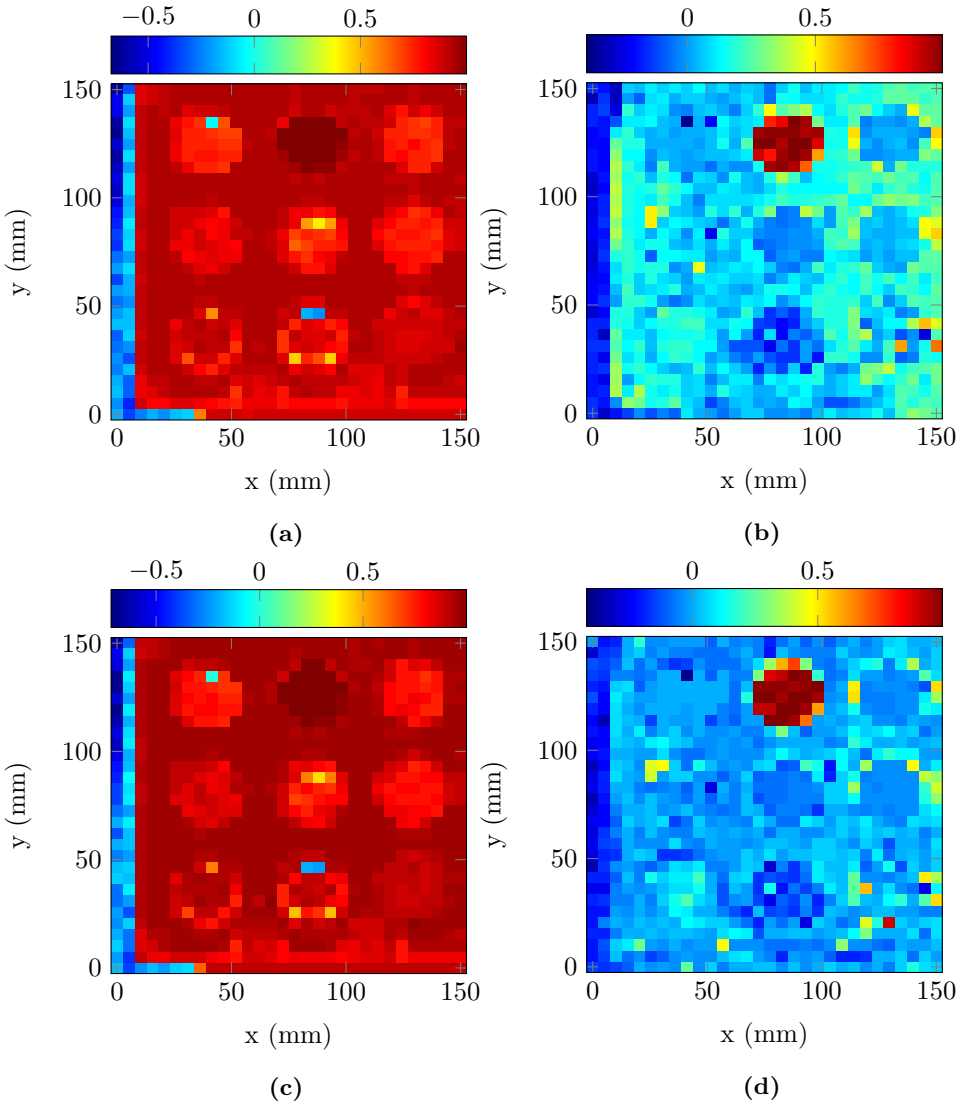


Figure B.7: Spectral correlation of Lactose with plastic barrier, using (a) $A(f)$ and $\delta t = 30$ ps (b) $dA(f)/df$ and $\delta t = 30$ ps (c) $A(f)$ and $\delta t = 15$ ps (d) $dA(f)/df$ and $\delta t = 15$ ps.

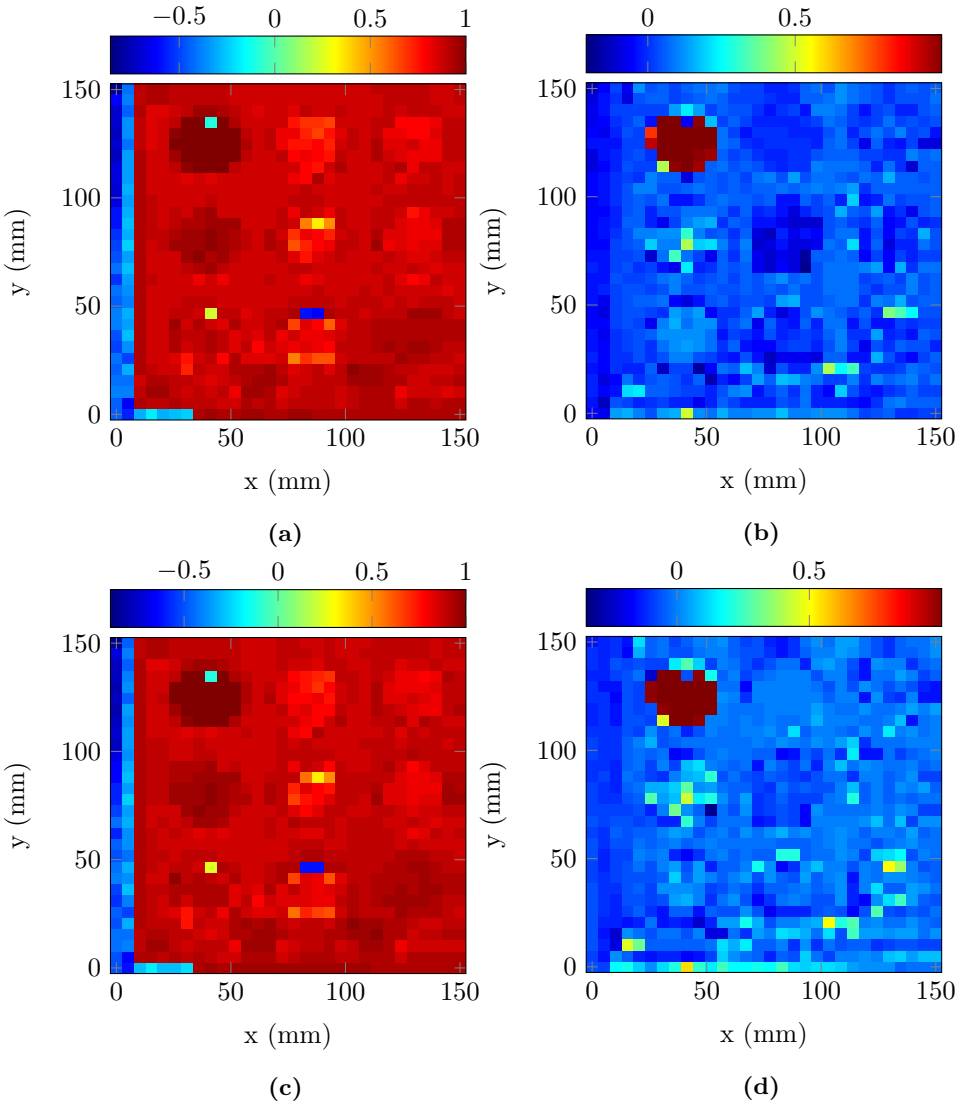


Figure B.8: Spectral correlation of RDX with plastic barrier, using (a) $A(f)$ and $\delta t = 30$ ps (b) $dA(f)/df$ and $\delta t = 30$ ps (c) $A(f)$ and $\delta t = 15$ ps (d) $dA(f)/df$ and $\delta t = 15$ ps

B.4 Cloth barrier (Image 5)

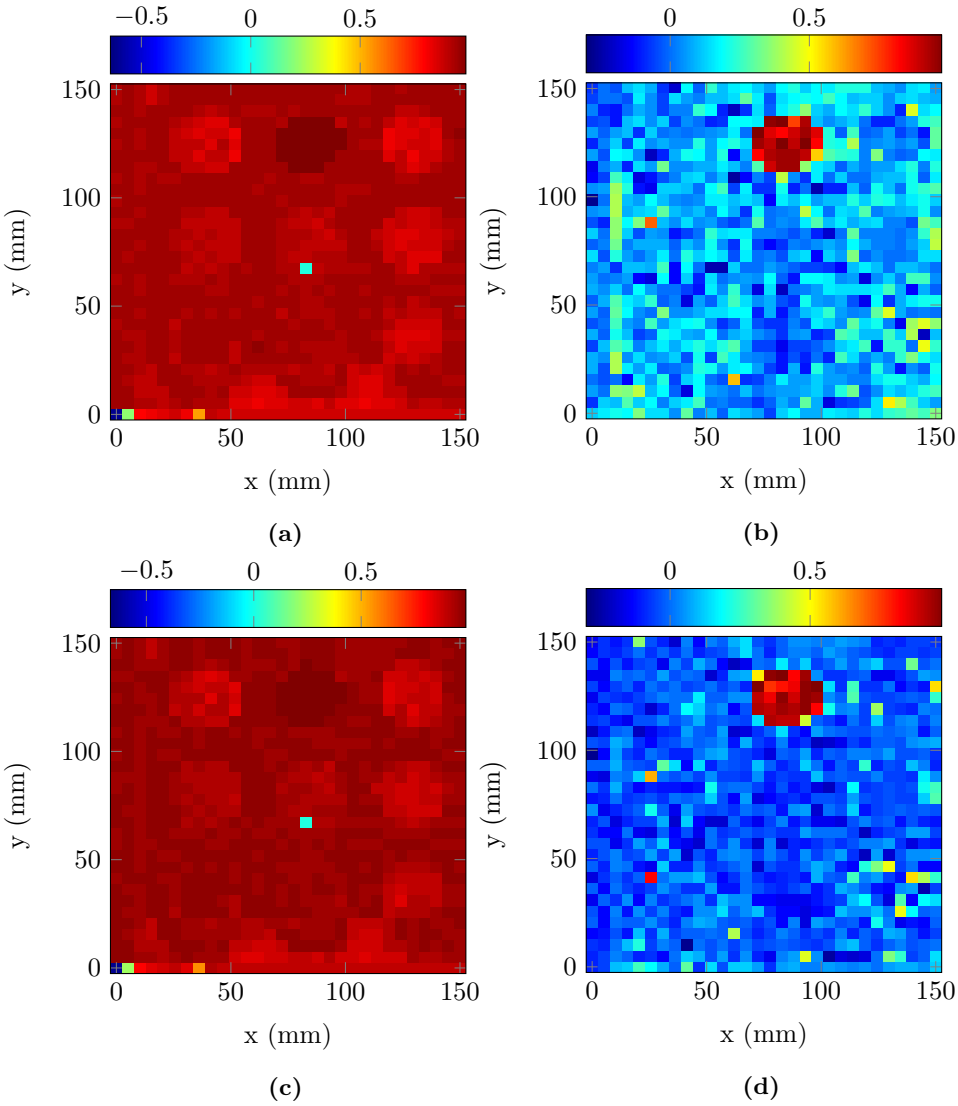


Figure B.9: Spectral correlation of Lactose with cloth barrier, using (a) $A(f)$ and $\delta t = 30$ ps (b) $dA(f)/df$ and $\delta t = 30$ ps (c) $A(f)$ and $\delta t = 15$ ps (d) $dA(f)/df$ and $\delta t = 15$ ps.

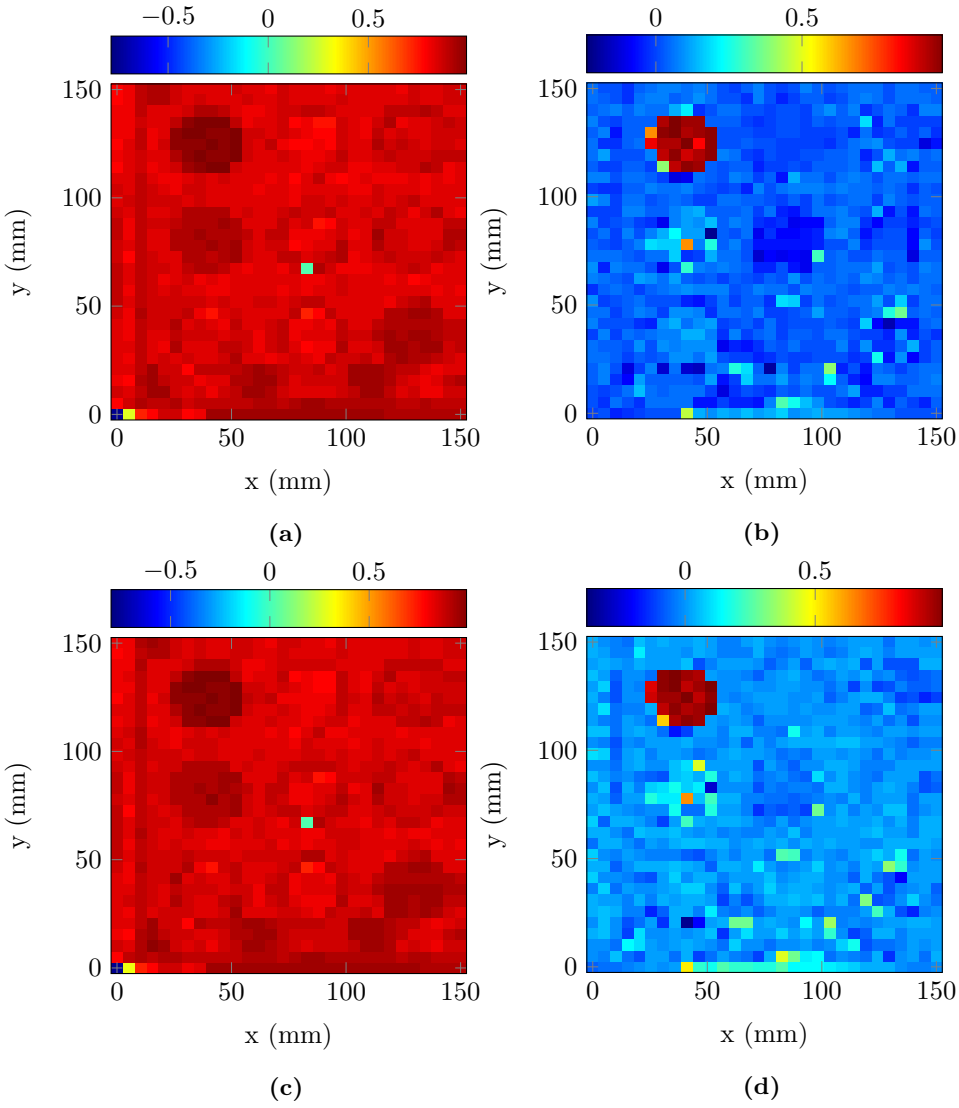


Figure B.10: Spectral correlation of RDX with cloth barrier, using (a) $A(f)$ and $\delta t = 30$ ps (b) $dA(f)/df$ and $\delta t = 30$ ps (c) $A(f)$ and $\delta t = 15$ ps (d) $dA(f)/df$ and $\delta t = 15$ ps.

Appendix C: ROC curves (SAM)

C.1 Paper barrier (Image 3)

Figures C.1 - C.3 show the ROC curves for Tartaric acid, Lactose and RDX, respectively, when covering the samples with a paper barrier.For Tartaric acid, an expected reduce in the performance is observed using dA(f)/df and $\delta t = 30$ ps. We see from the corresponding spectral correlation image (Figure B.3b in Appendix B) that the paper barrier causes lower correlation for all samples, especially Samples 4-6 compared to when no barrier is used. Note also that when using *A(f)*, the performance is actually better with the paper barrier than without. We see that the number of true positives is about the same at the point where the curves start flattening out, so the increased performance is mainly due to fewer false positives. It is hard to say exactly what causes this, but the paper might make e.g. the absorption spectra of the sample holder resemble the absorption spectra of the Tartaric acid samples less, thereby leading to a fewer number of false positives.

Also for Lactose, we observe that the performance is slightly increased for some cases with the paper barrier, compared to when no barrier is used. Only 26 pixels of the Image are from the Lactose sample, so the true positive rate can change noticeably if a few more/less pixels are labelled correctly/wrong. The exception is using *A(f)* with $\delta t = 15$ ps, where a decrease in the performance when introducing the paper barrier is seen, which is expected based on the corresponding spectral correlation image (Figure B.4d).

For RDX, the ROC curves are nearly ideal, as was the case using no barrier. Because the spectral correlation images (Figures **??** and B.5 respectively) were nearly identical for all combinations of spectral characteristic and window width, and the RDX sample was clearly identified, this is expected.

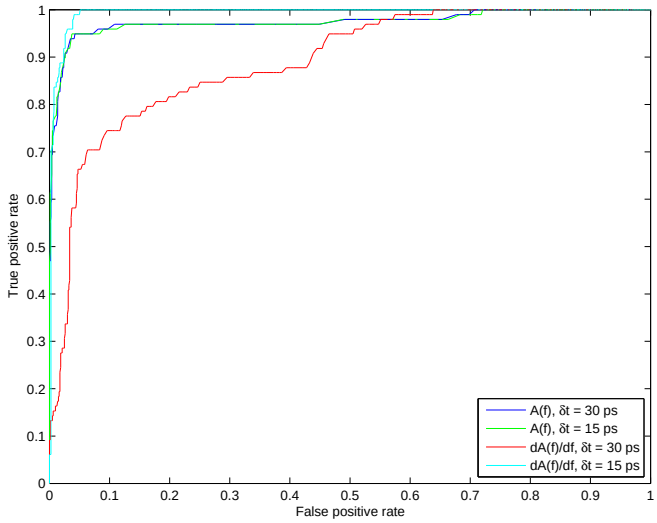


Figure C.1: ROC curves for identification of Tartaric acid using SAM, for the two spectral characteristics and window widths. Paper barrier.

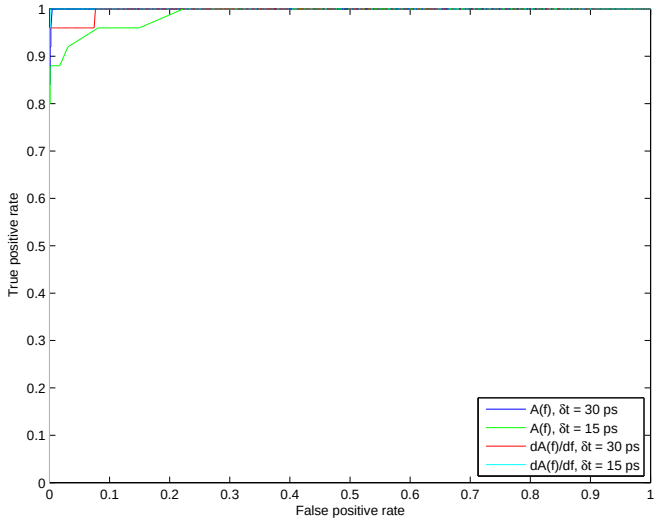


Figure C.2: ROC curves for identification of Lactose using SAM, for the two spectral characteristics and window widths. Paper barrier.

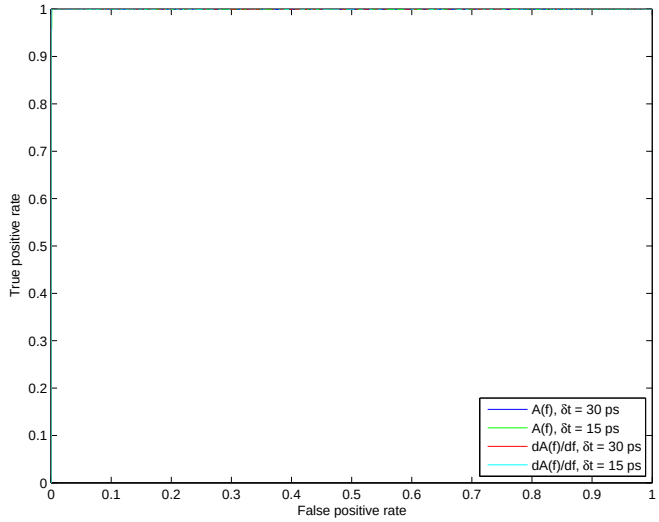


Figure C.3: ROC curves for identification of RDX using SAM, for the two spectral characteristics and window widths. Paper barrier.

C.2 Plastic barrier (Image 4) and cloth barrier (Image 5)

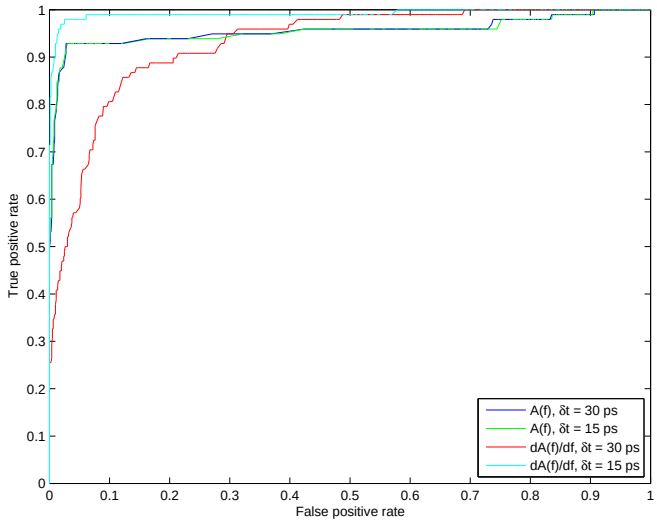


Figure C.4: ROC curves for identification of Tartaric acid using SAM, for the two spectral characteristics and window widths. Plastic barrier.

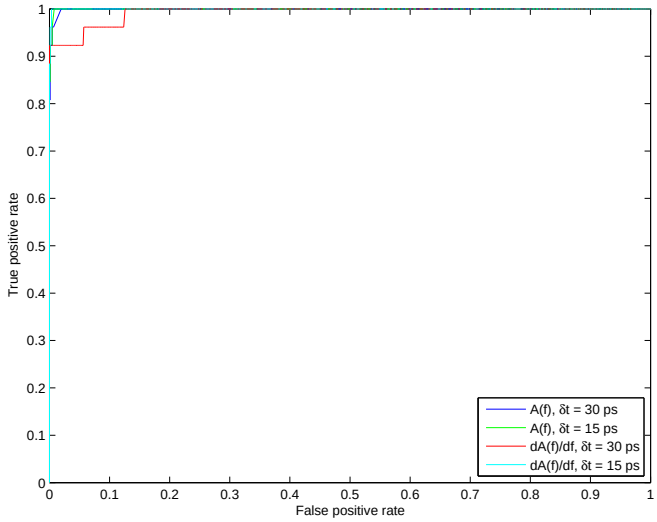


Figure C.5: ROC curves for identification of Lactose using SAM, for the two spectral characteristics and window widths. Plastic barrier.

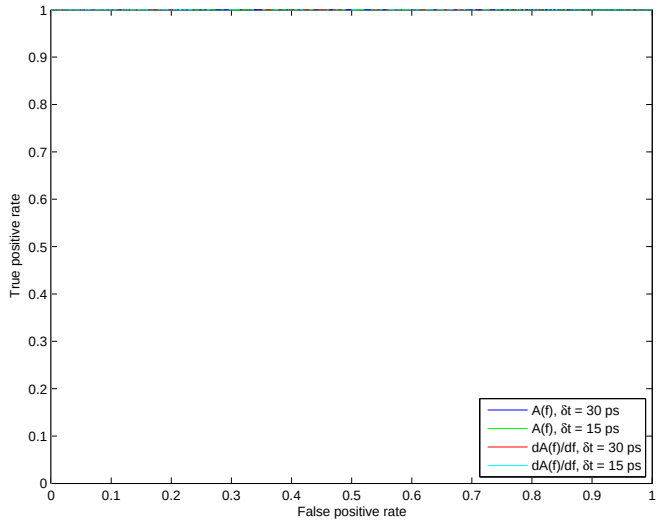


Figure C.6: ROC curves for identification of RDX using SAM, for the two spectral characteristics and window widths. Plastic barrier.

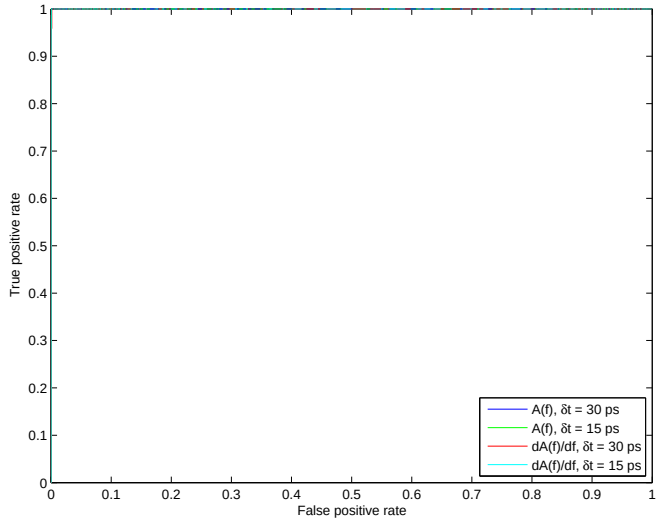


Figure C.7: ROC curves for identification of RDX using SAM, for the two spectral characteristics and window widths. Cloth barrier.

Appendix D: Score plots

D.1 No barrier (Image 2)

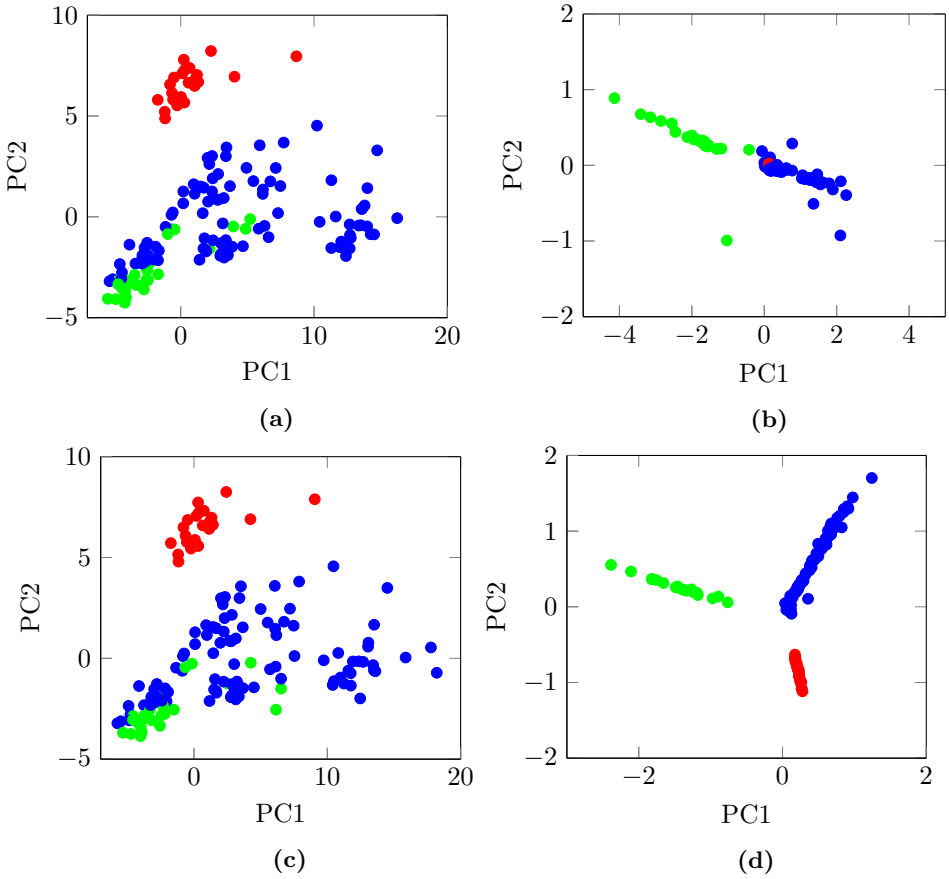


Figure D.1: Score plot of image with no barrier showing PC1 vs. PC2 for Tartaric acid (blue), Lactose (green) and RDX (red) using (a) *A(f)* and $\delta t = 30$ ps (b) *dA(f)/df* and $\delta t = 30$ ps (c) *A(f)* and $\delta t = 15$ ps (d) *dA(f)/df* and $\delta t = 15$ ps. Scores from surroundings omitted.

D.2 Cloth barrier (Image 5)

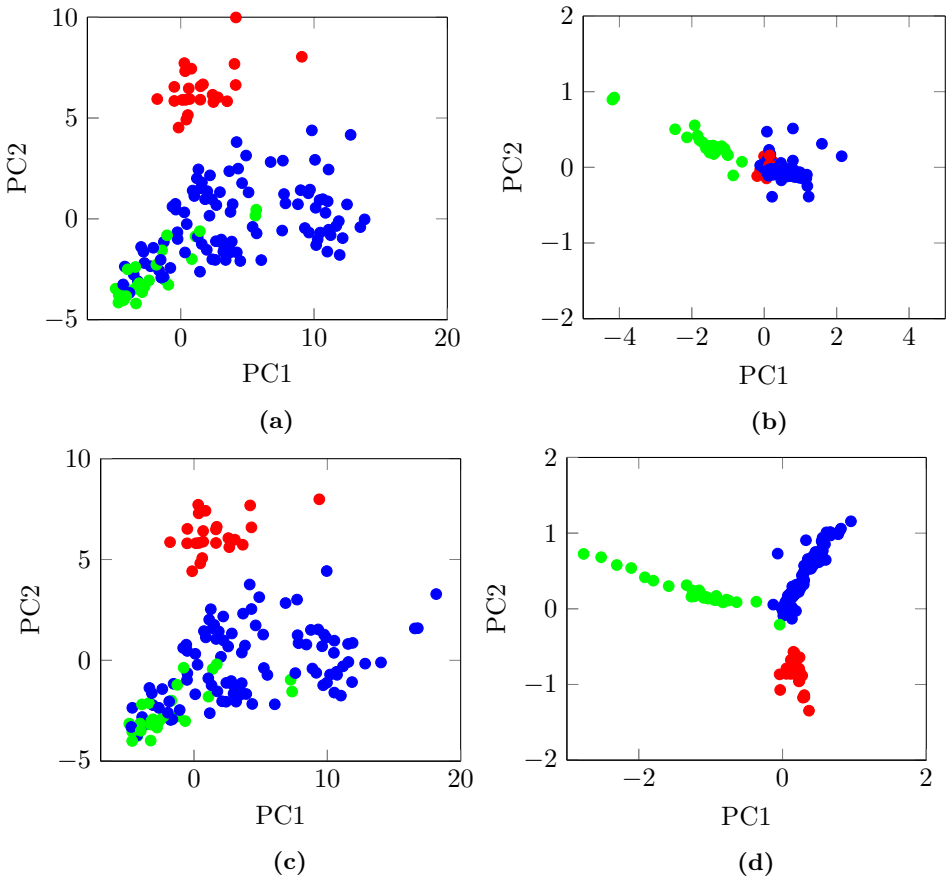


Figure D.2: Score plot of image with cloth barrier showing PC1 vs. PC2 for Tartaric acid (blue), Lactose (green) and RDX (red) using (a) *A(f)* and $\delta t = 30$ ps (b) *dA(f)/df* and $\delta t = 30$ ps (c) *A(f)* and $\delta t = 15$ ps (d) *dA(f)/df* and $\delta t = 15$ ps. Scores from surroundings omitted.

Appendix E: ROC curves (PCA)

E.1 No barrier (Image 2)

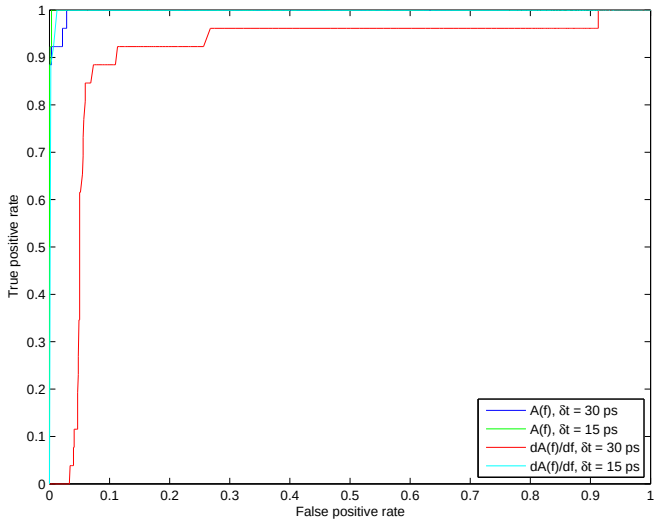


Figure E.1: ROC curves for identification of Lactose using PCA, for the two spectral characteristics and window widths. No barrier. Mahalanobis distance in three dimensions used for all cases.

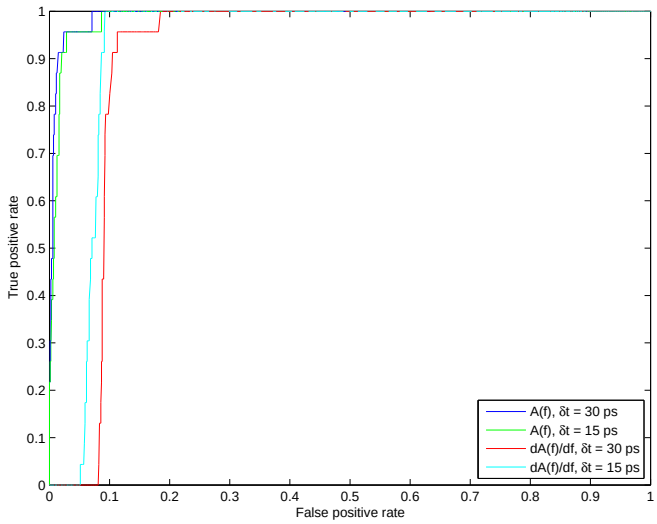


Figure E.2: ROC curves for identification of RDX using PCA, for the two spectral characteristics and window widths. No barrier. Mahalanobis distance in three dimensions used for all cases.

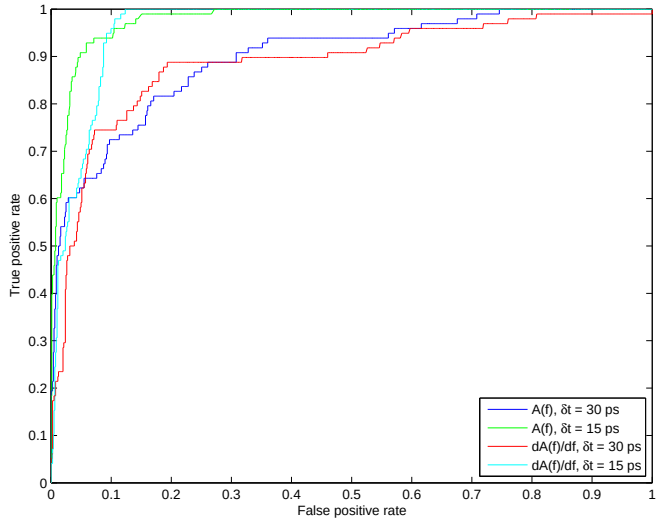
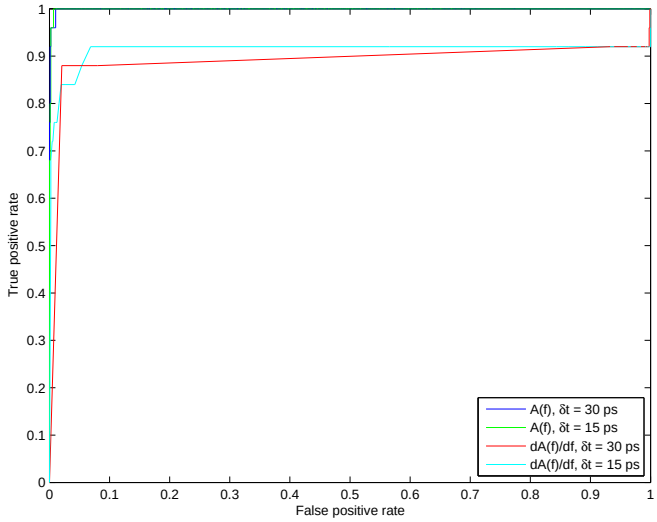
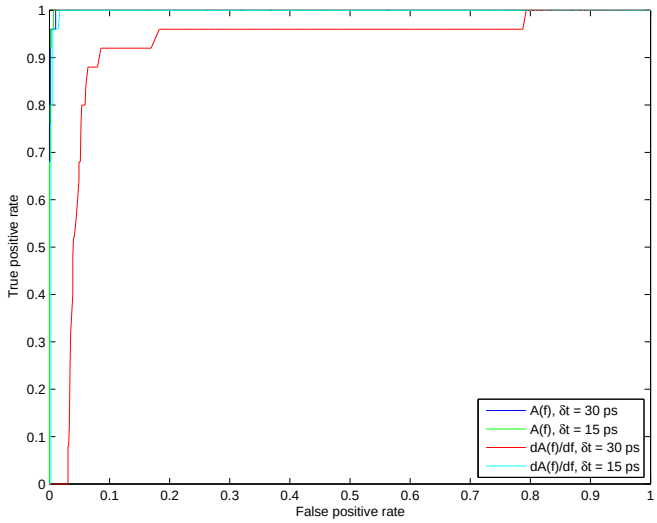


Figure E.3: ROC curves for identification of Tartaric acid using PCA, for the two spectral characteristics and window widths. Paper barrier.

E.2 Paper barrier (Image 3)

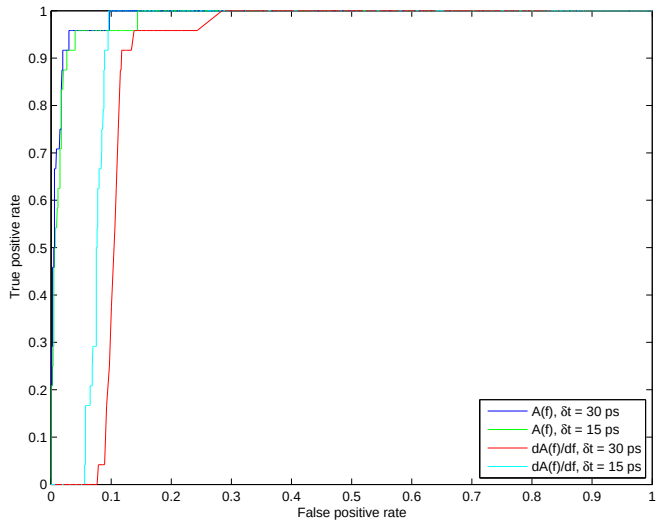


(a)

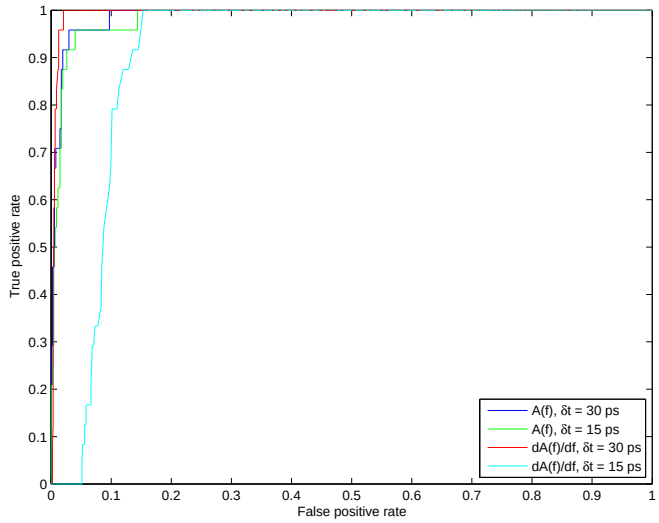


(b)

Figure E.4: ROC curves for identification of Lactose using PCA, for the two spectral characteristics and window widths. Paper barrier. (a) Mahalanobis distance in three dimensions for *A(f)* and one dimension for *dA(f)/df* (b) Mahalanobis distance in three dimensions for all cases.



(a)



(b)

Figure E.5: ROC curves for identification of RDX using PCA, for the two spectral characteristics and window widths. Paper barrier. (a) Mahalanobis distance in three dimensions for *A(f)* and one dimension for *dA(f)/df* (b) Mahalanobis distance in three dimensions for all cases.

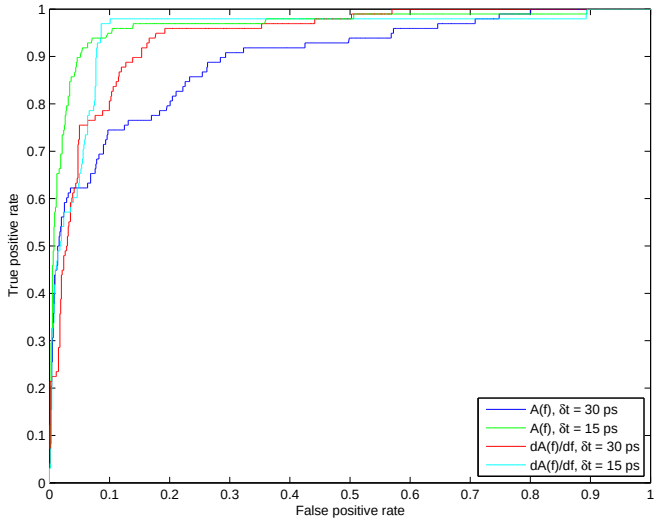
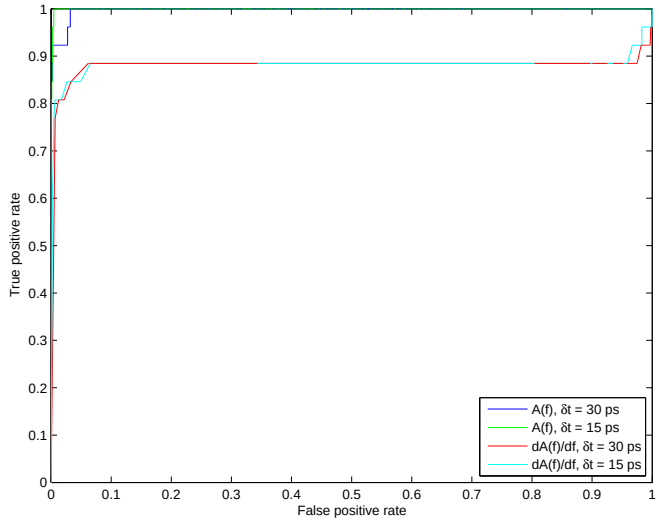
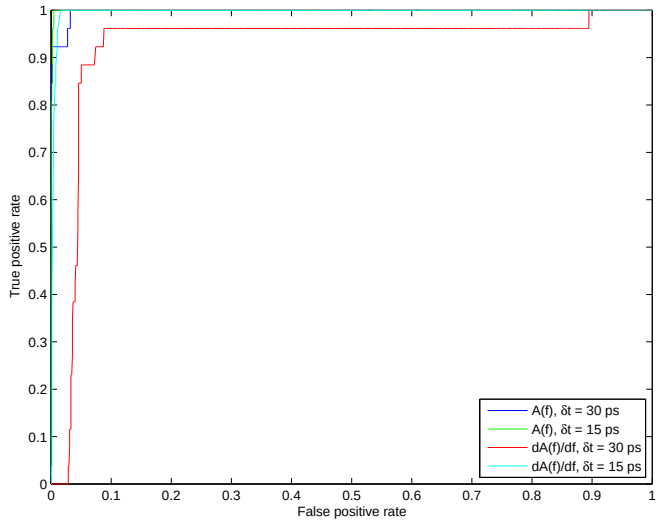


Figure E.6: ROC curves for identification of Tartaric acid using PCA, for the two spectral characteristics and window widths. Plastic barrier.

E.3 Plastic barrier (Image 4)

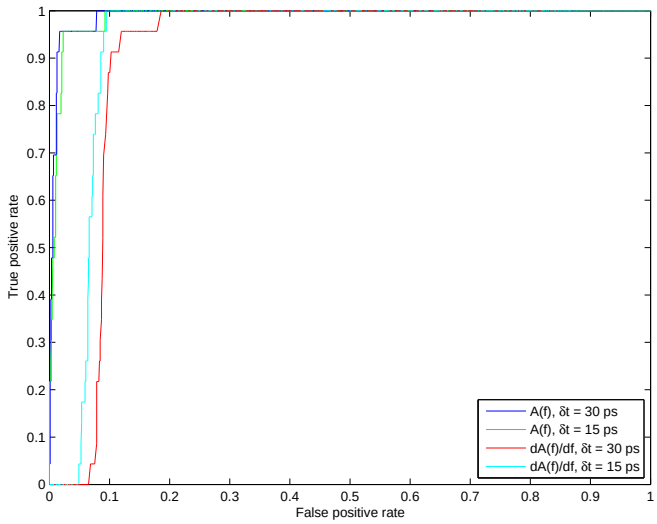


(a)

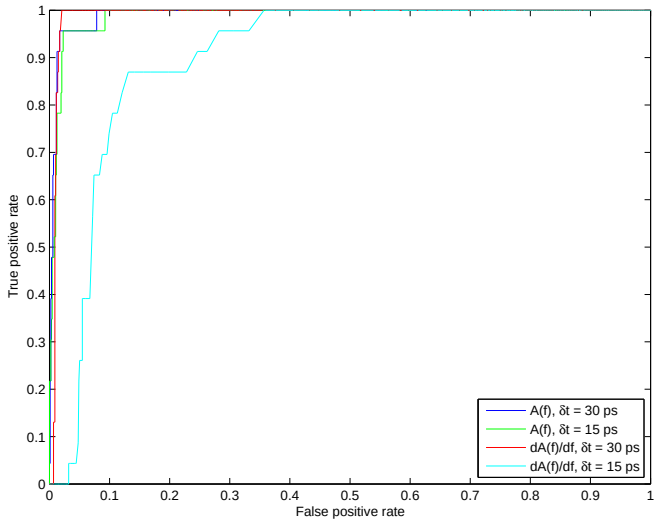


(b)

Figure E.7: ROC curves for identification of Lactose using PCA, for the two spectral characteristics and window widths. Plastic barrier. (a) Mahalanobis distance in three dimensions for *A(f)* and one dimension for *dA(f)/df* (b) Mahalanobis distance in three dimensions for all cases.



(a)



(b)

Figure E.8: ROC curves for identification of RDX using PCA, for the two spectral characteristics and window widths. Plastic barrier. (a) Mahalanobis distance in three dimensions for *A(f)* and one dimension for *dA(f)/df* (b) Mahalanobis distance in three dimensions for all cases.

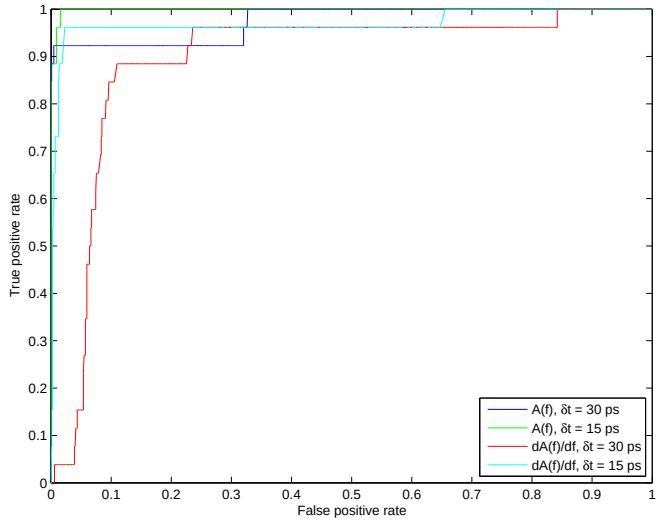


Figure E.9: ROC curves for identification of Lactose using PCA, for the two spectral characteristics and window widths. Cloth barrier. Mahalanobis distance in three dimensions used for all cases.

E.4 Cloth barrier (Image 5)

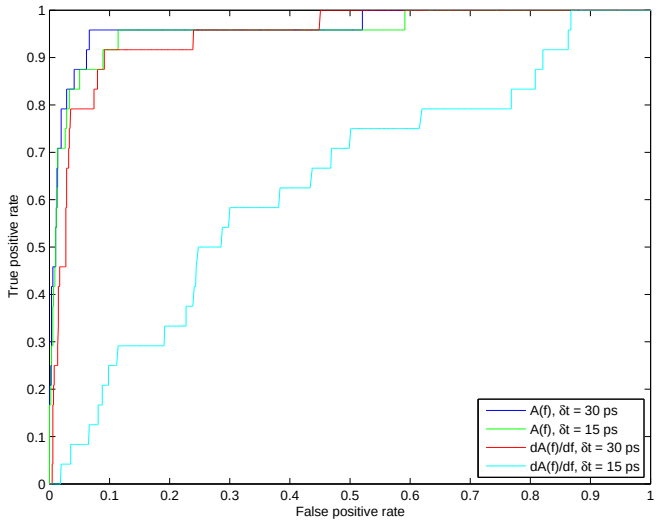


Figure E.10: ROC curves for identification of RDX using PCA, for the two spectral characteristics and window widths. Cloth barrier. Mahalanobis distance in three dimensions used for all cases.