Dissertation

zur Erlangung des Grades einer
Doktorin der Naturwissenschaften
der Technischen Universität Dortmund

Identification and Quantification of Peaks in Spectrometric Data

Der Fakultät Statistik der
Technischen Universität Dortmund
vorgelegt von

Sabine Bader

Dortmund 2008

Contents

Preliminaries

Theory

Spectra processing

List of Figures

List of Tables

List of Symbols

Ion mobility spectrometry

E	electric field strength
K	coefficient of ion mobility
l_D	length of the drift tube
n_D	number of drift time points
n_R	number of retention time points
$\mathbf{S}$	MCC/IMS spectra series matrix
$\mathbf{s}$	ion mobility spectrum
U_D	drift voltage
v_D	drift velocity
x	drift time value
x_i	ith drift time value of an ion mobility spectrum
y_j	jth retention time value of an MCC/IMS measurement
z_i	ith intensity value in an ion mobility spectrum
z_{ij}	value corresponding to the ith drift time and the jth retention time value

Wavelet transform

$\mathbf{A}_j$	matrix defining the jth step of the pyramid algorithm containing the circularly shifted scaling filter periodised to length $N_j - 1$
$\mathbf{A}_{j,i}$	matrix defining the jth step in the ith dimension of the two-dimensional pyramid algorithm containing the circularly shifted scaling filter periodised to length $n_i - 1$
$\tilde{\mathbf{A}}_j$	matrix defining the jth step of the MODWT pyramid algorithm containing the circularly shifted scaling filter periodised to length N

$\tilde{\mathbf{A}}_{j,i}$ matrix defining the jth step in the ith dimension of the two-dimensional MODWT pyramid algorithm containing the circularly shifted scaling filter periodised to length n_i

$\mathbf{B}_j$ matrix defining the jth step of the pyramid algorithm containing the circularly shifted wavelet filter periodised to length $N_j - 1$

$\mathbf{B}_{j,i}$ matrix defining the jth step in the ith dimension of the two-dimensional pyramid algorithm containing the circularly shifted wavelet filter periodised to length $n_i - 1$

$\tilde{\mathbf{B}}_j$ matrix defining the jth step of the MODWT pyramid algorithm containing the circularly shifted wavelet filter periodised to length N

$\tilde{\mathbf{B}}_{j,i}$ matrix defining the jth step in the ith dimension of the two-dimensional MODWT pyramid algorithm containing the circularly shifted wavelet filter periodised to length n_i

$\mathcal{D}_j$ jth level wavelet detail

$\mathcal{D}_j^{(HH)}$ jth level two-dimensional wavelet detail based on $\mathbf{W}_j^{(HH)}$

$\mathcal{D}_j^{(HL)}$ jth level two-dimensional wavelet detail based on $\mathbf{W}_j^{(HL)}$

$\mathcal{D}_j^{(LH)}$ jth level two-dimensional wavelet detail based on $\mathbf{W}_j^{(LH)}$

$\tilde{g}_i^\circ$ periodised scaling filter for the ith dimension of the two-dimensional MODWT

$\tilde{g}_l$ lth value of the MODWT scaling filter

$\tilde{g}_l^\circ$ lth value of the periodised MODWT scaling filter

g_i° periodised scaling filter for the ith dimension of the two-dimensional wavelet transform

g_l lth value of the scaling filter

g_l° lth value of the periodised scaling filter

$\tilde{h}_i^\circ$ periodised wavelet filter for the ith dimension of the two-dimensional MODWT

$\tilde{h}_l$ lth value of the MODWT wavelet filter

$\tilde{h}_l^\circ$ lth value of the periodised MODWT wavelet filter

h_i° periodised wavelet filter for the ith dimension of the two-dimensional wavelet transform

h_l lth value of the wavelet filter

h_l° lth value of the periodised wavelet filter

$\mathbf{I}_N$	identity matrix of dimension $(N \times N)$
λ	scaling factor
L	length of the wavelet and scaling filter
N	length of signal $\mathbf{s}$
n_i	ith dimension of the data matrix $\mathbf{S}$
N_j	number of wavelet coefficients associated with changes at scale τ_j
$\psi(\ )$	wavelet function
$\mathbf{P}_j$	orthonormal matrix allowing the synthesis of the signal $\mathbf{s}$ on step j of the pyramid algorithm
$\mathbf{w}$	vector of DWT coefficients
$\mathcal{S}_j$	jth level wavelet smooth
$\mathcal{S}_j^{(LL)}$	jth level two-dimensional wavelet smooth based on $\mathbf{W}_j^{(LL)}$
τ_j	jth downsampled scale
t	translation constant
t_i	translation constant for the ith dimension of a two-dimensional wavelet transform
$\mathbf{v}_{j\bullet}^T$	$(j+1)$th row of the matrix $\mathbf{V}_1$ of the pyramid algorithm
$\mathbf{V}_J$	submatrix of $\mathbf{W}$ defining the scaling coefficient
$\mathbf{V}_j$	matrix defining the jth step of the pyramid algorithm giving the base for obtaining the remaining wavelet coefficients on the successive stages
$\mathbf{v}_J$	subvector of $\mathbf{w}$ corresponding to the scaling coefficient
$\mathbf{v}_j$	vector of coefficients $V_{j,t}$
$\tilde{\mathbf{v}}_J$	subvector of the MODWT vector $\tilde{\mathbf{v}}$ corresponding to the scaling coefficient
$\tilde{\mathbf{v}}_j$	vector of MODWT coefficients $\tilde{V}_{j,t}$
$\tilde{\mathbf{V}}_J$	submatrix of the MODWT matrix $\tilde{\mathbf{V}}$ defining the scaling coefficient
$\tilde{V}_{j,t}$	MODWT coefficient constituted on the jth step of the pyramid algorithm
$V_{j,t}$	coefficient constituted on the jth step of the pyramid algorithm
$\mathbf{W}$	matrix defining the DWT
$\mathbf{w}_{j\bullet}^T$	$(j+1)$th row of the matrix $\mathbf{W}_1$ of the pyramid algorithm
$\mathbf{W}_j^{(HH)}$	jth level high-pass filtered matrix of a two-dimensional wavelet decomposition
$\mathbf{W}_j^{(HL)}$	jth level high/low-pass filtered matrix of a two-dimensional wavelet decomposition

$\mathbf{W}_j^{(LH)}$ jth level low/high-pass filtered matrix of a two-dimensional wavelet decomposition

$\mathbf{W}_j^{(LL)}$ jth level low-pass filtered matrix of a two-dimensional wavelet decomposition

$\mathbf{W}_j$ submatrix of $\mathbf{W}$ corresponding to scale τ_j

$\mathbf{w}_j$ subvector of $\mathbf{w}$ corresponding to scale τ_j

$\tilde{\mathbf{W}}_j^{(HH)}$ jth level high-pass filtered matrix of a two-dimensional MODWT decomposition

$\tilde{\mathbf{W}}_j^{(HL)}$ jth level high/low-pass filtered matrix of a two-dimensional MODWT decomposition

$\tilde{\mathbf{W}}_j^{(LH)}$ jth level low/high-pass filtered matrix of a two-dimensional MODWT decomposition

$\tilde{\mathbf{W}}_j^{(LL)}$ jth level low-pass filtered matrix of a two-dimensional MODWT decomposition

$\tilde{\mathbf{W}}_j$ submatrix of the MODWT matrix $\tilde{\mathbf{W}}$ corresponding to scale τ_j

$\tilde{\mathbf{w}}_j$ subvector of the MODWT vector $\tilde{\mathbf{w}}$ corresponding to scale τ_j

$\tilde{W}_{j,t_1,t_2}^{(HH)}$ wavelet coefficient of a two-dimensional MODWT of scale j at time combination t_1, t_2 in $\tilde{\mathbf{W}}_j^{(HH)}$

$\tilde{W}_{j,t_1,t_2}^{(HL)}$ wavelet coefficient of a two-dimensional MODWT of scale j at time combination t_1, t_2 in $\tilde{\mathbf{W}}_j^{(HL)}$

$\tilde{W}_{j,t_1,t_2}^{(LH)}$ wavelet coefficient of a two-dimensional MODWT of scale j at time combination t_1, t_2 in $\tilde{\mathbf{W}}_j^{(LH)}$

$\tilde{W}_{j,t_1,t_2}^{(LL)}$ wavelet coefficient of a two-dimensional MODWT of scale j at time combination t_1, t_2 in $\tilde{\mathbf{W}}_j^{(LL)}$

$\tilde{W}_{j,t}$ MODWT wavelet coefficient at time t constituted on the jth step of the pyramid algorithm

$W(\lambda, t)$ wavelet coefficient of scale λ at time t

$W_{j,t_1,t_2}^{(HH)}$ wavelet coefficient of a two-dimensional wavelet transform of scale j at time combination t_1, t_2 in $\mathbf{W}_j^{(HH)}$

$W_{j,t_1,t_2}^{(HL)}$ wavelet coefficient of a two-dimensional wavelet transform of scale j at time combination t_1, t_2 in $\mathbf{W}_j^{(HL)}$

$W_{j,t_1,t_2}^{(LH)}$ wavelet coefficient of a two-dimensional wavelet transform of scale j at time combination t_1, t_2 in $\mathbf{W}_j^{(LH)}$

$W_{j,t_1,t_2}^{(LL)}$	wavelet coefficient of a two-dimensional wavelet transform of scale j at time combination t_1, t_2 in $\mathbf{W}_j^{(LL)}$
W_n	nth wavelet coefficient of the DWT
$W_{j,t}$	wavelet coefficient at time t constituted on the jth step of the pyramid algorithm

Statistical methods

α	significance level of a statistical test
α_k	multiple significance level for k tests
$\mathbf{a}$	vector for the definition of a linear decision rule
$a(\omega_n)$	average distance between ω_n and all other objects of the same cluster
a_j	discriminant coefficient corresponding to the jth variable
a_j^*	standardised discriminant coefficient corresponding to the jth variable
$\mathbf{B}(\mathcal{C})$	matrix of variances between the clusters of the partition $\mathcal{C}$
$b(\omega_n)$	average distance between ω_n and all objects in the neighbour cluster
$\hat{\mathcal{C}}$	optimal cluster partition according to the variance criterion of the k-means method
$\hat{C}_j$	jth cluster of the optimal partition of a set of objects according to the variance criterion of the k-means method
$\mathcal{C}^j$	partition of a set of objects on stage j of a hierarchical cluster procedure
$C(g, \hat{g})$	cost function
C_j	jth cluster of a partition of a set of objects
$C_p(g, \hat{g})$	inversely proportional cost function
$C_s(g, \hat{g})$	simple symmetric cost function
$(d_2(\omega_n, \omega_m))^2$	squared Euclidean distance between object ω_n and ω_m
$\mathbf{D}$	distance matrix
$d(\omega_n, C)$	distance measure between object ω_n and cluster C
$d(\mathbf{x})$	single discriminant function for the case of only two classes
$D(C_g, C_h)$	distance measure between cluster C_g and C_h
$d_2(\omega_n, \omega_m)$	Euclidean or L_2 distance between object ω_n and ω_m
$d_g(\mathbf{x})$	discriminant function for class g
D_j	fusion level on step j of a hierarchical cluster procedure

$d_q(\omega_n, \omega_m)$	Minkowski q-metric or L_q distance between object ω_n and ω_m
$d_{q;r}(\omega_n, \omega_m)$	generalised Minkowski metric between object ω_n and ω_m
$\epsilon(e)$	(total) error rate
$\epsilon_{g,\hat{g}(e)}$	individual error rate
e	decision rule
Γ	set of all possible partitions of N objects in k clusters
$\hat{g}$	estimated class index
g	class index
$H(\mathcal{C}^j)$	homogeneity of clusters in a partition $\mathcal{C}^j$
H_0	null hypothesis of a statistical test
H_0^j	jth null hypothesis in a multiple test problem
$H_0^{(j)}$	jth ordered null hypothesis in a multiple test problem
H_1	alternative hypothesis of a statistical test
H_1^j	jth alternative hypothesis in a multiple test problem
$H_1^{(j)}$	jth ordered alternative hypothesis in a multiple test problem
I	a set of classification objects
k	number of clusters
$\hat{\mu}_g$	estimation of the expected value vector of class g
μ_g	expected value vector of class g
μ_i	expected value of the ith variable
ν	degrees of freedom of a χ^2 distribution
N	number of objects
n_j	number of objects belonging to cluster C_j
Ω	population of objects
Ω_i	subpopulation of objects from Ω
ω_n	nth classification object
$d(\omega_n, \omega_m)$	distance measure between object ω_n and ω_m
π_g	a priori probability of class g
p	number of variables
$p_{(i)}$	ith ordered p-value in a multiple test problem
p_i	ith p-value in a multiple test problem
q	standardisation parameter of the generalised Minkowski metric

r	weighting parameter of the generalised Minkowski metric
$R(\mathbf{a})$	ratio for the deduction of Fisher's linear discriminant analysis
$\bar{s}$	average silhouette width for all objects in a cluster analysis
$\hat{\mathbf{\Sigma}}_g$	estimatation of the variance matrix of class g
$\mathbf{\Sigma}_g$	variance of class g
$\mathcal{S}$	sample space
σ_i	standard deviation of the ith variable
σ_i^2	variance of the ith variable
$\mathbf{S}_g^2$	empirical covariance matrix of class g
$\mathbf{S}_P^2$	pooled covariance matrix
$s(\omega_n)$	silhouette width of the nth object in a cluster analysis
s_g^2	empirical variance of the gth class
s_i	empirical standard deviation of the ith variable
$s_i^{(q;r)}$	adjusted empirical standard deviation for usage with the generalised Minkowski metric
S_j^2	variance of the realisations of the jth variable
$s_{P_{jj}}^{\frac{1}{2}}$	pooled standard deviation of the jth variable
T	test statistic of the t test
$\mathbf{W}(\mathcal{C})$	matrix of variances within the clusters of the partition $\mathcal{C}$
$\mathbf{W}$	matrix of variances within the groups
$\bar{\mathbf{x}}$	mean vector of the observed variable vectors of all classification objects
$\bar{\mathbf{x}}_g$	mean vector of class g
$\bar{\mathbf{x}}_i$	mean vector of the values observed for the ith cluster
$\bar{X}_j$	mean of the realisations of the jth variable
$\bar{x}_j$	mean of the values observed for the jth variable
$\bar{X}_{ij}$	realisation of the jth variable of object ω_i
$\hat{\bar{\mathbf{x}}}_j$	mean vector of the observed variable vectors of objects belonging to the jth cluster of the optimal partition of a set of objects according to the variance criterion of the k-means method
$\mathbf{X}_g$	data matrix of class g
$\mathbf{x}_n$	vector of variable values of object ω_n
$\mathbf{x}_{gn}$	vector of variable values for object n in class g

$\tilde{X}_i$	standardised ith variable
$\tilde{x}_{ni}$	standardised value of the ith variable for object ω_n
$\tilde{x}_{ni}^{(e)}$	empirical standardised value of the ith variable for object ω_n
$\tilde{x}_{ni}^{(t)}$	theoretical standardised value of the ith variable for object ω_n
X_j	realisation of the jth variable
x_j	value of the jth variable
X_{ij}	ith realisation of the jth variable
x_{nj}	value of the jth variable for object ω_n
$\bar{y}_g$	mean value of discriminant values in class g
y	linear discriminant value
y_{gn}	discriminant value of object n determined with the discriminant function for class g

Spectra processing and analysis

α	angle corresponding to the area $\mathbf{A}$
$\mathcal{A}$	set of data points belonging to the region constituting peak A
$\mathbf{A}$	area of a two-dimensional measurement space containing no peaks
a	extent of a peak ellipse in the drift time dimension
a^G	extent of an ellipsoid general peak area in the drift time dimension
a^L	extent of a peak ellipse in the drift time dimension detected in a MRA matrix of a lower frequency level
A°	area of a peak ellipse
a_1	position of the first value in the drift time dimension for that a data point belonging to a peak of interest is observed
a_2	position of the last value in the drift time dimension for that a data point belonging to a peak of interest is observed
a_d	shrinkage factor of the lognormal detailing function
a_{BSL}	baseline shift constant
β	peak spreading angle of the practical peak area
b	extent of a peak ellipse in the retention time dimension
b^G	extent of an ellipsoid general peak area in the retention time dimension
b^L	extent of a peak ellipse in the retention time dimension detected in a MRA matrix of a lower frequency level

b_p	constant in P giving a cut-off point for diminishing the influence of the reactant ion peak
$\mathbf{C}$	area of a two-dimensional measurement space containing no peaks
c_D	compression level of wavelet smoothing in the drift time dimensions
C_P	practical two-dimensional peak capacity
c_R	compression level of wavelet smoothing in the retention time dimensions
C_T	theoretical two-dimensional peak capacity
C_{corr}	correlation between the two peak position vectors of a two-dimensional separation
C_{cv}	peak capacity in the dimension of compensation voltage
C_{rt}	peak capacity in the dimension of retention time
Δ_i	difference of a spectrum and $L(\mathbf{x})$
$\bar{E}$	mean of the data points in E
E	set of data points lying in an ellipsoid general peak area
f_h^{MRA}	height factor for the normalisation of MRA matrices for comparability of the intensity values with the original data
γ	angle corresponding to the area $\mathbf{C}$
$\mathbf{I}_k$	intensity interval corresponding to the kth stage of the growing interval merging algorithm
$\mathbf{I}_{n_D,n_R}$	matrix of dimension $(n_D \times n_R)$ with all elements equal to 1
i_1	lower position number defining a noise area in the drift time dimension
i_2	higher position number defining a noise area in the drift time dimension
i_k	lower limit of interval $\mathbf{I}_k$
i_{RIP}	position number of the RIP maximum in the drift time dimension
j_1	lower position number defining a noise area in the retention time dimension
j_2	higher position number defining a noise area in the retention time dimension
K_0	reduced mobility
K_0^r	reproducible reduced mobility
k_q	coefficient of the quadratic term of the retention time correction
k_T	coefficient of the linear term of the retention time correction respecting the column temperature deviation from the standard value
λ	threshold for wavelet denoising

$L(\mathbf{x})$	lognormal detailing function
L_{K_0}	lower limit of a rectangular general peak area in inverse reduced mobility
$L_{t_r^{30}}$	lower limit of a rectangular general peak area in retention time
med(a)	median of the ellipse extents in the drift time dimension of all points belonging to a cluster
med(b)	median of the ellipse extents in the retention time dimension of all points belonging to a cluster
m	scale parameter of the lognormal detailing function
n	sample size
n_D^*	drift time dimension extended to dyadic length
n_R^*	retention time dimension extended to dyadic length
n_D^c	drift time dimension after wavelet compression
N_k	number of data points in the intensity interval $\mathbf{I}_k$ corresponding to the kth stage of the growing interval merging algorithm
n_R^c	retention time dimension after wavelet compression
n_s	number of stages of the growing interval merging algorithm
$\mathbf{P}_k$	peak list emerging on the kth stage of the growing interval merging
P	penalty term for optimisation of the parameters of $L(\mathbf{x})$
p	ambient pressure
p_0	standard pressure
p_{abs}	factor in P assigning an absolute penalty for spectra points above $L(\mathbf{x})$
r_p	noise threshold in P allowing for little penalty-free variation of the spectrum around $L(\mathbf{x})$
$\bar{s}_a$	maximum of the average silhouette width for solutions resulting from the k-means methods with starting values derived from the average linkage
$\bar{s}_w$	maximum of the average silhouette width for solutions resulting from the k-means methods with starting values derived from Ward's method
σ_s	shape parameter of the lognormal detailing function
σ_{noise}	standard deviation in a pure noise area
$\mathbf{s}$	baseline corrected spectrum
$\mathbf{s}_{med}$	spectrum of medians per original drift time point across the spectra series
$\mathbf{S}^*$	baseline corrected MCC/IMS spectra series matrix

$\mathbf{s}_{i.}$	row vector of the MCC/IMS spectra series matrix containing the intensity values across spectra for one original drift time point x_i
s_{med}^{max}	maximum position of the median spectrum
θ	location parameter of the lognormal detailing function
T	ambient temperature
T_0	standard temperature
t_n	minimum level for peak detection
t_{grid}	shutter grid opening time
t_{RIP}	intensity threshold for the determination of a spectra truncation position
U_{K_0}	upper limit of a rectangular general peak area in inverse reduced mobility
$U_{t_r^{30}}$	upper limit of a rectangular general peak area in retention time
V_G	jth general peak variable
$\tilde{W}_{j,t_1,t_2}^{h}$	wavelet coefficient of the jth scale after hard denoising
$\tilde{W}_{j,t_1,t_2}^{s}$	wavelet coefficient of the jth scale after soft denoising
$\mathbf{x}$	values of the reproducible inverse reduced mobility
x^*	actual drift time after correction with $t_{grid}/2$
x_{RIP}^*	position of the RIP maximum in actual drift time
x_0	position of a peak ellipse in the drift time dimension
x_0^G	third quartile of the positions in the drift time dimension of all peak points belonging to a cluster
x_0^H	position of a peak ellipse in the drift time dimension detected in a MRA matrix of a higher frequency level
x_0^k	position of a peak ellipse in the drift time dimension detected on the kth stage of the growing interval merging
x_0^L	position of a peak ellipse in the drift time dimension detected in a MRA matrix of a lower frequency level
$x_{0_{0.25}}$	position of an ellipsoid general peak area in the drift time dimension
$x_{0_{0.75}}$	first quartile of the positions in the drift time dimension of all peak points belonging to a cluster
x_M^A	modus of all values in the drift time dimension of data points belonging to peak A
$x_{\max}$	maximum position of a spectrum in the drift time dimension
$x_{\max}^{log}$	maximum position of the detailing function in the drift time dimension

x_{RIP}	position of the reactant ion peak in the drift time dimension
$\mathbf{y}$	values of the corrected retention time
y^{30}	corrected retention time adjusting for varying column temperature
y_0	position of a peak ellipse in the retention time dimension
y_0^G	position of an ellipsoid general peak area in the retention time dimension
y_0^H	position of a peak ellipse in the retention time dimension detected in a MRA matrix of a higher frequency level
y_0^k	position of a peak ellipse in the retention time dimension detected on the kth stage of the growing interval merging
y_0^L	position of a peak ellipse in the retention time dimension detected in a MRA matrix of a lower frequency level
$y_{0_{0.25}}$	first quartile of the positions in the retention time dimension of all peak points belonging to a cluster
$y_{0_{0.75}}$	third quartile of the positions in the retention time dimension of all peak points belonging to a cluster
y_M^A	modus of all values in the retention time dimension of data points belonging to peak A

Preliminaries

Chapter 1

Introduction

Spectrometric methods offer a high variety of potential applications, not only in highly specialised laboratories, but also in everyday medical practices, airports, and even space shuttles. Coupled to other analytical methods to achieve more accurate separation results, they generate a high amount of complex data which is hard to analyse manually without automated algorithms. The central goal of this thesis is, therefore, to develop efficient processing methods for three-dimensional spectrometric data.

1.1 Motivation and objectives

"Sex, beer, and lung cancer" – this phrase, sounding like the vita of a rock'n'roll star, was actually part of a headline published to promote a new screening instrument for the monitoring of human breath.[1] Using ion mobility spectrometry, this device allows the detection of volatile organic compounds (VOCs) in exhaled air to give various information, for example about the sexual activity of men by the presence of pentane. While the same instrumentation can also be used for monitoring the fermentation processes of beer, it was now colocated in a lung hospital to analyse the composition of metabolites in human breath in correlation with different lung diseases.

In a collaboration with the ISAS - Institute for Analytical Sciences, Dortmund, and the lung hospital Hemer, the exhaled air of patients suffering from different lung diseases was studied to screen for characteristic patterns allowing an early diagnosis of bronchial maladies. After a pilot study comparing the analytic measurements of patients with lung

[1] Original German title "Künstliche Nasen riechen Sex, Bier und Lungenkrebs", published in "Stuttgarter Zeitung" on 14/11/2006.

cancer and those of healthy control persons yielded initial promising results presented in this work, the study period was extended for a further four years. Although more diseases were observed over the duration of this survey, this thesis concentrates on lung cancer – firstly comparing with a control group, and later with the aim to discriminate between different forms of tumors.

Ion mobility spectrometers (IMS) are especially suitable for this application as they can be used at ambient pressure, with air as the carrier gas, and offer short analysis times and low costs. By coupling the spectrometer with a multi-capillary column (MCC), the sample humidity can be separated at the very beginning of the analysis. The measurements of this two-dimensional separation, however, consist of a high amount of complex data, making it hard to extract the relevant information.

The objectives of this work, therefore, were the provision of data analysis methods allowing the efficient characterisation of MCC/IMS measurements via peak detection and quantification, as well as the development of a sufficient preprocessing strategy. Additionally, discriminating between different groups of patients and control persons, involved the preparation of the obtained measurement characterisations for further statistical evaluations.

1.2 Methodology

The methodology for peak detection and quantification can be best described with the aid of the following diagram:

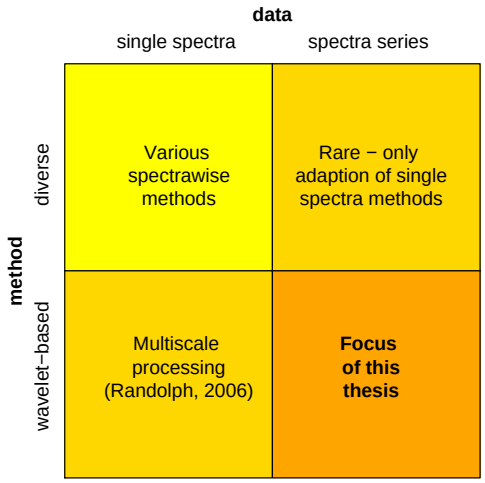

Figure 1.1: Scheme of current peak detection methodology.

The illustration shows how the research arrived at the focus area investigated in this thesis. Peak detection is already well researched in single spectra, although even for this field no reliable method exists. A promising new approach, however, was published by Randolph and Yasui (2006), based on multi resolution analysis by means of wavelets, allowing the detection of shoulder peaks that do not possess independent maxima. Methods for three-dimensional spectra series rarely exist, and the few that have been introduced only connect the results of algorithms developed for single spectra in an additional step to adapt for two-dimensional separations.

On the other hand, the idea of the peak detection method developed in this work, aims to directly grasp peaks in the three-dimensional data structure of devices, for example, coupled with chromatographic methods to benefit from the newly gained separation of signal peaks. This was realised by an enhancement of the wavelet-based method in the application with two-dimensional wavelet transforms to create a new, powerful peak detection method allowing the detection of peaks hidden in the depth of noisy data structures, or those which are poorly separated or covered by other peaks.

1.3 Achievements and contributions

The achievements accomplished in this work can be subdivided into two main sections: the preprocessing of raw spectra series for better processability, and the actual peak detection, which was linked closely to the application of the introduced methods in continuative analyses.

The following four main issues in the area of **preprocessing** were investigated:

- A better comparability of spectra in the IMS dimension was achieved by the development of a reproducible version of the (inverse) reduced mobility. The method is now used as a standard at the ISAS - Institute for Analytical Sciences, Dortmund.

- An improvement of the alignment of spectra series in the MCC dimension could be reached by the correction of retention time with respect to the value of the column temperature.

- An efficient data reduction and an increased signal-to-noise ratio was accomplished by a combined application of smoothing and denoising by the means of wavelets. This strategy was presented in an oral contribution at the Workshop "Recent

Progress in Wavelet Analysis and Frame Theory" (Bremen, Germany, January 2006).

- A detailing function was developed to adjust for the influence of a characteristic feature in the breath measurements, which was interfering with peak detection. This fitting of a modified lognormal function with a specially created penalty term was published together with the combined wavelet strategy for smoothing and denoising by Bader et al. (accepted in 2008).

The **peak detection** and further analyses comprised of the following items:

- The approach of 'merged peak cluster localisation' was initially developed. Together with the concurrent generation of general peak areas it provided a first method for peak characterisation and a continuative analysis, which was published in Bader et al. (2005) and Bader et al. (2006).
  The application of this procedure in the comparison of lung cancer patients with a healthy control group yielded a perfect discrimination between the groups, which was the subject of an analytical and a medical article (Baumbach et al., submitted in 2007; Westhoff et al., submitted in 2008) and was awarded with the Science Award of the German Association of Pneumology in 2006.

- The enhancement of this method to the 'growing interval merging' algorithm allowed a higher sensitivity in peak detection and a more sophisticated characterisation of peaks by ellipses. This method was presented together with the developed preprocessing strategy in two invited talks at the University of Barcelona (Barcelona, Spain, September 2007) and the 2006 Colloquium Series at the Department of Chemistry and Biochemistry, Ohio University (Athens, USA, November 2006), as well as in an oral contribution at the conference Compstat 2006 (Rome, Italy, August 2006).
  A first application of this algorithm was the analysis of data generated by pyrolysis-gaschromatography/differential mobility spectrometry (py-GC/DMS) of bacteria cultures during a research stay at the New Mexico State University, USA. Results from this project were presented in a joint poster contribution at the ASMS Conference on Mass Spectrometry (Indianapolis, USA, June 2007) and published in Prasad et al. (2007).

- Lastly, a wavelet-based peak detection method was developed allowing the detection of shoulder peaks without independent maxima, which was applied for the comparison of different forms of lung tumors. The approach is planned to be published from both the chemometric and medical point of view.

All calculations, the construction of figures, as well as the implementation of algorithms was performed in the statistical software package R (R Development Core Team, 2007).

1.4 Outline

This thesis is organised in 9 chapters, structured in 5 parts. In addition to the preliminary introduction, methods for the processing of spectra and the further analysis as well as the required theory are described and concluding remarks and proposals for potential future work are given.

Theory

Chapter 2 comprises of details concerning the IMS method, giving an idea of the physical background, the resulting data, and the connected problems.

Chapter 3 describes wavelet methods for one- and two-dimensional transforms as well as the multi resolution analysis.

Chapter 4 outlines the methods of cluster analysis, group comparisons, and discriminant analysis, which are later used for the processing of spectra series and subsequent analyses.

Spectra processing

Chapter 5 provides details on the preprocessing steps for three-dimensional data created during this thesis.

Chapter 6 introduces the developed peak detection methods and illustrates its beneficial outcome.

Data analysis

Chapter 7 enhances the processing of spectra series data for the analysis of entire studies by the creation of general peak areas and proves the applicability of the methods evolved from this work in two different applications.

Chapter 8 transfers the introduced methods to another spectrometric method, used in the analysis of bacterial measurements by py-GC/DMS.

Concluding remarks

Chapter 9 gives ideas for future work and summarises the main conclusions of this work that enabled the extraction of essential information from complex spectra series and allows the use of promising spectrometric data for various process analytical applications.

Theory

Chapter 2

Ion mobility spectrometry

Ion mobility spectrometry is a rapid, highly sensitive analytical method for the characterisation of gaseous samples with low detection limits. Whilst not generally applied for the identification of unknown compounds, it can be used in the quantification of analytes known to be involved in specific processes. Furthermore, instrumentation miniaturisation and the development of portable hand-held devices yielded a highly flexible applicability at a relatively low cost. Advantageous characteristics of this method include the operating at ambient pressure without the need for a vacuum, and the use of air as a potential carrier gas.

Originally developed for the detection of trace compounds such as gaseous pollutants in air, more than 70,000 IMS units are now in use worldwide, mostly applied to detect chemical warfare agents, explosives, or drugs, e.g. at international airports. Additionally, the use of IMS has advanced in the area of process analysis for applications such as monitoring of contamination in water, odoration of natural gas, human breath composition, and metabolites of bacteria (Baumbach, 2006).

The functioning of an IMS is based on the ionisation of gaseous analytes and the subsequent detection of the characteristic drift time of ion swarms through an electric field (Section 2.1). For complex biological samples, overlaps of signal peaks in the IMS spectra can constitute the coupling of an IMS with a gaschromatographic column to achieve an increase of separation. In this case complicated three-dimensional spectra series arise (Section 2.2) making the extraction of the essential information difficult (Section 2.3).

2.1 Instrumentation

The measurements from this investigation were generated with an ISAS custom-designed device (Fig. 2.1), coupling an IMS (Subsection 2.1.1) with a gaschromatographic column (Subsection 2.1.2).

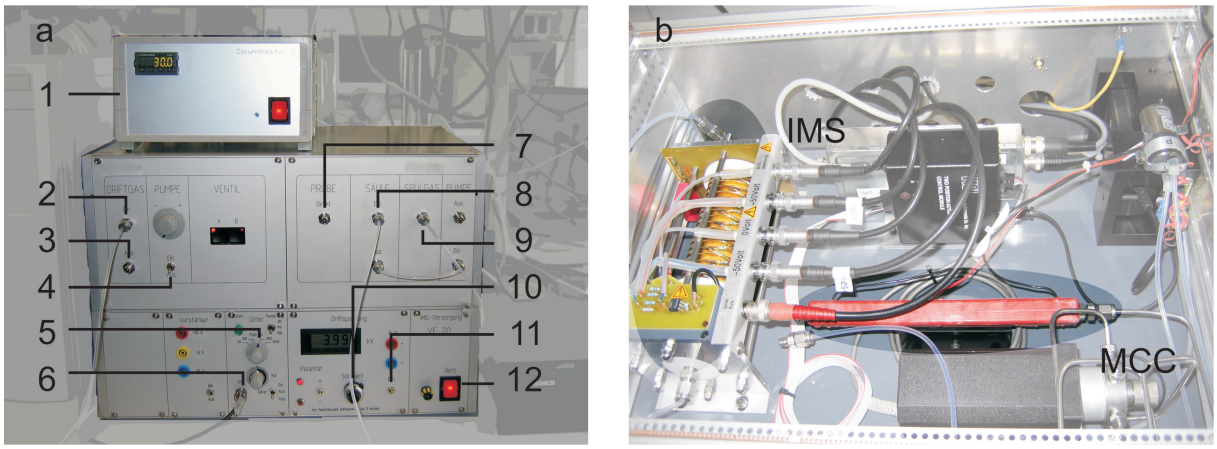

Figure 2.1: ISAS custom-made multi-capillary column/ion mobility spectrometer device: (a) exterior view with (1) temperature control for column heating, (2) drift gas inlet, (3) outlet for drift and sample gas, (4) pump switch, (5) control for shutter grid opening time, (6) measuring board connection, (7) sample inlet for use without column, (8) sample inlet into sample loop, (9) carrier gas inlet for column, (10) drift voltage control, (11) voltage switch, and (12) power switch; (b) interior view showing the multi-capillary column (MCC) and the ion mobility spectrometer (IMS)

2.1.1 Functionality of an ion mobility spectrometer

An IMS can be subdivided into two main parts: the ionisation and reaction region, where gaseous analytes are ionised; and the drift region, where their characteristic drift times are detected (Fig. 2.2, left).

A sample gas is introduced through the gas inlet into the ionisation and reaction region of the IMS, using synthetic air as carrier gas. With radioactive nickel (^{63}Ni) as the ionisation source, electrons in the form of beta rays are emitted with a maximum energy of 67 keV, therefore, no external power supply is required. This source allows measurements either in the negative or the positive mode, where only the latter will be regarded here. After ionisation, the ion swarms are released to the drift region by opening an ion shutter periodically, meanwhile, the continuous gas flow can leak across the gas outlet.

In the drift tube of length l_D, an electric field of strength $E = \frac{U_D}{l_D}$ with drift voltage U_D is established using drift rings for stabilisation. Collisions of the sample gas molecules with those of a drift gas such as air, flowing from the Faraday plate at the end of the drift tube towards the ion shutter, yield a constant drift velocity v_D. At the Faraday plate, ion swarms moving through the upstreamed aperture grid are converted into a voltage, whose intensity is measured at equidistant points in time.

Measuring the drift time x, required to pass through the drift tube, allows conclusions about the analytes to be made, as the characteristic drift velocity $v_D = \frac{l_D}{x}$ of an ion swarm is otherwise only dependent on the known drift length l_D. The velocity v_D is proportional to the electric field strength E with

$$v_D = KE,$$

where K can be formed as

$$\frac{l_D}{x} = K\frac{U_D}{l_D} \quad \Leftrightarrow \quad K = \frac{l_D^2}{U_D x}. \tag{2.1}$$

At constant measurement conditions the coefficient K, describing the mobility of ion swarms, is characteristic for the underlying analytes, explaining the terminology of ion mobility spectrometry.

Collisions of the high-energy electrons with carrier gas molecules yield so-called reactant ions. Using air as the carrier gas in the positive mode, mainly $H^+(H_2O)_n$ ions with $n \in \{1, \dots 7\}$ are formed. These result in the characteristic signal of the reactant ion peak (RIP) in the generated spectra. The overall loading of the reactant ions is dependent on the strength of the ionisation source and gives an upper limit for the number of molecules that can be ionised.

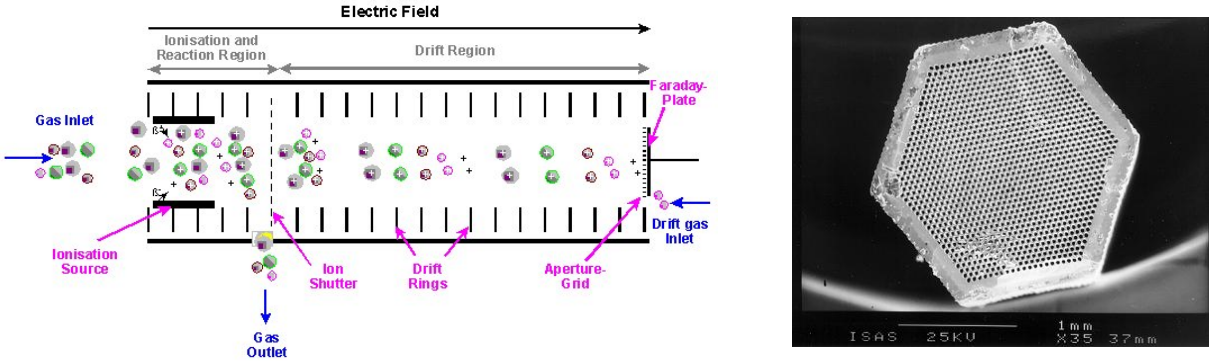

Figure 2.2: Scheme of an ion mobility spectrometer (left) and picture of a multi-capillary column (right), which allow an improved analyte peak separation in a combined set-up.

The standard reaction for molecule ionisation,

$$M + H^+(H_2O)_n \rightarrow MH^+(H_2O)_n \rightarrow MH^+(H_2O)_{n-m} + mH_2O,$$

where $n \in \{1, \ldots 7\}$ and $m \in \{0, 1\}$, can be divided into three main stages for simplification.

Firstly, the initial reaction of molecules, M, with the reactant ions takes place, resulting in the formation of an intermediate product, $MH^+(H_2O)_n$, which can be further broken down into a monomer product, $MH^+(H_2O)_{n-m}$, and water molecules, H_2O. Importantly, the reaction can only proceed if a number of criteria are met, e.g. that the molecules M possess a higher proton affinity than the reactant ions.

If the concentration of M is high, the monomer product ions, $MH^+(H_2O)_{n-m}$, can react with one another and produce proton bounded dimers, $M_2H^+(H_2O)_{n-m}$, following the reaction equation

$$MH^+(H_2O)_n + M \rightarrow M_2H^+(H_2O)_{n-m} + mH_2O.$$

The presence of monomers and dimers results in additional peaks in the ion mobility spectra, and in the ideal situation, a total separation of different analytes can be obtained when each takes a different time to pass the drift tube (Eiceman and Karpas, 2005).

2.1.2 Preseparation

For complex biological samples, an overlap of different analyte peaks can hinder the identification and quantification of sample constituents with an IMS. Coupling with a gaschromatographic column can, therefore, increase separation and, comparatively, a multi-capillary column (MCC) approved to be especially suitable for the investigation of human breath, as it separates sample humidity at the very beginning of the analysis (Ruzsanyi et al., 2005). In this, the number of interfering peaks, caused when high relative humidity of the carrier gas leads to the formation of ion clusters with water molecules, is diminished and the analysis of exhaled air, possessing a relative humidity of 100 %, can be improved.

Consisting of a single glass tube with 1000 parallel 40 μm capillaries (Fig. 2.2, right), an MCC provides short retention times with a high degree of separation and resolution. Additionally, it retains a high efficiency over a wide range of organic compounds, can work with any reasonable sample size, and appears to be particularly promising in the analysis of trace amounts of compounds. Thus, it substantially increases sensitivity and reduces analysis time.

2.2 Data

The data resulting from an IMS are spectra $\mathbf{s}$ of the form

$$\mathbf{s} = (z_1, \ldots, z_{n_D})^T ,$$

where the values z_i are the signal intensities of ions arriving at the end of the IMS drift region at equidistant drift times x_i, $i = 1, \ldots, n_D$ (Fig. 2.3 a). Each spectrum is generated by averaging several scans resulting in a higher signal-to-noise ratio.

When coupling an MCC to the IMS, the additional dimension of retention time when analytes pass from the column to the IMS is obtained, whose values are denoted by y_j, $j = 1, \ldots, n_R$. Then series of spectra are generated that can be represented in a matrix

$$\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_{n_R}) = (z_{ij})_{i=1,\ldots,n_D;\ j=1,\ldots,n_R} ,$$

consisting of intensity values z_{ij}. The resulting data can be displayed in a heatmap, where the axes define a grid of equidistant drift and retention times, and the signal intensity for each position is encoded by a colour scheme (Fig. 2.3 b).

The interesting features inherent in those data are signal peaks appearing as oval spots in the heatmap. The high peak apparent in the single spectra, but also in the heatmap of a spectra series as a bar across all retention times at a drift time of about 17 ms, is the afore-mentioned RIP. It can be seen as the reservoir of ions, as its height varies inversely to

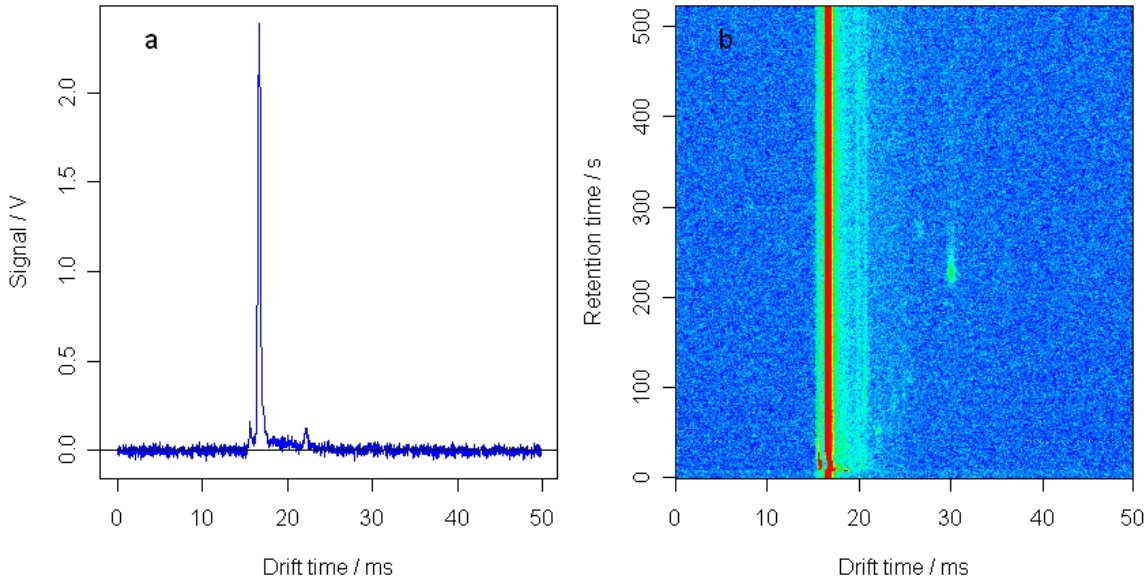

Figure 2.3: Graph of (a) a single spectrum and (b) a spectra series in a heatmap: local elevations and oval spots represent analyte peaks, respectively.

Table 2.1: Relevant measurement parameters of the ISAS custom-made MCC/IMS device

IMS		**Preseparation**	
Ionisation source	^{63}Ni (radioactive)	Column type	MCC, OV5
Polarity	positive	Column length	20 cm
Drift voltage	4 kV	Column temperature	30 °C
Drift tube length	12 cm	Carrier gas	synthetic air
Field strength	326 V/cm	Carrier gas flow	150 mL/min
Grid opening time	0.3 ms	Carrier gas humidity	0
Grid delay	100 ms	**Environment**	
Drift gas	synthetic air	Temperature	ambient
Drift gas flow	100 mL/min	Pressure	ambient
Drift gas humidity	0	**Software**	
Preamplifier	10^{10} V/A	Average	10
Sampling		Number of samples	2000
Sampling	pump	Frequency	40000
Sampling duration	10 s	Gain	1
Sample flow	350 mL/min	Data file suffices	0 to 500

the appearance of other peaks in the spectra. The information based on sample analytes, however, lies in the small peaks in the spectra part behind the RIP. This measurement part also contains the tailing of the RIP, which is responsible for the slow decrease of the signal intensity back to the baseline, causing varying heights for peaks in different parts of the drift time axis. The observed peak tailing is due to variations in the ion velocity from scan to scan which are caused by random ion-molecule reactions occurring in the drift tube and further compounded by the signal averaging process, as the contributions from each individual scan are recorded in the final spectrum.

The parameters of the instrumentation and the software, used in this work (Table 2.1), were kept as constant as possible, aiming the comparability of different measurements.

2.3 Problem

Measurements generated with an MCC/IMS result in large spectra series, which for the monitoring of human breath consist of more than one million data points. As most measurement parts consist of pure noise and even the interesting analyte peaks contain dozens

of data points, the high dimensionality involves an undesirable degree of redundancy. In addition, the data variation in all directions of its three-dimensional structure is problematic, as the height as well as the drift and retention time position of appearing peaks underlies limited reproducibility. The question is, therefore, not only what the essential information inherent in these data is, but also what the variables comparable between the measurements are.

Since peaks correspond to sample analytes, they are the relevant information, making an effective peak identification and characterisation necessary to give a base for further processing yielding potential results from the generated data. A reasonable peak detection procedure reduces data efficiently to meaningful peak characteristics ensuring little information loss, and is indispensable if questions with large sample sizes are investigated, as a manual analysis is time-consuming and to a certain amount subjective.

Although several peak detection algorithms exist, most are optimised for finding sharp peaks in mass spectrometry data, and even amongst these no method yields truely reliable results. Currently, few algorithms have been implemented for three-dimensional spectra series and those that have simply connect the results for single spectra in an additional step to adapt to two-dimensional separations. The idea of the peak detection method developed in this work on the other hand, aims to directly grasp the three-dimensional structure of the data to benefit from the newly gained separation of signal peaks.

Besides the general challenges common to peak detection methods for all three-dimensional spectra series, there are some additional problems to deal with for IMS data. Due to a small signal-to-noise ratio, peaks can sometimes not be distinguished from noise via signal intensity alone and despite preseparation, can still lie very close together resulting in shoulder peaks without a maximum; or for more severe overlaps, small peaks are even covered by larger ones. Another interfering feature of IMS data is the tailing of the RIP, making a simple threshold for separation between noise and peak areas difficult. These circumstances have to be respected to allow for an efficient IMS data preprocessing and peak identification.

Chapter 3

Wavelet transform

The wavelet transform is a tool for signal decomposition, allowing to locate features of a signal in frequency and time simultaneously. Therefore, it can be seen as an enhancement of the well-established Fourier transform.

The Fourier transform gives global frequency information without any localisation properties, since the underlying functions are big waves swinging from infinity to infinity (Fig. 3.1, top). To overcome this limitation, the windowed Fourier transform was developed,

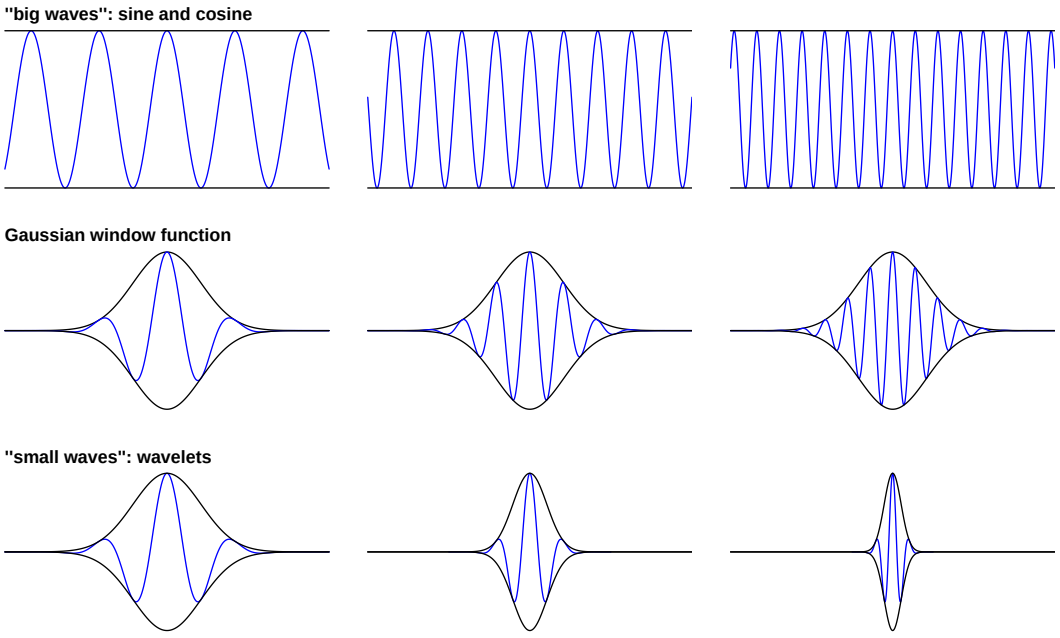

Figure 3.1: Scheme of the relationship of Fourier (top), windowed Fourier (middle), and wavelet transform (bottom): with the enhancement to wavelet transform the localisation properties are improved, while all frequencies can be considered accurately.

using a window function such as the Gaussian function to achieve localisation information in the decomposition (Fig. 3.1, middle). While the window size is fixed, the number of oscillations of a sine or cosine function is varied. This leads to the trade-off between precise localisation information, available only for high frequencies when a small window size is chosen, and a higher variety of considerable frequencies with, therefore, only vague localisation for larger windows. Keeping the number of oscillations fixed and varying the window size, the wavelet transform circumvents this problem and screens for details with high frequencies in small windows and the trend consisting of low frequencies in large windows (Fig. 3.1, bottom).

The functions used for these decompositions move essentially in a limited time interval, and can thus be described as small waves, leading to the denomination wavelets (Section 3.1). They are the base of various methods with different properties such as the discrete wavelet transform (DWT) (Section 3.2), the maximum-overlap discrete wavelet transform (MODWT) (Section 3.3), and the multi resolution analysis (MRA) (Section 3.4). The findings and the structure of the description of these methods are based on the presentation in Percival and Walden (2000); the notation, however, was conformed to the overall format of this work. In addition, the methodology can be extended to a two-dimensional variant of the wavelet transform, which is discussed here with particular respect to the MODWT and the MRA (Section 3.5).

3.1 Fundamentals

To introduce the wavelet transform, it is important to define the properties of the underlying wavelet functions as well as the resulting wavelet bases and coefficients.

A wavelet function $\psi(\ )$ is assumed to be a real-valued function defined over the real axis $(-\infty, \infty)$ with an integral of zero,

$$\int_{-\infty}^{\infty} \psi(u)du = 0,$$

and the square of $\psi(\ )$ integrating to unity,

$$\int_{-\infty}^{\infty} \psi^2(u)du = 1.$$

These requirements mean the function $\psi(\ )$ must possess some nonzero activity, limited to a relatively small interval, and balanced between negative and positive contributions.

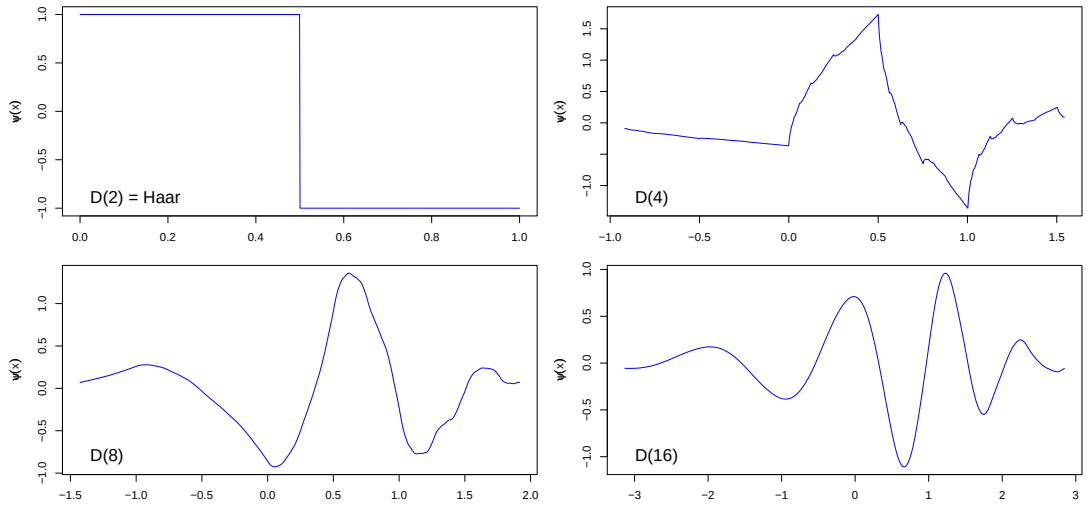

Figure 3.2: Four wavelet functions of the Daubechies family: wavelet functions are often irregular and asymmetric.

In contrast to the sine and cosine functions used for the Fourier transform, wavelet functions are often irregular and asymmetric (Fig. 3.2). There are also complex wavelet functions such as the Morlet wavelet, and depending on definition ones without compact support, but these are not considered here.

Coming from a mother wavelet function $\psi(\ )$, which fulfills the afore-mentioned conditions, a wavelet base can be formed by scaling and translating $\psi(\ )$ to

$$\psi_{\lambda,t}(u) \equiv \frac{1}{\sqrt{\lambda}}\psi\left(\frac{u-t}{\lambda}\right)$$

with $\lambda > 0$ and $-\infty < t < \infty$.

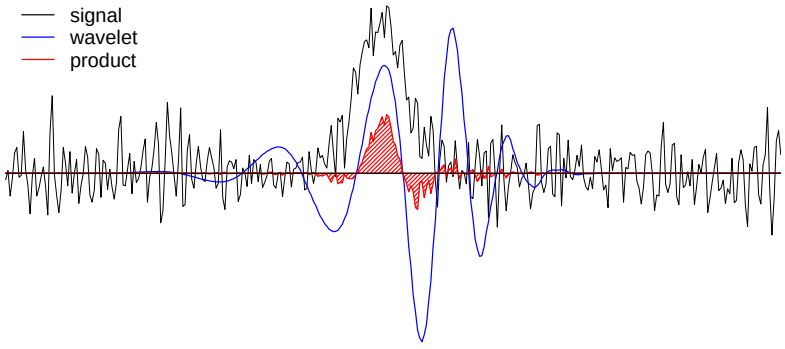

Figure 3.3: Scheme of the calculation of a coefficient of the continuous wavelet transform: The integral of the product of a signal with one of the functions of the wavelet base yields one coefficient of the resulting decomposition.

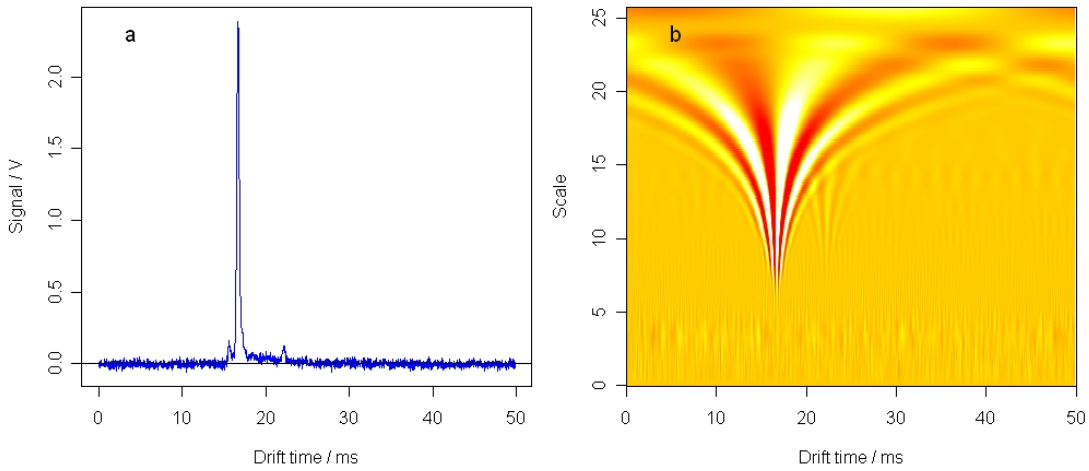

Figure 3.4: Plot of (a) a single spectrum and (b) the corresponding CWT decomposition: the coefficients are shown on a grid of time and scale with their height encoded by colour – high values are displayed in yellow and white, medium and low values in orange and red.

The integral of the product of a signal $\mathbf{s}$ and each of the members of the constructed wavelet base yields a collection of variables $\{W(\lambda, t) : \lambda > 0, -\infty < t < \infty\}$ with

$$W(\lambda, t) \equiv \int_{-\infty}^{\infty} \psi_{\lambda,t}(u)s(u)du,$$

defining the wavelet coefficients of the continuous wavelet transform (CWT). The coefficient $W(\lambda, t)$ can be interpreted as being proportional to a difference of adjacent weighted averages of scale λ around time t (Fig. 3.3).

As the wavelet base, used for the calculation of the coefficients, was created by scaling and translating, the CWT decomposition of a single spectrum results in a two-dimensional structure of coefficients, dependent on time and scale (Fig. 3.4). The position and height of a coefficient thus allows conclusions to be drawn considering changes at a particular location of the signal that are of the frequency corresponding to a specific scale.

The CWT of a signal preserves all information, also allowing for its recovery, however, because of its two-dimensional nature an image processing problem occurs as a lot of redundancy is included in the resulting decomposition of the signal.

3.2 Discrete wavelet transform

To obtain a representation containing no redundancy, the CWT coefficients $W(\lambda, t)$ can be downsampled by the dyadic factor 2. This yields the method of DWT, whose sparse

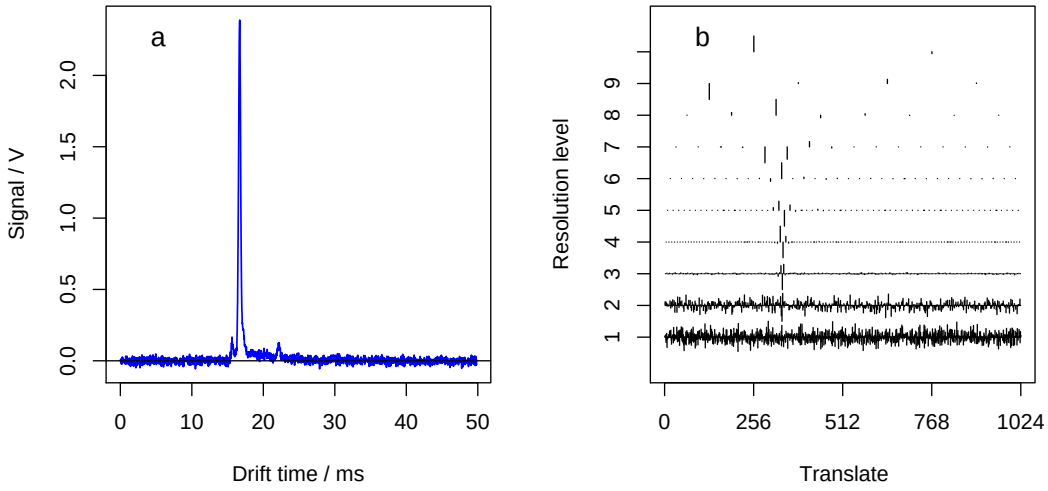

Figure 3.5: Plot of (a) a single spectrum and (b) the corresponding DWT decomposition: the coefficients are shown on a grid of time and scale, where – starting from a virtual zero line for each resolution level – each coefficient is given by a black line, indicating if a high or low, negative or positive contribution was determined.

nature qualifies it for fast algorithms, which can be amplified using a pyramid algorithm (Subsection 3.2.1), and optionally by limiting the calculations to a partial decomposition (Subsection 3.2.2).

As time and scale are not scanned continuously any longer, the subset of scales λ and the corresponding time values have to be specified. Assuming the signal $\mathbf{s}$ has length $N = 2^J$, the scales λ of the downsampled transform are chosen as $\tau_j \equiv 2^{j-1}$, $j = 1, \ldots, J$, resulting in $N_j \equiv N/(2\tau_j)$ DWT coefficients associated with changes on scale τ_j. Within a dyadic scale, the times t corresponding to these coefficients are separated by multiples of 2^j and set to $(2n+1)2^{j-1} - \frac{1}{2}$, $n = 0, 1, \ldots, N_j - 1$. Additionally, there is one scaling coefficient linked to an average of all the data.

The decomposition resulting from a DWT of a single spectrum (Fig. 3.5 a) can be illustrated as shown in Fig. 3.5 b: the lowest line displays the resolution level of the highest frequency and shows the finest details of the original signal; the topmost line corresponds to the lowest frequency and illustrates tendencies over a broad range of the spectrum. This representation is not only less redundant in the dimension of scales, but also contains a decreasing amount of coefficients on the higher scales.

Representing the DWT coefficients by $\{W_n : n = 0, \ldots, N-1\}$ for formalisation of the method, the transform can be written as

$$\mathbf{w} = \mathbf{W}\mathbf{s},$$

where $\mathbf{w}$ is a vector of length $N = 2^J$ whose nth element is the nth DWT coefficient W_n, and $\mathbf{W}$ is an $N \times N$ matrix defining the DWT and satisfying $\mathbf{W}^T\mathbf{W} = \mathbf{I}_N$.

The matrix $\mathbf{W}$ can be constructed by any wavelet base satisfying the properties of summation to zero and orthonormality, where the orthogonality allows the calculation of wavelet coefficients by an inner product and ensures a representation without redundancy. As the orthonormality condition does not yield a unique wavelet base, additional conditions such as 'extremal phase' or 'least asymmetric' must be demanded to obtain uniqueness (Daubechies, 1992).

The rows of the matrix $\mathbf{W}$ can be grouped into $J+1$ submatrices, each corresponding to a scale τ_j, which results in a partitioning of the vector $\mathbf{w}$ of DWT coefficients:

$$\mathbf{Ws} = \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \\ \vdots \\ \mathbf{W}_J \\ \mathbf{V}_J \end{bmatrix} \mathbf{s} = \begin{bmatrix} \mathbf{W}_1\mathbf{s} \\ \mathbf{W}_2\mathbf{s} \\ \vdots \\ \mathbf{W}_J\mathbf{s} \\ \mathbf{V}_J\mathbf{s} \end{bmatrix} = \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_J \\ \mathbf{v}_J \end{bmatrix} = \mathbf{w},$$

where $\mathbf{W}_j$ has dimension $N_j \times N$; $\mathbf{V}_J$ is $1 \times N$; $\mathbf{w}_j$ is a vector of length N_j; and $\mathbf{v}_J$ contains the last element of $\mathbf{w}$. Within the matrix $\mathbf{W}_j$ producing the wavelet coefficients for the

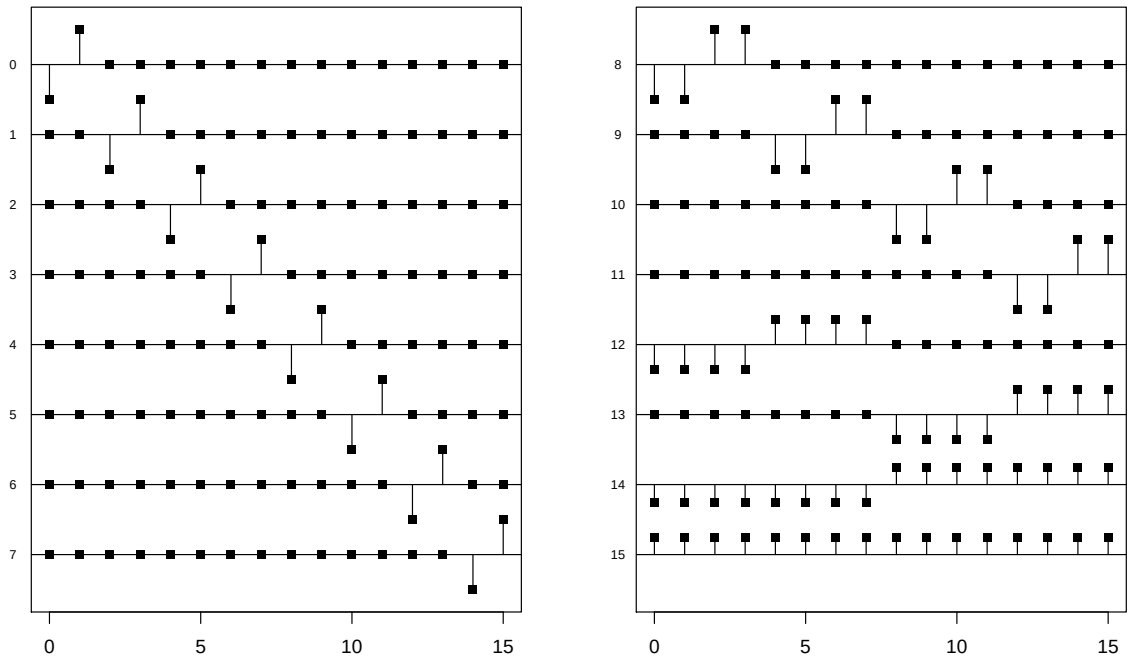

Figure 3.6: Row vectors $\mathbf{W}_{n\bullet}^T$ of the discrete wavelet transform matrix $\mathbf{W}$ based on the Haar wavelet for $N = 16$ with $n = 0$ to 7 (left) and $n = 8$ to 15 (right).

particular scale τ_j, the rows are circularly shifted versions of each other with the amount of the shift between adjacent rows being $2\tau_j = 2^j$.

Fig. 3.6 shows the row vectors of the matrix $\mathbf{W}$ for the Haar wavelet with $N = 16$, as its simple structure allows an easy understanding of the underlying scheme, how wavelet coefficients W_n are associated with particular scales and sets of times. The row vectors 0 to 7, 8 to 11, 12 to 13, and 14 constitute the matrices $\mathbf{W}_1$, $\mathbf{W}_2$, $\mathbf{W}_3$, and $\mathbf{W}_4$ respectively, while the row vector 15 provides $\mathbf{V}_4$.

Coming from such a wavelet decomposition the original signal $\mathbf{s}$ can be perfectly reconstructed without an error by

$$\mathbf{s} = \mathbf{W}^T\mathbf{w} = \sum_{j=1}^{J} \mathbf{W}_j^T\mathbf{w}_j + \mathbf{V}_J^T\mathbf{v}_J.$$

3.2.1 Pyramid algorithm

In practice the DWT matrix $\mathbf{W}$ is not formed explicitly, but rather $\mathbf{w}$ is computed using a pyramid algorithm that makes use of a wavelet filter and a scaling filter. Requiring only $O(N)$ multiplications, this algorithm is actually faster than the fast Fourier transform algorithm, calculated in $O(N \log_2 N)$ steps, that led to a widespread use of the discrete Fourier transform.

The first stage of the pyramid algorithm decomposes the signal $\mathbf{s} = \{z_t : t = 0, \ldots, N-1\}$ into two new series: $\{W_{1,t} : t = 0, \ldots, \frac{N}{2} - 1\}$ constituting the first half of the vector of wavelet coefficients $\mathbf{w}$; and $\{V_{1,t} : t = 0, \ldots, \frac{N}{2} - 1\}$ providing a basis for obtaining the remaining wavelet coefficients at successive stages of the pyramid algorithm.

Wavelet filter

To describe this algorithm, linear filtering operations will be discussed here, beginning with a real-valued wavelet filter $\{h_l : l = 0, \ldots, L-1\}$ with L giving an even length for the filter, meaning $h_0 \neq 0$, $h_{L-1} \neq 0$, and $h_l = 0$ for $l < 0$ and $l \geq L$. A wavelet filter is, therefore, an infinite sequence with at most L nonzero values, required to satisfy the following three basic properties

$$\sum_{l=0}^{L-1} h_l = 0, \quad \sum_{l=0}^{L-1} h_l^2 = 1, \quad \text{and} \quad \sum_{l=0}^{L-1} h_l h_{l+2n} = \sum_{-\infty}^{\infty} h_l h_{l+2n} = 0$$

for all nonzero integers n.

To obtain the $N/2$ wavelet coefficients for unit scale, the signal $\mathbf{s} = \{z_t\}$ is circularly filtered with $\{h_l\}$:

$$\sqrt{2}\tilde{W}_{1,t} \equiv \sum_{l=0}^{L-1} h_l z_{t-l \bmod N}, \quad t = 0, \ldots, N-1,$$

where $N = 2^J$ for some positive integer J.[1] The sequence $\{\sqrt{2}\tilde{W}_{1,t}\}$ is then downsampled to the $N/2$ values with odd indices to define the wavelet coefficients

$$W_{1,t} \equiv \sqrt{2}\tilde{W}_{1,2t+1}, \quad t = 0, \ldots, \frac{N}{2} - 1.$$

The first of the two subscripts on $W_{1,t}$ and $\tilde{W}_{1,t}$ states the scale $\tau_j = 2^{j-1}$ associated with the $N/2$ wavelet coefficients, with $j = 1$ as the index for the unit scale here. The square root of two is included to preserve energy following downsampling.

To connect the definition of $\{W_{1,t}\}$ to the matrix formulation, the coefficients can be derived directly as

$$W_{1,t} = \sum_{l=0}^{L-1} h_l z_{2t+1-l \bmod N} = \sum_{l=0}^{N-1} h_l^\circ z_{2t+1-l \bmod N}, \quad t = 0, \ldots, \frac{N}{2} - 1, \tag{3.1}$$

where $\{h_l^\circ\}$ is $\{h_l\}$ periodised to length N. These coefficients constitute the first $N/2$ coefficients of $\mathbf{w} = \mathbf{W}\mathbf{s}$, i.e., the elements of the subvector $\mathbf{w}_1 = \mathbf{W}_1\mathbf{s}$, where $\mathbf{W}_1$ is the $\frac{N}{2} \times N$ matrix containing the first $N/2$ rows of $\mathbf{W}$. The first row of $\mathbf{W}_1$ is given by

$$\mathbf{w}_{0\bullet}^T = \left[h_1^\circ, h_0^\circ, h_{N-1}^\circ, h_{N-2}^\circ, \ldots, h_2^\circ\right],$$

while the remaining $\frac{N}{2} - 1$ rows can be expressed as versions of $\mathbf{w}_{0\bullet}^T$ circularly shifted by the amount of 2.

Scaling filter

In preparation for forming the last $N/2$ rows of $\mathbf{W}$ via the pyramid algorithm, a second filter is required to construct the $\frac{N}{2} \times N$ matrix $\mathbf{V}_1$. Given the wavelet filter $\{h_l\}$, this so-called scaling filter is defined as

$$g_l \equiv (-1)^{l+1} h_{L-1-l},$$

[1] The modulo function mod is defined by

$$j \bmod N \equiv \begin{cases} j & , \text{ if } 0 \le j \le N-1 \\ j + nN \quad \text{with} \quad 0 \le j + nN \le N-1 & , \text{ else} \end{cases},$$

and was used to ensure indices lying in the defined range. For terms such as $2t+1-l \bmod N$, the function refers to the whole expression $2t + 1 - l$.

fulfilling the properties

$$\sum_{l=0}^{L-1} g_l = \sqrt{2}, \quad \sum_{l=0}^{L-1} g_l^2 = 1, \quad \sum_{l=-\infty}^{\infty} g_l g_{l+2n} = 0, \text{ and } \sum_{l=-\infty}^{\infty} g_l h_{l+2n'} = 0$$

for all nonzero integers n and all integers n'. The first condition can also be substituted by $\sum_{l=0}^{L-1} g_l = -\sqrt{2}$, but the convention chosen here simplifies the interpretation of scaling coefficients as being localised weighted averages.

As it is possible for the wavelet coefficients $W_{1,t}$, the $N/2$ first level scaling coefficients $V_{1,t}$ can also be calculated directly as

$$V_{1,t} = \sum_{l=0}^{L-1} g_l z_{2t+1-l \bmod N} = \sum_{l=0}^{N-1} g_l^\circ z_{2t+1-l \bmod N}, \qquad t = 0, \ldots, \frac{N}{2} - 1,$$

where $\{g_l^\circ\}$ is $\{g_l\}$ periodised to length N. These coefficients form $\mathbf{v}_1 = \mathbf{V}_1 \mathbf{s}$, where $\mathbf{V}_1$ is the $\frac{N}{2} \times N$ matrix whose rows are given by circularly shifted versions of

$$\mathbf{v}_{0\bullet}^T = \left[g_1^\circ, g_0^\circ, g_{N-1}^\circ, g_{N-2}^\circ, \ldots, g_2^\circ\right]$$

by the amount of 2.

As the rows of the matrix $\mathbf{V}_1$ constitute a set of orthonormal vectors, and because the scaling filter is orthogonal to the wavelet filter and all its even shifts, $\mathbf{V}_1$ and $\mathbf{W}_1$ are orthogonal. Thus the $N \times N$ matrix

$$\mathbf{P}_1 \equiv \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{V}_1 \end{bmatrix}$$

is orthonormal, allowing for synthesis of the signal $\mathbf{s}$ by

$$\mathbf{s} = \mathbf{P}_1^T \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{v}_1 \end{bmatrix} = \mathbf{W}_1^T \mathbf{w}_1 + \mathbf{V}_1^T \mathbf{v}_1.$$

Iterative stages of the pyramid algorithm

In general, the rows of $\mathbf{V}_1$ and the last $N/2$ rows of $\mathbf{W}$ are not identical, but $\mathbf{V}_1$ can be manipulated on the $J-1$ subsequent stages of the pyramid algorithm to obtain these rows.

For $j = 2, \ldots, J$, the jth stage transforms the vector $\mathbf{v}_{j-1}$ of length $N/2^{j-1}$ into the vectors $\mathbf{w}_j$ and $\mathbf{v}_j$, each of length $N/2^j$. Thereby $\mathbf{v}_{j-1}$ is treated in exactly the same way as the signal $\mathbf{s}$ on the first stage: the elements are filtered separately with $\{h_l\}$ and $\{g_l\}$,

and the filter outputs are downsampled to form the vectors $\mathbf{w}_j$ of wavelet coefficients and $\mathbf{v}_j$ of scaling coefficients for level j respectively. At the end of the Jth stage, the DWT coefficient vector $\mathbf{w}$ can be formed by joining the $J+1$ vectors $\mathbf{w}_1, \ldots, \mathbf{w}_J$ and $\mathbf{v}_J$.

More precisely, these stages are defined by $\mathbf{W}_j = \mathbf{B}_j \mathbf{V}_{j-1}$ and $\mathbf{V}_j = \mathbf{A}_j \mathbf{V}_{j-1}$ using the $N_j \times N_{j-1}$ matrices $\mathbf{B}_j$ and $\mathbf{A}_j$ containing the wavelet and scaling filters $\{h_l\}$ and $\{g_l\}$ periodised to length N_{j-1} and circularly shifted by the amount of 2, with $\mathbf{V}_0 = \mathbf{I}_N$ and thus $\mathbf{B}_1 = \mathbf{W}_1$ and $\mathbf{A}_1 = \mathbf{V}_1$.

Based on the $N_{j-1} \times N_{j-1}$ matrix

$$\mathbf{P}_j \equiv \begin{bmatrix} \mathbf{B}_j \\ \mathbf{A}_j \end{bmatrix},$$

the vector $\mathbf{v}_{j-1}$ can be recovered by

$$\mathbf{v}_{j-1} = \mathbf{P}_j^T \begin{bmatrix} \mathbf{w}_j \\ \mathbf{v}_j \end{bmatrix} = [\mathbf{B}_j^T \mathbf{A}_j^T] \begin{bmatrix} \mathbf{w}_j \\ \mathbf{v}_j \end{bmatrix} = \mathbf{B}_j^T \mathbf{w}_j + \mathbf{A}_j^T \mathbf{v}_j. \tag{3.2}$$

Recursive application yields the back-transform of the original signal by

$$\mathbf{s} = \mathbf{B}_1^T \mathbf{w}_1 + \mathbf{A}_1^T \mathbf{B}_2^T \mathbf{w}_2 + \cdots + \mathbf{A}_1^T \cdots \mathbf{A}_{j-1}^T \mathbf{B}_j^T \mathbf{w}_j + \mathbf{A}_1^T \cdots \mathbf{A}_{j-1}^T \mathbf{A}_j^T \mathbf{v}_j$$

on each of the $j = 1, \ldots, J$ stages of the algorithm.

3.2.2 Partial discrete wavelet transform

Stopping the pyramid algorithm after $J_0 < J$ repetitions leads to a level J_0 partial DWT of $\mathbf{s}$, whose coefficients are given by

$$\begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \\ \vdots \\ \mathbf{w}_j \\ \vdots \\ \mathbf{w}_{J_0} \\ \mathbf{v}_{J_0} \end{bmatrix} = \begin{bmatrix} \mathbf{W}_1 \\ \mathbf{W}_2 \\ \vdots \\ \mathbf{W}_j \\ \vdots \\ \mathbf{W}_{J_0} \\ \mathbf{V}_{J_0} \end{bmatrix} \mathbf{s} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \mathbf{A}_1 \\ \vdots \\ \mathbf{B}_j \mathbf{A}_{j-1} \cdots \mathbf{A}_1 \\ \vdots \\ \mathbf{B}_{J_0} \mathbf{A}_{J_0-1} \cdots \mathbf{A}_1 \\ \mathbf{A}_{J_0} \mathbf{A}_{J_0-1} \cdots \mathbf{A}_1 \end{bmatrix} \mathbf{s},$$

where $\mathbf{w}_j$, $j = 1, \ldots, J_0$, are subvectors of the DWT coefficient vector $\mathbf{w}$. The final subvector $\mathbf{v}_{J_0}$ of $N/2^{J_0}$ scaling coefficients replaces the last $N/2^{J_0}$ coefficients of $\mathbf{w}$, representing averages over the scale 2^{J_0} and comprising the large scale components in $\mathbf{s}$.

The partial DWT thus offers the flexibility to specify a scale beyond which a wavelet analysis into individual large scales is not of interest. In addition, the restriction of a dyadic signal length $N = 2^J$ can be eased to the condition of N being an integer multiple of 2^{J_0}.

3.3 Maximum overlap discrete wavelet transform

The result of the DWT can be strongly dependent on the starting point of the analysis, as a little shift in a spectrum can produce very different coefficients. This behaviour is not favourable, especially as spectra data are sometimes rarely aligned.

The modified method of MODWT, however, yields the property of translation invariance and possesses the additional advantage of being well defined for any sample size N. It also decomposes the signal on dyadic scales, but in contrast to the DWT it is not downsampled in time.

The (partial) MODWT of level J_0 is a highly redundant nonorthogonal transform defined by the $N \times N$ submatrices $\tilde{\mathbf{W}}_j$ and $\tilde{\mathbf{V}}_{J_0}$ yielding $\tilde{\mathbf{w}}_j = \tilde{\mathbf{W}}_j\mathbf{s}$ and $\tilde{\mathbf{v}}_{J_0} = \tilde{\mathbf{V}}_{J_0}\mathbf{s}$, $j = 1, \ldots, J_0$. The resulting column vectors $\tilde{\mathbf{w}}_1, \tilde{\mathbf{w}}_2, \ldots, \tilde{\mathbf{w}}_{J_0}$ and $\tilde{\mathbf{v}}_{J_0}$, each of dimension N and based on $\{W(\lambda, t) : \lambda = 2^{j-1}, j = 1, \ldots, J_0; -\infty < t < \infty\}$, contain the MODWT wavelet coefficients associated with changes in the signal $\mathbf{s}$ on a scale of $\tau_j = 2^{j-1}$, and the MODWT scaling coefficients representing averages at scales 2^{J_0} and higher respectively.

Practical considerations

The coefficients of a MODWT are continuous over time and can, therefore, be easily compared with the underlying spectrum. Since the noise contributions of the original signal are filtered out on the highest frequency levels of the decomposition, the other levels can give indications about hidden peaks, such as e.g. in Fig. 3.7 for level 5, where the coefficients show an additional potential peak in the end part of the RIP, whose position, however, varies between the considered scales.

Furthermore, an increasing shift of the different levels makes the location of features even harder. Although in this work the results of the MODWT were considered after back-transform of before manipulated coefficients to the time domain, it should, therefore, be remarked that an alignment with the original signal is necessary for comparison otherwise. As each of the coefficient vectors $\tilde{\mathbf{w}}_j$ and $\tilde{\mathbf{v}}_{J_0}$, however, has the same

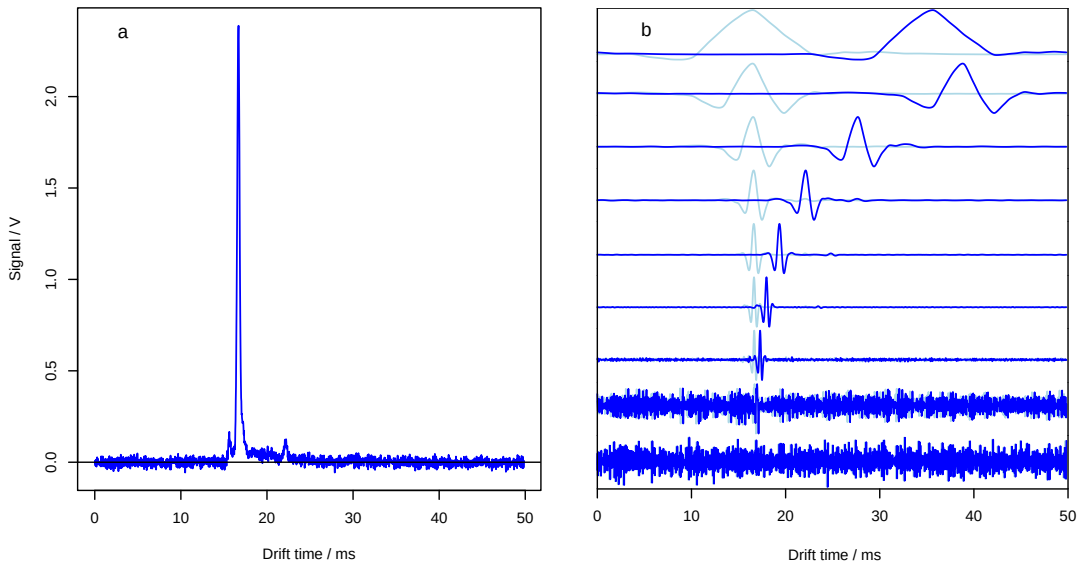

Figure 3.7: Plot of (a) a single spectrum and (b) the corresponding MODWT decomposition without (dark blue) and with (light blue) alignment to original signal: for each dyadic scale, the coefficients are illustrated by a continuous line over time, that was furthermore aligned with the original spectrum for simplified interpretation of peak indications, such as here the small hint for a potential hidden peak in the later part of the RIP e.g. on level 5.

number of elements as the spectrum $\mathbf{s}$, the coefficient vector can be easily aligned by advancing it, i.e., circularly shifting it to the left by absolute values $\nu_j^{(H)}$ and $\nu_{J_0}^{(G)}$ (Percival and Walden, 2000, p. 179 f).

Pyramid algorithm

In practice the MODWT coefficient vectors are generated via an efficient pyramid algorithm analogous to the DWT, using the wavelet filter $\{\tilde{h}_l\}$ and the scaling filter $\{\tilde{g}_l\}$ defined by $\tilde{h}_l \equiv h_l/\sqrt{2}$ and $\tilde{g}_l \equiv g_l/\sqrt{2}$.

The MODWT wavelet and scaling coefficients of the first level are then given by

$$\tilde{W}_{1,t} \equiv \sum_{l=0}^{L-1} \tilde{h}_l z_{t-l \bmod N} \quad \text{and} \quad \tilde{V}_{1,t} \equiv \sum_{l=0}^{L-1} \tilde{g}_l z_{t-l \bmod N}$$

with $t = 0, \ldots, N-1$. Equivalently, $\{\tilde{W}_{1,t}\}$ and $\{\tilde{V}_{1,t}\}$ can be regarded as the result of circularly filtering $\{z_t\}$ with $\{\tilde{h}_l^\circ\}$ and $\{\tilde{g}_l^\circ\}$, which are versions of the wavelet and the scaling filter periodised to length N. These filters also give the rows of the $N \times N$ matrix $\tilde{\mathbf{W}}_1$ and $\tilde{\mathbf{V}}_1$ in the matrix notation, if shifted circularly by the amount of 1. On the subsequent stages, the vector $\tilde{\mathbf{v}}_{j-1}$ is treated as the signal $\mathbf{s}$ on the first stage,

where in contrast to the DWT the resulting vectors $\tilde{\mathbf{w}}_j$ and $\tilde{\mathbf{v}}_j$ are all of length N for $j = 1, \ldots, J_0 \leq J$.

The original signal can be reconstructed from its MODWT via

$$\begin{aligned} \mathbf{s} &= \sum_{j=1}^{J_0} \tilde{\mathbf{W}}_j^T \tilde{\mathbf{w}}_j + \tilde{\mathbf{V}}_{J_0}^T \tilde{\mathbf{v}}_{J_0} \\ &= \tilde{\mathbf{B}}_1^T \tilde{\mathbf{w}}_1 + \tilde{\mathbf{A}}_1^T \tilde{\mathbf{B}}_2^T \tilde{\mathbf{w}}_2 + \cdots + \tilde{\mathbf{A}}_1^T \cdots \tilde{\mathbf{A}}_{j-1}^T \tilde{\mathbf{B}}_j^T \tilde{\mathbf{w}}_j + \tilde{\mathbf{A}}_1^T \cdots \tilde{\mathbf{A}}_{j-1}^T \tilde{\mathbf{A}}_j^T \tilde{\mathbf{v}}_j, \end{aligned}$$

on each of the stages $j = 1, \ldots, J_0 \leq J$ of the (partial) MODWT. The $N \times N$ matrices $\tilde{\mathbf{B}}_j$ and $\tilde{\mathbf{A}}_j$ consist of circularly shifted versions of $\{\tilde{h}_l\}$ and $\{\tilde{g}_l\}$ by the amount of one, periodised to length N after upsampling to width $2^{j-1}(L-1)+1$ by inserting $2^{j-1}+1$ zeros between each of the L values of the original wavelet and scaling filter, respectively.

Due to the existent redundancy the MODWT leads to a higher computational cost than the DWT, but using the pyramid algorithm it can be computed with the same complexity as the fast Fourier transform algorithm.

3.4 Multi resolution analysis

To analyse the different frequency contributions existent in a signal $\mathbf{s}$, the wavelet coefficients of each scale can be reconstructed separably. This proceeding is called an MRA: it results in different details $\mathcal{D}_j$ and a smooth function $\mathcal{S}_j$ and is defined by $\mathbf{s} = \sum_{j=1}^{J_0} \mathcal{D}_j + \mathcal{S}_{J_0}$.

The jth level wavelet detail $\mathcal{D}_j$ is defined as

$$\mathcal{D}_j \equiv \mathbf{W}_j^T \mathbf{w}_j = \mathbf{A}_1^T \cdots \mathbf{A}_{j-1}^T \mathbf{B}_j^T \mathbf{w}_j,$$

which is a vector of length N whose elements are associated with changes in $\mathbf{s}$ at scale $\tau_j = 2^{j-1}$ for $j = 1, \ldots, J_0$.

For $0 \leq j \leq J_0 - 1$, the jth level wavelet smooth $\mathcal{S}_j$ is defined as

$$\mathcal{S}_j \equiv \sum_{k=j+1}^{J_0} \mathcal{D}_k + \mathcal{S}_{J_0} = \mathbf{A}_1^T \cdots \mathbf{A}_{j-1}^T \mathbf{A}_j^T \mathbf{v}_j,$$

where $\mathcal{S}_{J_0} \equiv \mathbf{V}_{J_0}^T \mathbf{v}_{J_0}$ represents averages of the scale $\lambda_{J_0} = 2^{J_0}$. $\mathcal{S}_j$ is a smoothed version of $\mathbf{s}$, since the difference with the original signal, $\mathbf{s} - \mathcal{S}_j = \sum_{k=1}^{j} \mathcal{D}_k$ for $j \geq 1$, contains only details at the high frequency scales, $\tau_j = 2^{j-1}$ and smaller. The wavelet smooth at the highest level J is a constant vector with all elements equal to the signal mean $\bar{\mathbf{s}}$.

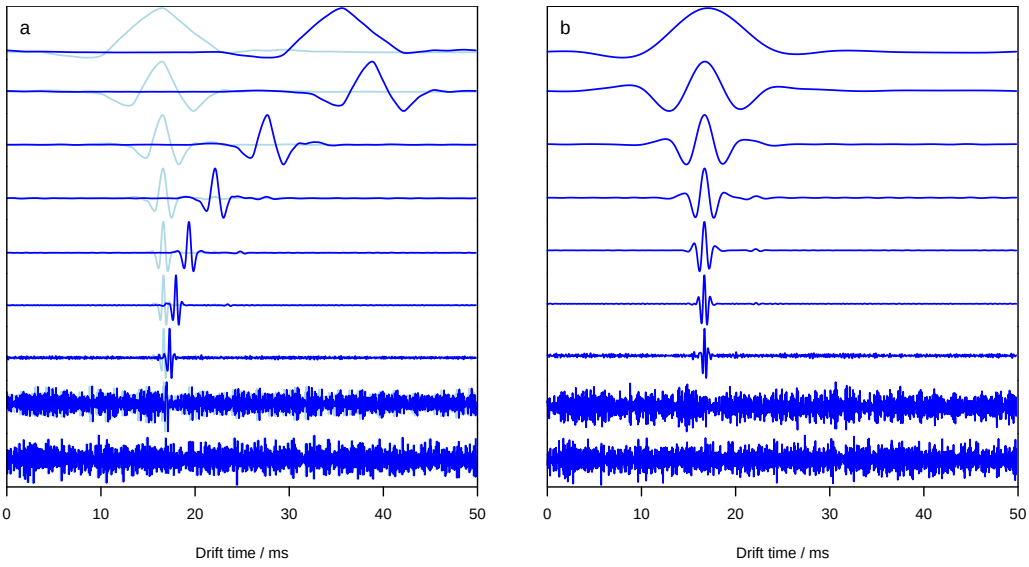

Figure 3.8: Plot of (a) the MODWT decomposition of a single spectrum without (dark blue) and with (light blue) alignment to original signal and (b) the corresponding MRA decomposition based on the MODWT coefficients: the details, illustrated by a continuous line over time for each dyadic scale, show peak shapes more similar to IMS spectra, and little peaks around the main peak appear to be higher and expressed more clearly compared to the underlying MODWT decomposition.

As true for the DWT, the MODWT can be used to form an MRA analogously. In contrast to the usual DWT, the details $\tilde{\mathcal{D}}_j$ and the smooth $\tilde{\mathcal{S}}_{J_0}$ of this MRA are such that circularly shifting the time series by any amount will circularly shift each detail and smooth by a corresponding amount. An alignment with the original signal as mentioned for the MODWT coefficients is not necessary for the details $\tilde{\mathcal{D}}_j$ and the smooth $\tilde{\mathcal{S}}_{J_0}$.

Because the details $\tilde{\mathcal{D}}_j$ and the smooth $\tilde{\mathcal{S}}_{J_0}$ are continuous over time, the result of an MRA can be compared easily with the original signal, but also with the underlying wavelet coefficients, if an MODWT was used as the base for the formation of the MRA (Fig. 3.8). After separate back-transform for the different levels, peak shapes are more similar to those in the original IMS spectrum than for the underlying MODWT coefficients. In addition, the peak indications around the main peak appear to be higher and expressed more clearly than in the MODWT decomposition.

3.5 Separable two-dimensional wavelet transform

To investigate the three-dimensional structure in spectra series data, a two-dimensional wavelet transform such as the separable two-dimensional DWT can be useful, which ap-

plies a common one-dimensional DWT first to all rows, and subsequently to all columns. The proceeding and notations can be transferred from the one-dimensional case, but have to take respect of matrix operations instead of vectors now.

Furthermore, filtering the data matrix $\mathbf{S}$ by the wavelet and the scaling filter in each consecutive combination, a decomposition into four matrices $\mathbf{W}_1^{(HH)}$, $\mathbf{W}_1^{(HL)}$, $\mathbf{W}_1^{(LH)}$, and $\mathbf{W}_1^{(LL)}$ is achieved, where H indicates high-pass and L low-pass filtering with the wavelet and the scaling filter respectively.

Like in the one-dimensional case, further iterations on the next levels are applied only to the low-pass part $\mathbf{W}_1^{(LL)}$. This proceeding results in a matrix system of $3J_0 + 1$ matrices for a partial decomposition of level J_0.

Two-dimensional discrete wavelet transform

For the two-dimensional DWT, assuming the dimensions n_1 and n_2 of the data matrix $\mathbf{S}$ are dyadic, the wavelet coefficients of the two-dimensional decomposition are given by

$$
\begin{aligned}
W_{j,t_1,t_2}^{(HH)} &= \sum_{l_1=0}^{\frac{n_1}{2^{j-1}}-1} \sum_{l_2=0}^{\frac{n_2}{2^{j-1}}-1} \left(h_1^{\circ} h_2^{\circ T}\right)_{l_1,l_2} W_{j-1,2t_1+1-l_1 \bmod \frac{n_1}{2^{j-1}},2t_2+1-l_2 \bmod \frac{n_2}{2^{j-1}}}^{(LL)}, \\
W_{j,t_1,t_2}^{(HL)} &= \sum_{l_1=0}^{\frac{n_1}{2^{j-1}}-1} \sum_{l_2=0}^{\frac{n_2}{2^{j-1}}-1} \left(h_1^{\circ} g_2^{\circ T}\right)_{l_1,l_2} W_{j-1,2t_1+1-l_1 \bmod \frac{n_1}{2^{j-1}},2t_2+1-l_2 \bmod \frac{n_2}{2^{j-1}}}^{(LL)}, \\
W_{j,t_1,t_2}^{(LH)} &= \sum_{l_1=0}^{\frac{n_1}{2^{j-1}}-1} \sum_{l_2=0}^{\frac{n_2}{2^{j-1}}-1} \left(g_1^{\circ} h_2^{\circ T}\right)_{l_1,l_2} W_{j-1,2t_1+1-l_1 \bmod \frac{n_1}{2^{j-1}},2t_2+1-l_2 \bmod \frac{n_2}{2^{j-1}}}^{(LL)}, \\
W_{j,t_1,t_2}^{(LL)} &= \sum_{l_1=0}^{\frac{n_1}{2^{j-1}}-1} \sum_{l_2=0}^{\frac{n_2}{2^{j-1}}-1} \left(g_1^{\circ} g_2^{\circ T}\right)_{l_1,l_2} W_{j-1,2t_1+1-l_1 \bmod \frac{n_1}{2^{j-1}},2t_2+1-l_2 \bmod \frac{n_2}{2^{j-1}}}^{(LL)},
\end{aligned}
$$

for $j = 1, \ldots, J$ with $J = \log_2(\min(n_1, n_2))$ and $\mathbf{W}_0^{(LL)} = \left(W_{0,t_1,t_2}^{(LL)}\right)_{t_1=1,\ldots,n_1;t_2=1,\ldots,n_2} = \mathbf{S}$.

This formulation is closely related to the calculation of the coefficients of the one-dimensional DWT (Equ. 3.1, p. 26). Equivalently, the sequences $\left\{h_i^{\circ}\right\}$ and $\left\{g_i^{\circ}\right\}$ are $\{h\}$ and $\{g\}$ periodised to length $n_i/2^{j-1}$, $i = 1, 2$, and applied to filter the low-pass contribution of the level before.

The $\frac{n_1}{2^j} \times \frac{n_2}{2^j}$ wavelet coefficient matrices $\mathbf{W}_j^{(HH)}$, $\mathbf{W}_j^{(HL)}$, and $\mathbf{W}_j^{(LH)}$ describe respectively the bi-directional, horizontal, and vertical edge information inherent to the data on scale $\tau_j = 2^{j-1}$, whereas $\mathbf{W}_j^{(LL)}$ contains the low-frequency contributions.

The matrix $\mathbf{W}_{j-1}^{(LL)}$ can be reconstructed analogously to the one-dimensional case (Equ. 3.2, page 28) by

$$\mathbf{W}_{j-1}^{(LL)} = \mathbf{B}_{j,1}^T \mathbf{W}_j^{(HH)} \mathbf{B}_{j,2} + \mathbf{B}_{j,1}^T \mathbf{W}_j^{(HL)} \mathbf{A}_{j,2} + \mathbf{A}_{j,1}^T \mathbf{W}_j^{(LH)} \mathbf{B}_{j,2} + \mathbf{A}_{j,1}^T \mathbf{W}_j^{(LL)} \mathbf{A}_{j,2},$$

where $\mathbf{A}_{j,i}$ and $\mathbf{B}_{j,i}$ are the $\frac{n_i}{2^j} \times \frac{n_i}{2^{j-1}}$ matrices containing circularly shifted versions of the wavelet and scaling filters $\{h\}$ and $\{g\}$ by the amount of 2, and periodised to length $\frac{n_i}{2^{j-1}}$, $i = 1, 2$, respectively. Recursive application allows the back-transform to the original data matrix $\mathbf{S}$.

Two-dimensional maximum overlap discrete wavelet transform

To achieve the advantageous properties of the applicability of the two-dimensional wavelet transform to signal matrices $\mathbf{S}$ of arbitrary dimensions, and translation invariance of the resulting coefficients, the method can also be applied on the base of the MODWT. Then the coefficients are derived by

$$\begin{aligned}
\tilde{W}_{j,t_1,t_2}^{(HH)} &= \sum_{l_1=0}^{n_1-1} \sum_{l_2=0}^{n_2-1} \left(\tilde{h}_1^\circ \tilde{h}_2^{\circ T} \right)_{l_1,l_2} \tilde{W}_{j-1,t_1-l_1 \bmod n_1, t_2-l_2 \bmod n_2}^{(LL)}, \\
\tilde{W}_{j,t_1,t_2}^{(HL)} &= \sum_{l_1=0}^{n_1-1} \sum_{l_2=0}^{n_2-1} \left(\tilde{h}_1^\circ \tilde{g}_2^{\circ T} \right)_{l_1,l_2} \tilde{W}_{j-1,t_1-l_1 \bmod n_1, t_2-l_2 \bmod n_2}^{(LL)}, \\
\tilde{W}_{j,t_1,t_2}^{(LH)} &= \sum_{l_1=0}^{n_1-1} \sum_{l_2=0}^{n_2-1} \left(\tilde{g}_1^\circ \tilde{h}_2^{\circ T} \right)_{l_1,l_2} \tilde{W}_{j-1,t_1-l_1 \bmod n_1, t_2-l_2 \bmod n_2}^{(LL)}, \\
\tilde{W}_{j,t_1,t_2}^{(LL)} &= \sum_{l_1=0}^{n_1-1} \sum_{l_2=0}^{n_2-1} \left(\tilde{g}_1^\circ \tilde{g}_2^{\circ T} \right)_{l_1,l_2} \tilde{W}_{j-1,t_1-l_1 \bmod n_1, t_2-l_2 \bmod n_2}^{(LL)},
\end{aligned}$$

where $\left\{ \tilde{h}_i^\circ \right\}$ and $\left\{ \tilde{g}_i^\circ \right\}$ are $\{h\}$ and $\{g\}$ periodised to length n_i, $i = 1, 2$, respectively, after upsampling to the width $2^{j-1}(L-1)+1$ by inserting $2^{j-1}+1$ zeros between each of the L values of the original wavelet and scaling filter.

The resulting wavelet coefficient matrices $\tilde{\mathbf{W}}_j^{(HH)}$, $\tilde{\mathbf{W}}_j^{(HL)}$, $\tilde{\mathbf{W}}_j^{(LH)}$, and $\tilde{\mathbf{W}}_j^{(LL)}$ are of dimension $n_1 \times n_2$, and can be used for the recovery of the same-sized matrix $\tilde{\mathbf{W}}_{j-1}^{(LL)}$, proceeding equivalently to the two-dimensional inverse DWT, using the $n_i \times n_i$ matrices $\tilde{\mathbf{A}}_{j,i}$, $\tilde{\mathbf{B}}_{j,i}$ composed of circularly shifted versions of the filters $\left\{ \tilde{h}_i^\circ \right\}$ and $\left\{ \tilde{g}_i^\circ \right\}$, $i = 1, 2$, by the amount of 1.

In Fig. 3.9 a, a partial two-dimensional MODWT decomposition of level 5 is illustrated in a system of heatmaps, where the MODWT coefficient matrices are arranged in a way that the upper left matrix is iteratively decomposed in the range of the transform.

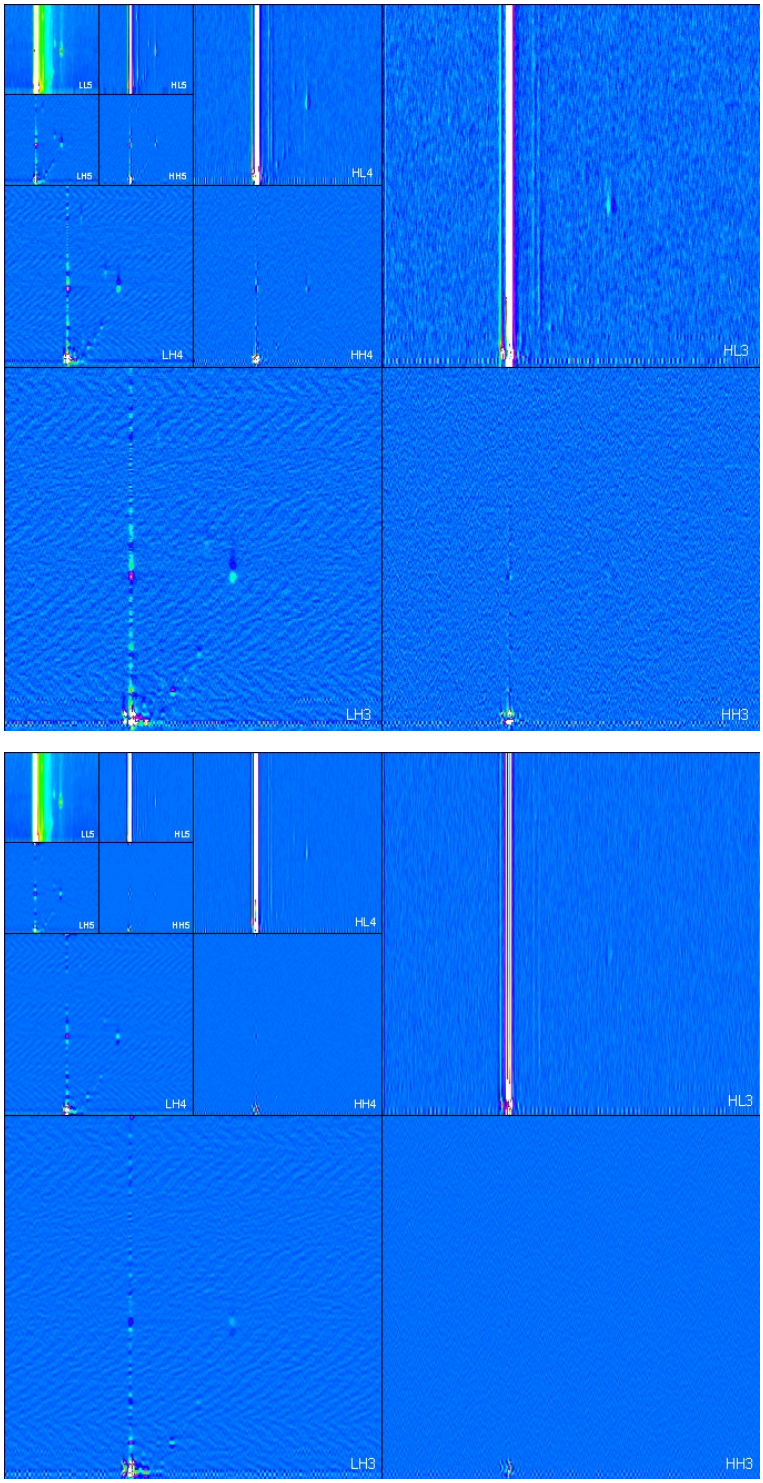

Figure 3.9: Plot of the levels 3, 4, and 5 of (a) the partial 2D-MODWT decomposition of level 5 of a spectra series and (b) the corresponding 2D-MRA decomposition: for both methods the matrices are arranged in a way that the upper left matrix is iteratively decomposed in the range of the transforms. In opposite to the 2D-MODWT, no alignment of the matrices with the original data matrix is necessary for the 2D-MRA, and although peaks appear less high, their shape is more similar to the original data for this method.

In addition to the different frequency levels, the original signal matrix was furthermore divided into matrices $\tilde{\mathbf{W}}_j^{(HH)}$, $\tilde{\mathbf{W}}_j^{(HL)}$, and $\tilde{\mathbf{W}}_j^{(LH)}$ for each level, describing the bi-directional, horizontal, and vertical edge information inherent to the data on scale $\tau_j = 2^{j-1}$, respectively. While the HH matrices are more or less uninformative, the HL matrices are disguised by a stretching effect in the dimension of the retention time. The LH matrices on the other hand hold the main part of the peak information inherent in the original spectra series. The matrix $\mathbf{W}_5^{(LL)}$ contains the low-frequency contributions of this partial decomposition.

Two-dimensional multi resolution analysis

If the signal contributions of specific scales are of interest, an MRA can be computed for the three-dimensional data $\mathbf{S}$ resulting in a matrix convolution according to that of the two-dimensional wavelet transform.

The MRA, therefore, results in three wavelet details on the jth level defined as

$$\begin{aligned}
\mathcal{D}_j^{(HH)} &\equiv \mathbf{A}_{1,1}^T \cdots \mathbf{A}_{j-1,1}^T \mathbf{B}_{j,1}^T \mathbf{W}_j^{(HH)} \mathbf{B}_{j,2} \mathbf{A}_{j-1,2} \cdots \mathbf{A}_{1,2}, \\
\mathcal{D}_j^{(HL)} &\equiv \mathbf{A}_{1,1}^T \cdots \mathbf{A}_{j-1,1}^T \mathbf{B}_{j,1}^T \mathbf{W}_j^{(HL)} \mathbf{A}_{j,2} \mathbf{A}_{j-1,2} \cdots \mathbf{A}_{1,2}, \\
\mathcal{D}_j^{(LH)} &\equiv \mathbf{A}_{1,1}^T \cdots \mathbf{A}_{j-1,1}^T \mathbf{A}_{j,1}^T \mathbf{W}_j^{(LH)} \mathbf{B}_{j,2} \mathbf{A}_{j-1,2} \cdots \mathbf{A}_{1,2},
\end{aligned}$$

which are $n_1 \times n_2$ dimensional matrices whose elements are associated with changes in $\mathbf{S}$ at a scale $\tau_j = 2^{j-1}$ for $j = 1, \ldots, J$.

The jth level wavelet smooth is defined as

$$\mathcal{S}_j^{(LL)} \equiv \mathbf{A}_{1,1}^T \cdots \mathbf{A}_{j-1,1}^T \mathbf{A}_{j,1}^T \mathbf{W}_j^{(LL)} \mathbf{A}_{j,2} \mathbf{A}_{j-1,2} \cdots \mathbf{A}_{1,2},$$

for $0 \leq j \leq J-1$, representing averages of the scale $\lambda_j = 2^j$.

The signal reconstruction can then be formulated as

$$\mathbf{S} = \sum_{j=1}^{J} \left(\mathcal{D}_j^{(HH)} + \mathcal{D}_j^{(HL)} + \mathcal{D}_j^{(LH)} \right) + \mathcal{S}_J^{(LL)}.$$

The MRA can also be achieved using the (partial) MODWT analogously, leading to translation-invariant results for an arbitrary sized matrix $\mathbf{S}$.

Fig. 3.9 b displays a two-dimensional MRA based on the coefficient matrices of a MODWT (Fig. 3.9 a) – again for the case of a partial decomposition of level 5. The matrices are

arranged as introduced before, and the findings for the MODWT can mostly be transferred for the MRA result as well, but in opposite to the 2D-MODWT, no alignment of the matrices with the original data matrix is necessary now. In addition, it can be observed that although peaks appear less high for the MRA, their shape is more similar to the original data than for the MODWT coefficients.

Chapter 4

Statistical methods

In the practical application of spectrometric devices, the detection of analytic peaks is usually inevitable, but often this preparatory step only gives the basis for a subsequent statistical analysis, such as in the comparison of groups of measurements (Section 4.2) or classification tasks (Section 4.3).

Before these methods can be applied, the performance of an additional general variable creation step by the means of cluster analysis (Section 4.1) was found to be useful, thus enabling further analysis on the basis of a common set of problem-related peak variables.

4.1 Cluster analysis

The multivariate method of cluster analysis has the main aim of joining a set of classification objects to homogeneous groups. The basic idea of a homogeneous group is determined by internal cohesion (homogeneity within a cluster) and external isolation (heterogeneity between the clusters).

The field of assignment principles in this work was limited to deterministic methods that result in an explicit assignment with a probability of 1, which can be divided into the two groups of hierarchical (Subsection 4.1.3) and partition (Subsection 4.1.4) clustering methods. Furthermore, the presentation was restricted to object-oriented cluster analysis especially for quantitative variables, in contrast to a variable-oriented analysis where groups of variables are clustered.

Before the actual cluster analysis can be performed, a measure for the distance of objects is required (Subsection 4.1.2), and the standardisation of the classification variables can

be useful to allow for comparability (Subsection 4.1.1). To judge the quality of a cluster solution, certain figures of merits can be considered (Subsection 4.1.5) that also allow decisions on the optimal number of clusters (Kaufman and Rousseeuw, 1990).

4.1.1 Standardisation

The variables x_i, $i = 1, \ldots, p$, involved in a deterministic cluster analysis of classification objects ω_n with observed variable vectors $\mathbf{x}_n$, $n = 1, \ldots, N$, have to be formally comparable, which is not fulfilled, for example, if variables are given in different units or possess values showing a different order of magnitude or variability. Then a transformation of variables can ensure equal influence on the resulting cluster formation.

Commonly used data transformations are the theoretical and empirical standardisation

$$\tilde{x}_{ni}^{(t)} = \frac{x_{ni} - \mu_i}{\sigma_i} \quad \text{and} \quad \tilde{x}_{ni}^{(e)} = \frac{x_{ni} - \bar{x}_i}{s_i} \quad \text{for} \quad n = 1, \ldots, N,\ i = 1, \ldots, p, \tag{4.1}$$

respectively, where $\tilde{x}_{ni}$ gives the standardised value of object ω_n in the standardised variable $\tilde{X}_i$; x_{ni} is the value of object ω_n for variable X_i before standardisation with theoretical mean μ_i and standard deviation σ_i; $\bar{x}_i$ and s_i are the empirical pendants of these moments.

While the parameters of the distribution of the variables have to be defined for the theoretical transformation, an empirical standardisation can always be applied, achieving equal weight for all variables, as only the part of the variable domain that is actually represented in the data is included. If all variables contribute to the separation of classes this effect is positive.

4.1.2 Distance measures

In a cluster analysis, homogeneous groups are characterised by the similarity or dissimilarity of classification objects within or between clusters, requiring a measure of similarity or dissimilarity that respects the data and the aim of the analysis. As object-oriented cluster analysis is usually based on distance measures, only these were considered here:

Definition 4.1 Let $I = \{\omega_1, ..., \omega_N\}$ be a set of N objects. A function $d : I \times I \to \mathbb{R}$ is called *distance measure*, if

$$d(\omega_n, \omega_n) = 0, \quad d(\omega_n, \omega_m) \geq 0, \quad \text{and} \quad d(\omega_n, \omega_m) = d(\omega_m, \omega_n), \quad n, m = 1, \ldots, N.$$

The symmetrical $N \times N$ matrix $\mathbf{D} = (d(\omega_n, \omega_m))$ is called *distance matrix*.

Measures fulfilling the triangle inequality

$$d(\omega_n, \omega_m) \leq d(\omega_n, \omega_l) + d(\omega_m, \omega_l) \qquad \text{for} \qquad n, m, l = 1, ..., N,$$

possess the advantageous property of being metric, and are preferable as they are consistent with our spatial sense. For quantitative data, such distance measures are often derived from the generalised Minkowski metric,

$$d_{q;r}(\omega_n, \omega_m) = \left(\sum_{i=1}^{n} |x_{ni} - x_{mi}|^r \right)^{\frac{1}{q}}, \quad q, r \geq 1, \quad n, m = 1, \ldots, N,$$

to keep the reduction of information inherent to every calculation of distances from a data matrix low. Using the generalised Minkowski metric, the parameter r causes a weighting of the differences in the single variables with increasing r, resulting in a higher influence of big differences in a few variables, compared to small differences in many variables. The parameter q, on the other hand, effects a back standardisation to the original scale unit if $r = q$, yielding the special case of L_q distances or Minkowski q-metrics:

$$d_q(\omega_n, \omega_m) = \left(\sum_{i=1}^{n} |x_{ni} - x_{mi}|^q \right)^{\frac{1}{q}}, \quad q \geq 1, \quad n, m = 1, \ldots, N.$$

In addition to translation invariance, implying the independence of distances from the choice of origin, Minkowski q-metrics possess the property of being metric.

(Generalised) Minkowski metrics are, however, not scale invariant, so that distances are dependent from the unit of variables. The standardisation of variables before calculation of the distance matrix $\mathbf{D}$ is, therefore, recommended to adjust for a different amount of influence. In this context, the empirical standardisation given in Equ. 4.1 is commonly used, substituting the empirical standard deviation s_i by

$$s_i^{(q;r)} = \left(\frac{1}{N} \sum_{i=1}^{N} |x_{ni} - \bar{x}_i|^r \right)^{\frac{1}{q}}, \quad q, r \geq 1, \quad n = 1, \ldots, N. \tag{4.2}$$

A commonly used special case of the Minkowski q-metric is, in particular, the Euclidean distance (L_2 distance):

$$\begin{aligned} d_2(\omega_n, \omega_m) &= ((\mathbf{x}_n - \mathbf{x}_m)'(\mathbf{x}_n - \mathbf{x}_m))^{\frac{1}{2}} \\ &= \|\mathbf{x}_n - \mathbf{x}_m\| \end{aligned}, \quad n, m = 1, \ldots, N. \tag{4.3}$$

A crucial advantage of the Euclidean distance is the invariance property regarding orthogonal transforms: the distance $d_2(\omega_n, \omega_m)$ remains unchanged when $\mathbf{x}_n$ and $\mathbf{x}_m$ are

substituted by $\mathbf{C}\mathbf{x}_n$ and $\mathbf{C}\mathbf{x}_m$, if $\mathbf{C}$ is an orthogonal matrix. In essence this property means that the Euclidean distance neglects rotations and mirroring of the coordinate system. Additionally, for the Euclidean distance the value $s_i^{(q,r)}$ in Equ. 4.2 coincides with the empirical standard deviation s_i, and the recommended standardisation for Minkowski metrics, therefore, conforms with the empirical standardisation (Subsection 4.1.1).

The squared Euclidean distance

$$\left(d_2(\omega_n, \omega_m)\right)^2 = (\mathbf{x}_n - \mathbf{x}_m)'(\mathbf{x}_n - \mathbf{x}_m) = \|\mathbf{x}_n - \mathbf{x}_m\|^2 \ , \quad n, m = 1, \ldots, N,$$

which is often used because of its arithmetical simplicity, is not a Minkowski q-metric as $r{=}2 \neq q{=}1$, and, therefore, no metric distance measure.

4.1.3 Hierarchical cluster methods

Hierarchical cluster methods result in a sequence of partitions of the set of classification objects $I = \{\omega_1, ..., \omega_N\}$, where each object belongs to exactly one of k clusters $C_1, ..., C_k$. For each partition $\mathcal{C} = \{C_1, ..., C_k\}$ it is essential that

$$\bigcup_{g=1}^{k} C_g = I, \quad C_g \cap C_h = \emptyset, \quad g \neq h, \quad h = 1, \ldots, k;$$

one such partition is generated for each possible cluster quantity $1 \leq k \leq N$. For this, the clusters are constructed in a way that the distances between the objects within one cluster are possibly low, while the distances between the objects of different clusters on the other hand are to be possibly high, which is quantified by the means of a distance measure (Subsection 4.1.2).

Hierarchical cluster methods can proceed in two opposite directions: agglomerative methods successively merge those clusters which show the smallest distance, resulting in a decreasing number of clusters; divisive methods vice versa execute a stepwise decomposition into subclusters and, therefore, yield an increasing number of clusters in the course of the procedure. The relation between single cluster solutions is hierarchical, as once a merging (agglomerative) or division (divisive) of clusters has taken place, it can not be reversed in a subsequent step.

Amongst the divisive methods, monothetic and polythetic methods are differentiated. Monothetic methods only consider one variable per step for the division of clusters, so that the result of the classification is undesirably dependent on the order of the respected variables. Polythetic methods, on the other hand, search for the partition, dividing one

of the established clusters in two smaller clusters in a way that the difference between the variance in the subdivided cluster and the sum of the variances in the two evolving clusters is maximal, using the variance as an indicator for homogeneity. As there are $2^{|C_k|-1} - 1$ possibilities to divide a cluster C_k with $|C_k|$ objects into two clusters, the polythetic methods have the disadvantage of a high computational cost. For this reason only agglomerative methods are further discussed.

Different agglomerative methods differ only in the measure $D(C_g, C_h)$ of the distance between the subsets of objects. As the measure D is furthermore based on the distance measure $d(\omega_n, \omega_m)$ quantifying the distance between two objects, the obtained cluster solution is dependent on the cluster method as well as the chosen distance measure.

The construction of a sequence of partitions is carried out according to the following scheme:

(1) In the original partition $\mathcal{C}^0 = \{\{\omega_1\}, ..., \{\omega_N\}\}$ every classification object constitutes an independent cluster.
The cluster number k corresponds with the number of classification objects N.

(2) Every following partition $\mathcal{C}^j$ $(j \geq 1)$ is derived by merging the two clusters in $\mathcal{C}^{j-1}$, that minimise the measure D implied by the specified cluster procedure.
The cluster number k amounts to $N - j$.

(3) Step 2 is iterated till $\mathcal{C}^j = \mathcal{C}^{N-1} = \{I\}$ contains all classification objects.
The cluster number k is then 1.

On each hierarchical step the so-called fusion level

$$D_j = \min_{g \neq h} D(C_g, C_h) \ , \quad j = 1, \ldots, N-1, \quad g, h = 1, \ldots, k,$$

is determined, allowing conclusions to be drawn on the homogeneity of the cluster formed in the jth hierarchical step. These values D_j can be used to decide how many classes are underlying the considered data (Subsection 4.1.5).

The most common agglomerative cluster methods are presented on the following pages, starting with the two oldest and simplest procedures constituted by the single and the complete linkage, which are also denominated as basis models.

Single linkage

For the single linkage the distance between two clusters C_g and C_h is set to the smallest distance between an object from C_g and an object belonging to C_h:

$$D(C_g, C_h) = \min_{\substack{\omega_n \in C_g, \\ \omega_m \in C_h}} \{d(\omega_n, \omega_m)\} , \quad g, h = 1, \ldots, k.$$

This method has the property of being able to detect clusters of diverse shape, and is, therefore, especially capable for non-spherical, e.g. long stretched clusters, if the clusters underlying the data are externally isolated, but not internally coherent. If the data, however, contains clusters that are not clearly isolated, a disadvantageous linkage effect can be observed, causing the fusion of clusters that are generally well-separated if only a few objects fill the space between the clusters. The reason for this contraction of the classification space are the low constraints of the method considering the homogeneity within a cluster.

Complete linkage

For the complete linkage the distance of two clusters C_g and C_h is chosen as the greatest distance of an object in C_g and and object in C_h:

$$D(C_g, C_h) = \max_{\substack{\omega_n \in C_g, \\ \omega_m \in C_h}} \{d(\omega_n, \omega_m)\} , \quad g, h = 1, \ldots, k.$$

A disadvantage of the complete linkage is the so-called dilatation effect, as originally long stretched clusters are covered by a string of small spherical clusters, generally yielding too many clusters by reason of the strong constraints for the homogeneity within the clusters.

Both complete and single linkage are qualified only for small sets of classification objects, as they require the entire distance matrix to be kept in memory. Furthermore they are especially sensitive considering sources of errors as outliers (Punj and Stewart, 1983). For this reason some modifications of the two basis models are now introduced, which compensate their disadvantageous effects.

Average linkage

The average linkage calculates the distance D of two clusters C_g and C_h as the average of all distances between objects in C_g and C_h,

$$D(C_g, C_h) = \frac{1}{n_g n_h} \sum_{\omega_n \in C_g} \sum_{\omega_m \in C_h} d(\omega_n, \omega_m), \quad n_i = |C_i|, \quad g, h, i = 1, \ldots, k,$$

with n_k giving the number of objects belonging to cluster C_k. For the fusion of two clusters it is, therefore, sufficient, if the objects of both clusters are satisfactorily, on average, similar, so that the dissimilarity of objects which lie further apart can be balanced by small distances of particularly closely located objects. In this, the linkage and dilatation effect of the two base models are avoided.

Weighted average linkage

The weighted average linkage is another modification of the single and the complete linkage, developed to balance the disadvantageous effects of these two base models. In comparison to the average linkage method, however, it is harder to interpret.

Assuming that the cluster C_g was composed after cluster C_h and additionally that C_{g_1} and C_{g_2} are the two clusters which were merged to cluster C_g, the distance of the clusters C_g and C_h is determined as the mean of the distances between C_h and C_{g_i} for $i = 1, 2$:

$$D(C_g, C_h) = \frac{1}{2} \sum_{i=1}^{2} D(C_{g_i}, C_h) \ , \quad g, h = 1, \ldots, k.$$

The distance measure $d(\omega_m, \omega_n)$, which is chosen independently of the applied cluster method, influences the calculations only if two singletons are merged. The fusion level D_j here equals the smallest average pairwise dissimilarities between two clusters C_v and C_w, if the clusters in the preceding fusion step are considered as being of equal size and, therefore, possessing no direct relation to the empirical distances.

As the term of weighted averages can either refer to the distances between clusters or to the distances of objects of the merged clusters, the two average methods hold a wide variety of different names in literature, where they are sometimes even denominated vice versa.

Ward's method

In opposite to the cluster methods previously introduced allowing a free choice of the distance measure $d(\omega_n, \omega_m)$, Ward's method implies the usage of the Euclidean distance for the determination of distances between classification objects. The criterion for the two clusters that are to be merged on a step of the cluster procedure on the other hand is dependent on the loss of homogeneity observed if two clusters C_g and C_h are merged,

where the homogeneity of a cluster solution $H(\mathcal{C}^{j-1})$ is measured by the sum of variances within the clusters:

$$H(\mathcal{C}^{j-1}) = \sum_{g=1}^{k} \sum_{\omega_n \in C_g} \|\mathbf{x}_n - \bar{\mathbf{x}}_g\|^2.$$

If the partition $\mathcal{C}^j$ is obtained by the fusion of the clusters C_v and C_w, the homogeneity of the cluster solution of this step of the procedure is calculated as

$$H(\mathcal{C}^j) = \sum_{g \neq v,w} \sum_{\omega_n \in C_g} \|\mathbf{x}_n - \bar{\mathbf{x}}_g\|^2 + \sum_{\omega_n \in C} \|\mathbf{x}_n - \bar{\mathbf{x}}_C\|^2, \quad \text{with} \quad \begin{array}{l} C = C_v \cup C_w \ , \\ g, v, w = 1, \ldots, k. \end{array}$$

The difference of the homogeneities $H(\mathcal{C}^j) - H(\mathcal{C}^{j-1})$ is then closely related to the measure, which Ward's method uses for the determination of distances between clusters, where the actual value of $D(C_g, C_h)$ is given by the square root of this difference:

$$D(C_g, C_h) = \sqrt{\frac{2 n_g n_h}{n_g + n_h} \|\bar{\mathbf{x}}_g - \bar{\mathbf{x}}_h\|^2} \ , \quad g, h = 1, \ldots, k. \tag{4.4}$$

The factor 2 in Equ. 4.4 was inserted to ensure for the special case of two singletons C_g and C_h that $D(C_g, C_h)$ coincides with the original Euclidean distance between the objects (Kaufman and Rousseeuw, 1990).

4.1.4 Partition cluster methods

For partition cluster methods, previously established partitions of classification objects are not irrevocable – objects can change into a different cluster on each consecutive iteration step. The number of clusters constituting the resulting partition, however, has to be stated before the start of the performance. Coming from a set of starting values, the stepwise optimisation of a specified objective function is then aimed, until a partition results that can not be further improved by shifting any classification objects to another cluster.

The principle of the iterative proceeding can be described as follows:

(1) After the calculation or the input of starting values, an initial partition $\mathcal{C}^0$ is established.

(2) For each classification object in partition $\mathcal{C}^j$ it is checked if the value of the objective function can be improved by shifting it to another cluster.

(3) Depending on the procedure, either all objects that cause an improvement or just the object causing the most improvement are shifted to the according cluster, resulting in the new partition $\mathcal{C}^{j+1}$.

(4) Steps 2 and 3 are iterated till no change is effected any longer.

The result of partition cluster methods is not necessarily the partition that optimises the given objective function – in many cases the shifting procedure stops at a suboptimal solution. For the identification of the global maximum, the entirety of possible partitions would have to be formed to choose the one with the optimal value of the performance index. There is, however, no exact method that solves this optimisation task with reasonable computational costs, as the number of potential partitions $\frac{1}{k!}\sum_{g=0}^{k}(-1)^g\binom{k}{g}(k-g)^N$ gets extremely high already for rather small numbers of classification objects N and cluster numbers k. For the example of $N = 100$ objects, there are $2.756 \cdot 10^{93}$ possible partitions that divide these into $k = 10$ not empty disjunct cluster sets. In most applications a complete enumeration is, therefore, not possible.

The common way of overcoming this problem are heuristic methods, which yield a locally optimised solution based on a set of starting values, so that better solutions can only be found for different starting values. A consequence of this procedure is the dependency of the cluster solution from the starting values, which are often chosen randomly. To counter the sensitivity of the methods against the choice of different starting values, it can be useful to calculate the partition method for different initial partitions and choose the best arising cluster solution. Another less arbitrary alternative is the use of the result from a hierarchical cluster method to create an initial partition $\mathcal{C}^0$.

The most common partition cluster procedure is the k-means method, which will now be introduced and compared with the method of k-medoids.

k-means method

The k-means method is the most common and frequently used partition cluster method. It is based on the objective function of the variance within the clusters, which should be small for clusters with similar objects. For the distance measure $d(\omega_n, \omega_m)$, the use of the (squared) Euclidean distance is implied. An efficient version of the k-means cluster algorithm is described in Hartigan and Wong (1979).

The variance criterion can be formulated as

$$\begin{aligned} tr\mathbf{W}(\mathcal{C}) &= \sum_{g=1}^{k} \sum_{\omega_n \in C_g} \|\mathbf{x}_n - \bar{\mathbf{x}}_g\|^2 \\ &= \sum_{g=1}^{k} \frac{1}{2n_g} \sum_{\omega_n \in C_g} \sum_{\omega_m \in C_g} \|\mathbf{x}_n - \mathbf{x}_m\|^2 \rightarrow \min_{\mathcal{C} \in \Gamma}, \end{aligned}$$

where $\mathbf{W}(\mathcal{C}) = \sum_{g=1}^{k} \sum_{\omega_n \in C_g} (\mathbf{x}_n - \bar{\mathbf{x}}_g)(\mathbf{x}_n - \bar{\mathbf{x}}_g)^T$ gives the matrix of variances within the clusters of the partition $\mathcal{C}$, and Γ is the set of all possible partitions $\mathcal{C} = \{C_1, ..., C_k\}$ of N objects in k clusters. The sum of variances within the clusters can be interpreted as the amount of variance that is not explained by the cluster solution.

This figure of merit of the k-means method offers the important minimal distance property: in an optimal partition $\hat{\mathcal{C}} = \{\hat{C}_1, ..., \hat{C}_k\}$ according to the variance criterion it is

$$\|\mathbf{x}_n - \hat{\bar{\mathbf{x}}}_g\|^2 \leq \|\mathbf{x}_n - \hat{\bar{\mathbf{x}}}_h\|^2, \quad \text{for } \omega_n \in \hat{C}_g, \quad g, h = 1, ..., k.$$

In essence this means that the squared Euclidean distance to the center $\hat{\bar{\mathbf{x}}}_g$ of the cluster $\hat{C}_g$ that an object belongs to is smaller than, or equal to, the squared Euclidean distance of this object to the other cluster centers.

k-medoids method

An alternative partition method is the k-mediods method, which considers objects instead of mean values as cluster centers. Kaufman and Rousseeuw (1990) call these centers medoids and the method accordingly k-medoids. This robust method is superior to the k-means method in some special applications, though for the examination of large amounts of data the k-medoids method involves high computational cost. Furthermore, the existence of quantitative data in a data matrix generally implies the use of the k-means method, for which reason the k-medoids method is not closer elucidated here.

4.1.5 Performance indices

In cluster analysis the true class memberships of classification objects is unknown. A cluster solution can, therefore, not be categorised as right or wrong, but only as more or less useful. Two performance indices that allow to evaluate the usability of cluster solutions for both hierarchical and partition cluster methods are, therefore, presented.

Calinski and Harabasz (1974) approach the choice of cluster solutions in terms of the multivariate analysis of variances and introduce the variance ratio criterion (VRC) as an index that is based on the variance matrices,

$$\mathbf{B}(\mathcal{C}) = \sum_{g=1}^{k} n_g(\bar{\mathbf{x}}_g - \bar{\mathbf{x}})(\bar{\mathbf{x}}_g - \bar{\mathbf{x}})^T \quad \text{and} \quad \mathbf{W}(\mathcal{C}) = \sum_{g=1}^{k} \sum_{\omega_n \in C_g} (\mathbf{x}_n - \bar{\mathbf{x}}_g)(\mathbf{x}_n - \bar{\mathbf{x}_g})^T,$$

between and within the clusters of a partition $\mathcal{C} = \{C_1, ..., C_k\}$ respectively. Number n_g and mean $\bar{\mathbf{x}}_g$ correspond to the objects of cluster C_g, while $\bar{\mathbf{x}}$ is the mean of all classification objects. The index of Calinski and Harabasz is than derived as

$$VRC = \frac{tr(\mathbf{B}(\mathcal{C}))}{tr(\mathbf{W}(\mathcal{C}))} \frac{N-k}{k-1}. \tag{4.5}$$

Comparing different cluster solutions, a high value of the VRC indicates the solution that is most appropriate to describe the examined data.

In the comparison of 30 methods for the identification of good cluster solutions by Milligan and Cooper (1985), the VRC was the index that performed best. This examination, however, only included hierarchical methods for the determination of the evaluated partitions.

An alternative criterion for the goodness of a cluster solution is proposed by Rousseeuw (1987) as the average silhouette width $\bar{s}$, which is based on the term of the silhouette, originally used for displaying cluster solutions of arbitrary methods. For its calculation a value $s(\omega_n)$, evaluating the goodness of classification is assigned to each object ω_n. This requires the determination of the measure $a(\omega_n)$, quantifying the average dissimilarity between $\mathbf{x}_n$ and all other variable vectors of objects of the cluster that object ω_n belongs to. Furthermore, the average distance $d(\omega_n, C)$ between $\mathbf{x}_n$ and all objects in an other cluster C is calculated for all other clusters. The smallest value of these distances $d(\omega_n, C)$ is then denominated as $b(\omega_n) = \min_C d(\omega_n, C)$ and can be interpreted as the average distance between $\mathbf{x}_n$ and the variable vectors of the objects of its neighbour cluster. The distances that are necessitated for computing the measures $a(\omega_n)$ and $b(\omega_n)$ are determined according to the distance measure which was used for the construction of the evaluated cluster partition. The silhouette width $s(\omega_n)$ is then defined as:

$$s(n) = \frac{b(\omega_n) - a(\omega_n)}{\max\{a(\omega_n), b(\omega_n)\}}, \quad n = 1, \ldots, N.$$

If object ω_n is well-classified, $s(\omega_n)$ has a value near 1, while small values of $|s(\omega_n)|$ near zero indicate a location of the object between two clusters. Classification objects possessing a negative $s(\omega_n)$ value are likely to be assigned to the wrong cluster.

A measure for the goodness of a cluster solution can be derived by considering all silhouette widths simultaneously in the average silhouette width of all classification objects:

$$\bar{s} = \sum_{i=1}^{N} s(\omega_i). \tag{4.6}$$

Comparing cluster solutions with a different amount of clusters or resulting from different cluster methods, the solution with the highest value of this criterion is preferable.

While hierarchical methods directly produce a sequence of partitions, the iterative procedure of partition cluster methods has to be applied for each considered cluster number k separately, if an evaluation of the best cluster number is aspired to. Kaufman and Rousseeuw (1990) define the silhouette coefficient SC for this purpose, which is based on the average silhouette width and is determined as

$$SC = \max_k \bar{s}(\omega_k) \ , \quad k = 1, \ldots, N.$$

At the same time, this coefficient can be seen as an indicator for the existence of an underlying cluster structure.

4.2 Group comparison

The comparison of two sets of samples is a common problem in practical applications, where the same variable is usually measured for different conditions or populations, and the expected value is subject of a location comparison. The most popular test for the described situation is the t-test, assuming a normal distribution of the variables. If this assumption is not fulfilled, its also common non-parametric equivalent, the Wilcoxon test, is often a reasonable choice.

For the data examined in this work the Wilcoxon test could not be applied, as the necessary assumption of equal shape of distributions in both sample sets could not be verified. For this reason and because of the sufficiently high sample size implying an approximate validity of the t-test, the application of this parametric test was preferred (Subsection 4.2.1).

As a high number of comparisons for different variables between the same populations was involved in this work, the problem of multiple testing is furthermore introduced and respected by means of the Bonferroni-Holm method (Subsection 4.2.2).

4.2.1 *t*-test

The t-test introduced here is a parametric test for the comparison of two independent samples with unknown parameters of the underlying distributions. Both expectation and variance, therefore, have to be estimated from the data. While the equality of the expected values is the actual issue of the test, the unknown variances were supposed to be unequal.

It is presumed that the data of the two variables X_1 and X_2 are independent realisations $X_{11}, \ldots, X_{n_1 1}$ and $X_{12}, \ldots, X_{n_2 2}$ and identically distributed as X_1 and X_2 respectively. Furthermore, it is assumed that the entire sample set $X_{11}, \ldots, X_{n_1 1}$, $X_{12}, \ldots, X_{n_2 2}$, is independent, and the two variables X_1 and X_2 follow the normal distributions $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$.

In the following, a two-sided t-test also denominated as Student's t-test is described, checking the hypotheses

$$H_0: \quad \mu_1 = \mu_2 \qquad \text{versus} \qquad H_1: \quad \mu_1 \neq \mu_2.$$

For the evaluation of these hypotheses the mean values $\bar{X}_1 = \frac{1}{n_1} \sum_{i=1}^{n_1} X_{i1}$ and $\bar{X}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} X_{j2}$ are considered, as they give an unbiased estimation of the expected values μ_1 and μ_2. The difference of these means $\bar{X}_1 - \bar{X}_2$ can be seen as an indicator for the equality or inequality of the location parameters of interest, it is, however, not qualified as a test statistic, as the meaning of the absolute differences between the two means is dependent on the variances. A standardisation with the variance of the considered difference,

$$\mathrm{Var}(\bar{X}_1 - \bar{X}_2) = \mathrm{Var}(\bar{X}_1) + \mathrm{Var}(\bar{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2},$$

as $Cov(\bar{X}_1, \bar{X}_2) = 0$ because of the claimed independence, is, therefore, reasonable and necessary.

If the variances are unknown, σ_1^2 and σ_2^2 are estimated and substituted by the empirical variances S_1^2 and S_2^2:

$$S_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_{i1} - \bar{X}_1)^2 \qquad \text{and} \qquad S_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (X_{j2} - \bar{X}_2)^2.$$

The test statistic T resulting after standardisation of the difference: $\bar{X}_1 - \bar{X}_2$ is, therefore,

$$T = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_1^2}{n_2}}},$$

which is not normally distributed by reason of the approximative standardisation, but can be proved to follow a t-distribution assuming the null hypothesis of equal means for both variables. For this, the fraction in the test statistic is expanded by $1 \Big/ \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$, which yields a numerator that is $N(0,1)$-distributed. Furthermore, the denominator is now closely related with a χ^2-distribution with ν degrees of freedom, where the Welch-Satterthwaite approximation implies

$$\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{S_2^2}{n_2}\right)^2}.$$

As T can be expressed as a ratio of a normally distributed random variable and the square of a χ^2-distributed random variable divided by the corresponding degrees of freedom, it possesses a t-distribution.

Consequently, the null hypothesis of the test is rejected if $|T| > t_{1-\alpha}(\nu)$ with $t_{1-\alpha}(\nu)$ giving the $(1-\alpha)$ quantile of the t-distribution with ν degrees of freedom.[1] For large sample sets ($n_1, n_2 > 30$) the test statistic T is approximately standard normal distributed; the quantiles of the introduced decision rule can then be substituted by the corresponding quantiles of the $N(0,1)$-distribution.

4.2.2 Multiple testing

If more than one test problem is examined simultaneously on the basis of the same data set, the problem of multiple testing arises, as the probability of a wrong decision for the alternative hypothesis is no longer controlled. A multiple test problem appears for example, if the same hypothesis is tested in several non-overlapping groups (independent tests), or if more than one variable is compared for the same samples (dependent tests).

For a single test problem, the probability to reject the null hypothesis wrongly is limited by the chosen significance level α, ensuring that a decision for the alternative hypothesis is true with a high probability $1-\alpha$. If, however, k pairs of hypothesis are tested simultaneously for stochastic independent tests with level α, the probability to reject at least one of the null hypotheses wrongly if all null hypotheses are assumed to be true is only limited by $\alpha_k = 1-(1-\alpha)^k$. For $\alpha = 0.05$ and $k = 100$, the probability for at least one falsely significant result is, therefore, $\alpha_k = 0.994$. Accordingly, a simple correction for multiple testing can be arranged for independent tests with the Šidák method, ensuring

[1] The number of degrees of freedom ν has to be rounded down to ensure a valid value input.

the control of the multiple significance level by examining the k single tests each with the level $1-(1-\alpha)^{1/k}$.

For dependent tests the calculation of the error probability α_k is more complicated and no general equation for α_k can be formulated. Several methods, however, exist to constrain the level α_k by an upper limit to avoid exceeding the specified error probabilities. Importantly, the corrected procedure should be not too conservative to retain a possibly high power of the tests.

The easiest way to controll the significance level of simultaneous tests is, in the case of stochastic dependency, the Bonferroni correction, which indicates the execution of each test to the level α/k. This method ensures that the multiple level α is not exceeded, but the rough estimation of α_k reduces the power of the test considerably, because of the strong decrease of the level for the single tests.

An alternative method yielding a less conservative test procedure is the Bonferroni-Holm method (Holm, 1979). The application of this methods involves the ordering of the k elementary null hypotheses $H_0^1, \ldots, H_0^k$ after the calculation of the p-values $p_1, \ldots, p_k$ for the single test problems by the ordered p-values $p_{(1)}, \ldots, p_{(k)}$ to $H_0^{(1)}, \ldots, H_0^{(k)}$. Then a stepwise rejection of hypotheses $H_0^{(i)}$ is executed till the inequality

$$p_{(i)} \leq \frac{\alpha}{k-(i-1)} \ , \quad i = 1, \ldots, k, \tag{4.7}$$

is not fulfilled in the $(i+1)$th step. In this case, the hypothesis $H_0^{(i+1)}$ and all following null hypotheses are retained and the procedure is stopped. If the inequality 4.7 is valid for all p-values $p_{(1)}, \ldots, p_{(k)}$, all null hypotheses can be rejected.

Both introduced multiple test procedures for the case of dependent tests are very flexible, as they have no demands on the executed tests. The Bonferroni-Holm method possesses the advantage that the probability to reject a sequence of (false) null hypotheses is higher than for the easier Bonferroni correction. The amount of gained power by the application of the procedure proposed by Holm (1979) is dependent on the number of false null hypotheses.

4.3 Discriminant analysis

The discriminant analysis is a method for pattern recognition and separation of data. It is, therefore, a versatile tool, offering both the possibility for descriptive characterisation of group differences and the classification of objects with unknown group membership.

The supervised method differs fundamentally from the cluster analysis, as the division into classes is given a priori, whereas the cluster analysis has the aim of creating classes.

One way of introducing the discriminant analysis is the decision theoretical, probabilistic approach that tries to assign objects according to the expected density of considered variables, commonly assuming a normal distribution. An alternative entry is the classic approach of Fisher that claims little assumptions and is based on the linear combination of data that allows the best separation between the groups respecting its variance-covariance structure. Under certain, later specified conditions the results of both approaches are equivalent.

The following description concentrates on the case of the linear discriminant analysis (LDA) for the existence of two classes, first introducing the general assumptions (Subsection 4.3.1) and giving an overview over different decision rules (Subsection 4.3.2) for the decision theoretic problem (Subsection 4.3.3). After the following presentation of Fisher's approach (Subsection 4.3.4), some concluding remarks to error rate estimations (Subsection 4.3.5) and stepwise selection (Subsection 4.3.6) are given, which are important for the establishment of efficient and reasonably interpretable classification rules.

4.3.1 Assumptions

Concentrating on the description of the discriminant analysis for two classes, the considered population Ω is divided into $k = 2$ disjoint classes and accordingly subpopulations Ω_1 and Ω_2. Each classification object ω is assigned not only the value of a p-dimensional variable vector $\mathbf{x}$, but also the true class membership g, where $\omega \in \Omega_g$, g=1,2.

Requirement for the applicability of the discriminant function is the demand for a sufficiently high sample size N of the training set, that has to exceed the number of variables p, which in turn has to be higher than the number of classes g. For the case of the LDA the different classes are furthermore assumed to possess equal covariance matrices, as quadratic discriminant functions are indicated otherwise.

The aim of the discriminant analysis is the explicit assignment of an object $\omega \in \Omega$ with unknown class index g to one of the groups Ω_g, g=1,2, on the basis of the observed variable vector $\mathbf{x}$. For this, a decision rule e is searched that assigns an estimated class index $\hat{g} \in \{1, 2\}$ to each $\mathbf{x}$ in the sample space $\mathcal{S} \in \mathbb{R}^p$:

$$\begin{aligned} e : \mathcal{S} &\rightarrow \{1, 2\}, \\ \mathbf{x} &\mapsto \hat{g} = e(\mathbf{x}). \end{aligned}$$

The number of wrong decisions of the rule $e(\mathbf{x})$, meaning $\hat{g} \neq g$, should be small for the optimal decision rule, where three different kinds of error rates can be considered. The individual error rate,

$$\varepsilon_{g,\hat{g}}(e) = P\big(e(\mathbf{x}) = \hat{g}|g\big), \quad g \neq \hat{g}, \quad g, \hat{g} = 1, 2,$$

is also denominated as a confusion probability and gives the conditional probability that an object belonging to class g is assigned to class $\hat{g} \neq g$. The conditional error rate,

$$\varepsilon(e|\mathbf{x}) = P\big(g \neq e(\mathbf{x})|\mathbf{x}\big), \quad g = 1, 2, \tag{4.8}$$

coincides with the conditional probability of a wrong decision, if the variable vector $\mathbf{x}$ was observed. Lastly, the (total) error rate,

$$\varepsilon(e) = P\big(e(\mathbf{x}) \neq g\big) = \sum_{g=1}^{2} \sum_{\substack{\hat{g}=1 \\ \hat{g} \neq g}}^{2} \varepsilon_{g\hat{g}}(e)\pi_g = \int_S \varepsilon(e|\mathbf{x}) f(\mathbf{x}) d\mathbf{x}, \tag{4.9}$$

corresponds to the unconditional probability of obtaining a wrong decision based on the decision rule e.[2]

The foundation for the establishment of this rule is the characterisation of $\mathbf{x}$ and g by the a priori probability π_g that $\omega \in \Omega_g$ and the class distribution $f(\mathbf{x}|g)$ of $\mathbf{x}$ in Ω_g, when $\mathbf{x}$ and g are interpreted as random variables. The mixed distribution of the a priori probability and the conditional distribution of $\mathbf{x}$ for given g results in the unconditional distribution of $\mathbf{x}$ on Ω:

$$f(\mathbf{x}) = \sum_{g=1}^{2} \pi_g f(\mathbf{x}|g).$$

Based on the a priori probability, the class distribution, and this unconditional distribution, the probability that an object with observed $\mathbf{x}$ belongs to class g can be determined. It is denominated as the a posteriori probability which is of special interest for classification problems and can be calculated after Bayes's theorem as

$$p(g|\mathbf{x}) = \frac{\pi_g f(\mathbf{x}|g)}{f(\mathbf{x})}, \quad g = 1, 2.$$

4.3.2 Decision rules

On the basis of the introduced probabilities and distributions, several decision rules can be introduced that are optimal in different respects.

[2]For discrete distributions $f(\mathbf{x})$ the integral in the right part of Equ. 4.9 has to be substituted by a sigma sign.

Firstly, the intuitive Bayes decision rule is presented, which makes a decision on the assignment of an object for the class $\hat{g}$ for which the probability of the class membership is maximal for given $\mathbf{x}$:

Definition 4.2 According to the *Bayes decision rule*, $e(\mathbf{x}) = \hat{g}$ assigns an object with variable value $\mathbf{x}$ to class $\hat{g}$ if

$$p(\hat{g}|\mathbf{x}) \geq p(\tilde{g}|\mathbf{x}) \qquad \text{respectively} \qquad \pi_{\hat{g}} f(\mathbf{x}|\hat{g}) \geq \pi_{\tilde{g}} f(\mathbf{x}|\tilde{g}) \qquad \text{for} \quad \hat{g}, \tilde{g} = 1, 2.$$

An object is, therefore, assigned to the class possessing the highest a posteriori probability for the observed vector $\mathbf{x}$. For the comparison with other decision rules that are introduced later on, the consideration of the transformation

$$\frac{f(\mathbf{x}|\hat{g})}{f(\mathbf{x}|\tilde{g})} \geq \frac{\pi_{\tilde{g}}}{\pi_{\hat{g}}} \ , \quad \hat{g}, \tilde{g} = 1, 2,$$

of the Bayes decision rule is beneficial.

The Bayes rule offers the smallest total error rate amongst all decision rules and is optimal on this note. This fact can be shown by examination of the conditional error rate $\varepsilon(e|x)$ in Equ. 4.8, which is now expressed by the a posteriori probabilities:

$$\varepsilon(e|\mathbf{x}) = 1 - P\big(g = e(\mathbf{x})|\mathbf{x}\big) = 1 - p\big(e(\mathbf{x})|\mathbf{x}\big) \ , \quad g = 1, 2.$$

As the conditional error rate of the Bayes rule for given $\mathbf{x}$ is never greater than that of another rule, the integral of the conditional error rates weighted by $f(\mathbf{x})$, which coincides with the total error rate, is also minimal.[3]

For the deduction of another decision rule, the class membership g is not seen as a random variable, but as an unknown parameter which characterises the distribution of $\mathbf{x}$. Furthermore, the distribution attributing the highest probability to the observed value $\mathbf{x}$ is searched according to the maximum likelihood principle of parameter estimation:

Definition 4.3 According to the *maximum likelihood decision rule*, $e(\mathbf{x}) = \hat{g}$ assigns an object with variable value $\mathbf{x}$ to class $\hat{g}$, if

$$f(\mathbf{x}|\hat{g}) \geq f(\mathbf{x}|\tilde{g}) \qquad \text{respectively} \qquad \frac{f(\mathbf{x}|\hat{g})}{f(\mathbf{x}|\tilde{g})} \geq 1 \qquad \text{for } \hat{g}, \tilde{g} = 1, 2.$$

[3] For discrete distributions $f(\mathbf{x})$ the integral has to be substituted by the weighted sum.

A decision is, therefore, made in favour of the class $\Omega_{\hat{g}}$, $\hat{g} = 1, 2$, that possesses the highest value of the likelihood function $L(\hat{g}; \mathbf{x}) \equiv f(\mathbf{x}|\hat{g})$ for the observed variable vector $\mathbf{x}$. This decision rule can be seen as a special case of the Bayes rule, occurring if equal a priori probabilities π_1=π_2 are assumed, and is, therefore, optimal in this case. For this reason it also seems to be a reasonable choice if no information on π_g, $g = 1, 2$, is available.

While the two introduced decision rules are optimal in the sense of the minimisation of the error rate judging misclassification, independent of the group an object belongs or is assigned to, the cost of a misclassification often differs between the groups in practical applications. For example, it is usually a more substantial error if a medical diagnosis for severe diseases judges a patient as healthy so that the disease will not be treated, as if a healthy person is classified as ill and gets a needless therapy. This differential evaluation of misclassifications can be incorporated by the enhancement of decision rules by a cost function:

Definition 4.4 Let $C(g, \hat{g})$ be the costs incurring if g is the true class index of object ω and the decision $\hat{g}$ is made. According to the *cost optimal decision rule*, $e(\mathbf{x}) = \hat{g}$ assigns an object with variable value $\mathbf{x}$ to class $\hat{g}$, if

$$\pi_{\hat{g}} C(g, \tilde{g}) f(\mathbf{x}|\hat{g}) \geq \pi_{\tilde{g}} C(g, \hat{g}) f(\mathbf{x}|\tilde{g}) \quad \text{respectively} \quad \frac{f(\mathbf{x}|\hat{g})}{f(\mathbf{x}|\tilde{g})} \geq \frac{\pi_{\tilde{g}} C(g, \hat{g})}{\pi_{\hat{g}} C(g, \tilde{g})}, \quad g, \hat{g}, \tilde{g} = 1, 2.$$

An object is thus assigned to the class for which minimal costs incur with observed variable vector $\mathbf{x}$.

Two important special cases of the cost function are the simple symmetric cost function and the inversely proportional cost function. The simple symmetric cost function,

$$C_s(g, \hat{g}) = \begin{cases} 0 & \text{for} \quad g = \hat{g} \\ C > 0 & \text{for} \quad g \neq \hat{g} \end{cases}, \quad g, \hat{g} = 1, 2,$$

evaluates each misclassification with the same cost. The according cost optimal decision rule is the Bayes rule. The inversely proportional cost function,

$$C_p(g, \hat{g}) = \begin{cases} 0 & \text{for} \quad g = \hat{g} \\ C/\pi_g & \text{for} \quad g \neq \hat{g} \end{cases}, \quad g, \hat{g} = 1, 2, \tag{4.10}$$

strongly increases the cost for a wrong assignment of objects belonging to classes with small a priori probabilities. As the a priori probabilities usually express differences in the size of the considered populations, the cost function $C_p(g, \hat{g})$ appears to be reasonable, for example in medical diagnoses for rare, but severe diseases. The cost optimal decision rule for the inversely proportional cost function is the maximum likelihood rule. Under

certain conditions, both the Bayes and the maximum likelihood decision rule, therefore, also possess the property of cost optimality.

The different decision rules correspond with specific discriminant functions:

$$\begin{aligned} d_g(\mathbf{x}) &:= \pi_g f(\mathbf{x}|g) && \text{results in the Bayes rule,} \\ d_g(\mathbf{x}) &:= f(\mathbf{x}|g) && \text{in the maximum likelihood rule, and} \\ d_g(\mathbf{x}) &:= -C(g|\mathbf{x}) && \text{in the cost optimal rule for arbitrary cost functions,} \\ & && g = 1, 2. \end{aligned}$$

For the assignment of an object ω with observed variable vector $\mathbf{x}$ to a class $\Omega_{\hat{g}}$, the values d_1 and d_2 of the discriminant functions $d_1, d_2 : \mathcal{S} \to \mathbb{R}$ are determined and $\hat{g} = e(\mathbf{x})$ is chosen so that

$$d_{\hat{g}}(\mathbf{x}) \geq d_{\tilde{g}}(\mathbf{x}) \ , \quad \hat{g}, \tilde{g} = 1, 2.$$

As the classification rules are invariant concerning transformations with strictly monotonic increasing functions, adequately chosen transformations of the original discriminant functions $d_1(\mathbf{x})$ and $d_2(\mathbf{x})$ can yield considerable simplifications without an influence on the classification result. Furthermore, for the case of two classes a switch to a single discriminant function d is possible with

$$d(\mathbf{x}) = d_1(\mathbf{x}) - d_2(\mathbf{x}),$$

where objects with the variable vectors $\mathbf{x} \in \mathcal{S}$ with $d(\mathbf{x}) \geq 0$ are assigned to class Ω_1, all others to class Ω_2.

The following descriptions are based on the use of the maximum likelihood rule, as this proved to be superior for the problems addressed in this work, especially under the aspect of cost optimisation.

4.3.3 Discriminant analysis for normal distribution

For classification problems based on quantitative data, the discriminant analysis is in many cases applied with the assumption of an underlying normal distribution, which is often justified because of the central limit theorem and the flexible statistical properties of the normal distribution. The class distribution of the variable vector $\mathbf{x}$ is then

$$f(\mathbf{x}|g) = \frac{1}{(2\pi)^{p/2}(\det \mathbf{\Sigma}_g)^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \mu_g)^T \mathbf{\Sigma}_g^{-1}(\mathbf{x} - \mu_g)\right\}, \quad g = 1, 2,$$

where μ_g and $\mathbf{\Sigma}_g$ denote the class specific expected values and covariance matrices.

Utilising the invariance concerning strictly monotonic increasing transformations, the maximum likelihood rule can be simplified to

$$d_g(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_g)^T\mathbf{\Sigma}_g^{-1}(\mathbf{x}-\mu_g) - \frac{1}{2}\ln(\det\mathbf{\Sigma}_g), \qquad g = 1,2, \tag{4.11}$$

by taking the natural logarithm and neglecting the common additive term $-\frac{p}{2}\ln(2\pi)$. If classwise identical covariance matrices are assumed, the discriminant function 4.11 can be further simplified, as $\det\mathbf{\Sigma}_g$ and the term $-\frac{1}{2}\mathbf{x}'\mathbf{\Sigma}^{-1}\mathbf{x}$ which are independent from g can be omitted, yielding the linear discriminant functions

$$d_g(\mathbf{x}) = \mu_k^T\mathbf{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\mu_k^T\mathbf{\Sigma}^{-1}\mu_k\ , \qquad g = 1,2.$$

For the special case of two classes the linear discriminant function is, therefore,

$$d(\mathbf{x}) = d_1(\mathbf{x}) - d_2(\mathbf{x}) = \left[\mathbf{x} - \frac{1}{2}(\mu_1+\mu_2)\right]^T \mathbf{\Sigma}^{-1}(\mu_1-\mu_2),$$

resulting in the assignment of an object ω with the observed variable vector $\mathbf{x}$ to class Ω_1 if $d(\mathbf{x}) \geq 0$, and its categorisation as member of class Ω_2 if $d(\mathbf{x}) < 0$.

If the underlying normal distributions of the data are unknown, the corresponding parameters have to be estimated on the base of a training set for which both variable vector and true class memberships are available. For this, the common unbiased estimators of arithmetic mean $\hat{\mu}_g = \bar{\mathbf{x}}_g$ and empirical covariance $\hat{\mathbf{\Sigma}}_g = \mathbf{S}_g^2$ for observations from class Ω_g, g=1,2, can be used. For the case of classwise identical covariance matrices, an unbiased estimator for the interclass parameter $\mathbf{\Sigma}$ is the pooled covariance matrix,

$$\mathbf{S}_P^2 = \frac{1}{N-2}\sum_{g=1}^{2}\sum_{n=1}^{n_g}(\mathbf{x}_{gn}-\bar{\mathbf{x}}_g)(\mathbf{x}_{gn}-\bar{\mathbf{x}}_g)^T, \tag{4.12}$$

where $\mathbf{x}_{gn}$, $n = 1,...,n_g$, denominates the observations in class g.

The linear decision rule can, therefore, be defined by the assignment of an object ω with variable vector $\mathbf{x}$ to class Ω_1, if and only if

$$\left\{\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2)\right\}^T \mathbf{a} \geq 0 \qquad \text{with} \quad \mathbf{a} = {\mathbf{S}_P^2}^{-1}(\bar{\mathbf{x}}_1-\bar{\mathbf{x}}_2). \tag{4.13}$$

4.3.4 Fisher's linear discriminant analysis

The decision theoretical considerations introduced in the previous subsection are based on the assumption of a normal distribution. The prior approach of Fisher (1936) goes without this assumption and concerns the case of unknown parameters.

Coming from the classwise data matrices,

$$\mathbf{X}_1^T = (\mathbf{x}_{11}, \ldots, \mathbf{x}_{1n_1}) \quad \text{and} \quad \mathbf{X}_2^T = (\mathbf{x}_{21}, \ldots, \mathbf{x}_{2n_2}),$$

containing the variable vectors of a training set, it is aimed to find a linear combination

$$y = \mathbf{a}'\mathbf{x} \ , \quad \mathbf{a} = (a_1, \ldots, a_p)^T,$$

of the observations of the variable vector $\mathbf{x}$ that reflects the group structure in one dimension. For this, the vector $\mathbf{a}$ that maximises the variation between the groups and minimises the variation within the groups is investigated.

These requirements result in the ratio

$$R(\mathbf{a}) = \frac{(\bar{y}_1 - \bar{y}_2)^2}{s_1^2 + s_2^2}, \tag{4.14}$$

which is maximal for the required $\mathbf{a}$. The difference of the classwise arithmetic y means $\bar{y}_g = \mathbf{a}^T\bar{\mathbf{x}}_g$, g=1,2, can be seen as the distance between the groups, while the variation within the groups is composed of the sum of squared deviations $s_g^2 = \sum_{n=1}^{n_g}(y_{gn} - \bar{y}_g)^2$, g=1,2, in both classes.

The ratio $R(\mathbf{a})$ can be rearranged as

$$R(\mathbf{a}) = \frac{(\bar{y}_1 - \bar{y}_2)^2}{\mathbf{a}^T\mathbf{W}\mathbf{a}} \qquad \text{with} \qquad \mathbf{W} = (n_1 + n_2 - 2)\mathbf{S}^2 = \sum_{g=1}^{2}\sum_{n=1}^{n_g}(\mathbf{x}_{gn} - \bar{\mathbf{x}}_g)^2,$$

as it is

$$\begin{aligned} s_1^2 + s_2^2 &= \sum_{g=1}^{2}\sum_{n=1}^{n_g}(y_{gn} - \bar{y}_g)^2 = \sum_{g=1}^{2}\sum_{n=1}^{n_g}(\mathbf{a}^T\mathbf{x}_{gn} - \mathbf{a}^T\bar{\mathbf{x}}_g)^2 \\ &= \mathbf{a}^T\left(\sum_{g=1}^{2}\sum_{n=1}^{n_g}(\mathbf{x}_{gn} - \bar{\mathbf{x}}_g)(\mathbf{x}_{gn} - \bar{\mathbf{x}}_g)^T\right)\mathbf{a} \ . \end{aligned}$$

For the maximisation of $R(\mathbf{a})$, $\mathbf{a}$ is differentiated, yielding the necessary condition

$$\frac{\partial R(\mathbf{a})}{\partial \mathbf{a}} = \frac{2(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\mathbf{a}^T\mathbf{W}\mathbf{a} - 2\mathbf{W}\mathbf{a}(\mathbf{a}^T\bar{\mathbf{x}}_1 - \mathbf{a}^T\bar{\mathbf{x}}_2)}{(\mathbf{a}^T\mathbf{W}\mathbf{a})^2} = 0, \tag{4.15}$$

for the extreme value which can be transformed to

$$\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2 = \mathbf{W}\mathbf{a}\left(\frac{\mathbf{a}^T\bar{\mathbf{x}}_1 - \mathbf{a}^T\bar{\mathbf{x}}_2}{\mathbf{a}^T\mathbf{W}\mathbf{a}}\right).$$

It is, therefore,

$$\mathbf{a} = \mathbf{W}^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2),$$

if the factor of proportionality $\frac{\mathbf{a}^T\bar{\mathbf{x}}_1 - \mathbf{a}^T\bar{\mathbf{x}}_2}{\mathbf{a}^T\mathbf{W}\mathbf{a}}$, which is constant for given $\mathbf{a}$ and does, therefore, not influence the direction of $\mathbf{a}$, is neglected. This vector $\mathbf{a}$ from Fisher's LDA $y = \mathbf{a}^T\mathbf{x}$ is hence equal to the vector $\mathbf{a}$ in inequality 4.13 except for a multiplicative factor, since $\mathbf{W} = (n_1 + n_2 - 2)\mathbf{S}^2$.

The classification of object ω with observed variable vector $\mathbf{x}$ and unknown class index is performed by examining the distance of $y = \mathbf{a}'\mathbf{x}$ to the two class means $\bar{\mathbf{y}}_1$ and $\bar{\mathbf{y}}_2$, where the object is assigned to class Ω_1 if

$$|y - \bar{y}_1| < |y - \bar{y}_2| \qquad \Leftrightarrow \qquad y > \frac{1}{2}(\bar{y}_1 - \bar{y}_2).$$

This results in the decision rule that assigns an object ω with the observed variable vector $\mathbf{x}$ to class Ω_1, if and only if

$$\left\{\mathbf{x} - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\right\}^T \mathbf{a} \geq 0.$$

Comparing Fishers linear discriminant function with inequality 4.13, shows that the same classification result is obtained as for the maximum likelihood rule that was introduced in subsection 4.3.3. While the latter, however, requires classwise normal distribution with equal covariance matrices, Fisher's approach offers the same rule on the basis of a reasonable distributionless criterion. This indicates a relatively robust behaviour of the LDA against violations of its assumptions and a capability in a broad range of situations.

4.3.5 Error rate estimation

For the evaluation of the quality of a discriminant function, the error rate can be considered which usually has to be estimated on the basis of the data.

The easiest method to estimate the error rate is the determination of the resubstitution error rate, conforming with the ratio of falsely classified training objects. As this estimation considers the same data for the determination of the error rate that were used for the establishment of the decision rule, the resubstitution error rate underestimates the true error rate in most cases.

To circumvent this negative effect, the sample set is often divided into training and test samples. The objects of the test sample are then classified on the basis of the decision rule that was created with the training sample and the amount of misclassification is determined, giving an estimation of the total error rate. This method is, however, only

recommended if large sample sets are available, as the training sample is otherwise to small to generate a meaningful decision rule.

An alternative that can also be used if smaller sample sets are considered is the leave-one-out method (Lachenbruch and Mickey, 1968), constituting a special case of the cross validation. The cross validation departs the complete sample set repeatedly into training and test sets to avoid a bias of the estimation by a single division, and averages the obtained error rates. For the leave-one-out method, which is also denominated as the jackknife method, only one object is put into the test set at a time. On the basis of the remaining $N - 1$ objects, a classification rule is then determined, which is used for the assignment of the excluded object to one of the classes Ω_1 and Ω_2. This is done for all elements before the error rate is estimated by the ratio of misclassified objects, yielding a robust, and in contrast to the resubstitution error rate, nearly unbiased estimator of the expected actual error rate. Furthermore, it is beneficial that the leave-one-out method can be applied for the LDA without complete recalculation for each object (Ripley, 1996). For small N the estimator of the leave-one-out method, however, yields possibly high variances.

4.3.6 Stepwise variable selection

For several reasons it is advantageous not to include all available variables in the construction of a classification rule. On the one hand it is desirable to exclude unimportant variables in the establishment of a decision rule to ensure that only really necessary variables have to be observed for the future assignment of objects, while at the same time a representation based on only the most relevant variables allows an easier interpretation of the decision rule. On the other hand, also methodical reasons speak for the explanation of the characteristic structure of the underlying classes with preferably few variables, as then a smaller number of parameters have to be estimated, yielding a higher quality of the estimation.

The variable selection is furthermore motivated by a special behaviour of the (expected) actual error rate: in contrast to the optimal error rate of the Bayes rules with known parameters, it can occur that the actual error rate increases when another variable is added to the model, although the theoretical error rate improves if this variable can actually contribute to the class separation. This behaviour is caused by the raised complexity of the model, as the number of parameters that have to be estimated increases. If the additional variable contains too little discriminatory information, the class separation,

therefore, gets worse. For estimated decision rules the error rate usually decreases in the beginning when variables are added successively, but rises again after a certain number of variables. This phenomenon is denominated as the bath tub effect.

To find the optimal linear discriminant function concerning the error rate, a complete model search should be ideally performed, which is, however, very demanding for high numbers of variables. It is alternatively possible to arrange a stepwise choice by successively adding or discarding variables in a forward or backward procedure.

In the case of a stepwise backward elimination, starting with the discriminant function which is based on all available variables, the error rate can be used as the optimisation criterion for variable selection. For each step the error rates of all discriminant functions that occur if one of the variables is excluded are estimated. If the discriminant function obtained without one of the examined variables offers a lower error rate as the best function of the previous step, the corresponding variable is eliminated. If the error rate decreases for more than one of the constructed discriminant function, the variable with the lowest error rate is chosen. If the minimal observed error rate of a new step increases, the procedure is stopped. In many cases, however, it is recommended to perform one more step to check this tendency. The error rate of the described proceeding can, for example, be estimated by the leave-one-out method.

It is possible that more than one variable shows the same maximum improvement of the error rate when excluded from the model, which makes it unclear as to which variable should be eliminated. Instead of making a random choice, the standardised discriminant coefficients a_j^* can be determined, standardising the original discriminant coefficients with the pooled standard deviation $s_{P_{jj}}^{\frac{1}{2}}$ of the corresponding variable x_j:

$$a_j^* = a_j \; s_{P_{jj}}^{\frac{1}{2}} \; , \quad j = 1, \ldots, p. \tag{4.16}$$

In this way, the influences of potentially different magnitudes or units of the variables on the height of the discriminant coefficients are eliminated. The weight of the different variables on the separation of classes and the classification is, therefore, directly apparent, and in the case of doubt when the same value of the minimal estimated error rate appears for more than one variable, the one with the lowest influence on the separation of classes corresponding to the lowest standardised discriminant coefficient can be excluded.

Spectra processing

Chapter 5

Preprocessing of ion mobility spectra

Instrumental and environmental factors, which can influence both drift and retention time, and an occuring baseline shift hinder the comparability of different measurements. In addition, strong RIP tailing and high levels of noise corrupt general peak clarity, which makes an efficient preprocessing strategy for MCC/IMS data necessary. Consequently the development of specific axes transformations allows a better alignment in the different measurement dimensions (Section 5.1). Fitting a lognormal function to the strong tailing of the RIP (Section 5.2) and using the discrete wavelet transform for data compression and denoising (Section 5.3) yields the successful elimination of the RIP tailing, a data reduction to a quarter or less, and a significant increase in the signal-to-noise ratio. In conclusion, the developed preprocessing strategy offers the desired outcome of smooth peaks lying on a common base level.

5.1 Comparability of measurements

Varying environmental and instrumental factors such as ambient pressure, temperature, or electric field strength, can affect the characteristic drift time of analyte peaks and, therefore, complicate the comparison of measurements. Due to deviations of the column temperature from the standard value of 30 °C a similar effect occurs, influencing the retention time of peak positions; and a baseline shift distorts the comparability in the dimension of signal intensity.

A better alignment of different measurements was therefore sort, using different kinds of transformations for each axis of the three-dimensional data structure. For the drift time, an enhanced version of the common reduced mobility was found to yield reproducible

results (Subsection 5.1.1), while a quadratic function was chosen to balance deviations in the retention time (Subsection 5.1.2), and a baseline correction, ensuring the variation of intensity values around zero in noise areas, was executed (Subsection 5.1.3).

5.1.1 Reproducible inverse reduced mobility

Differing measurement conditions, such as ambient pressure and temperature, can be compensated by the transformation of the drift time values to the reduced mobility K_0. At the same time, this term adjusts parameter values of the instrumentation, such as electric field strength and the drift tube length, allowing the comparability of measurements from instruments with a different setup.

Before transforming to reduced mobility, the drift times x_i, $i = 1, \ldots, n_D$, had to be shifted by the half opening time t_{grid} of the shutter grid to obtain peaks varying around the actual drift time $x^* = x - \frac{t_{grid}}{2}$ of an analyte. The ion mobility K was then calculated by the equation

$$K = \frac{l_d}{x^* \; E} \cdot 1000,$$

where l_d denotes the length of the drift tube (cm) and E is the electric field strength (V/cm). [1]

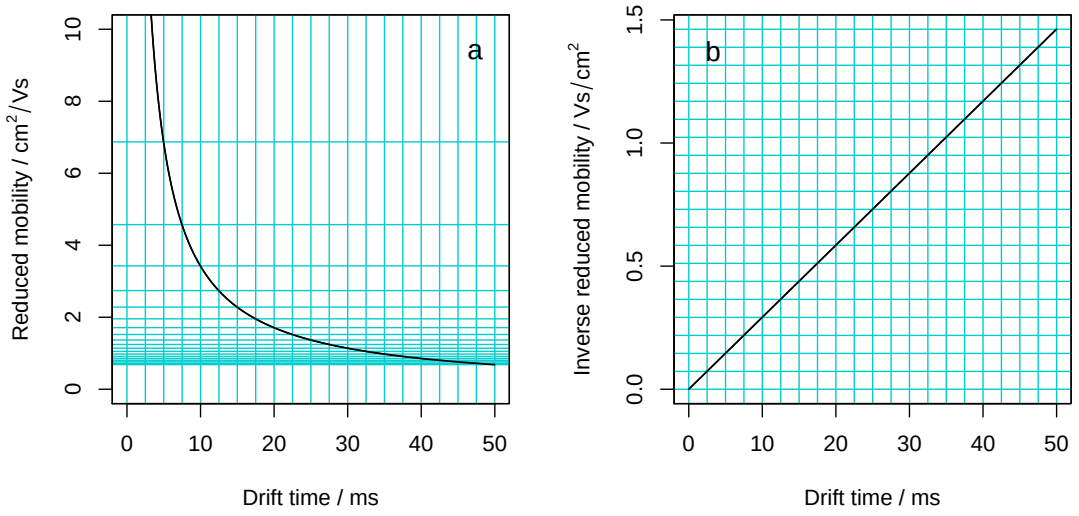

Figure 5.1: Drift time values in relation to reduced mobility (a) without and (b) with inversion: the linear relationship of the inversed values allows for an easier processing and interpretation.

[1] The factor 1000 is necessary to convert the unit of drift time from ms to s.

Adjusting for ambient pressure p (hPa) and temperature T (K), the reduced mobility K_0 is following determined by

$$K_0 = K \cdot \frac{p}{p_0} \cdot \frac{T_0}{T},$$

where the standard pressure $p_0 = 1013.25$ hPa and temperature $T_0 = 273.15$ K are included for normalisation.

The consideration of the inverse reduced mobility

$$\frac{1}{K_0} = x^* \cdot \frac{E}{1000\ l_d} \cdot \frac{p_0\ T}{p\ T_0}$$

instead of K_0, was found to be even more beneficial, as it yielded a linear relationship with the drift time, making values easier to process and interpret (Fig. 5.1).

As a large amount of variation in IMS peak positions persisted even after calculation of the inverse reduced mobility in the current investigations (Fig. 5.2), a reproducible version of the inverse reduced mobility $1/K_0^r$ was developed, leading to a better alignment of spectra. For this, the RIP was used as a basing point in the drift dimension, since it appeared in

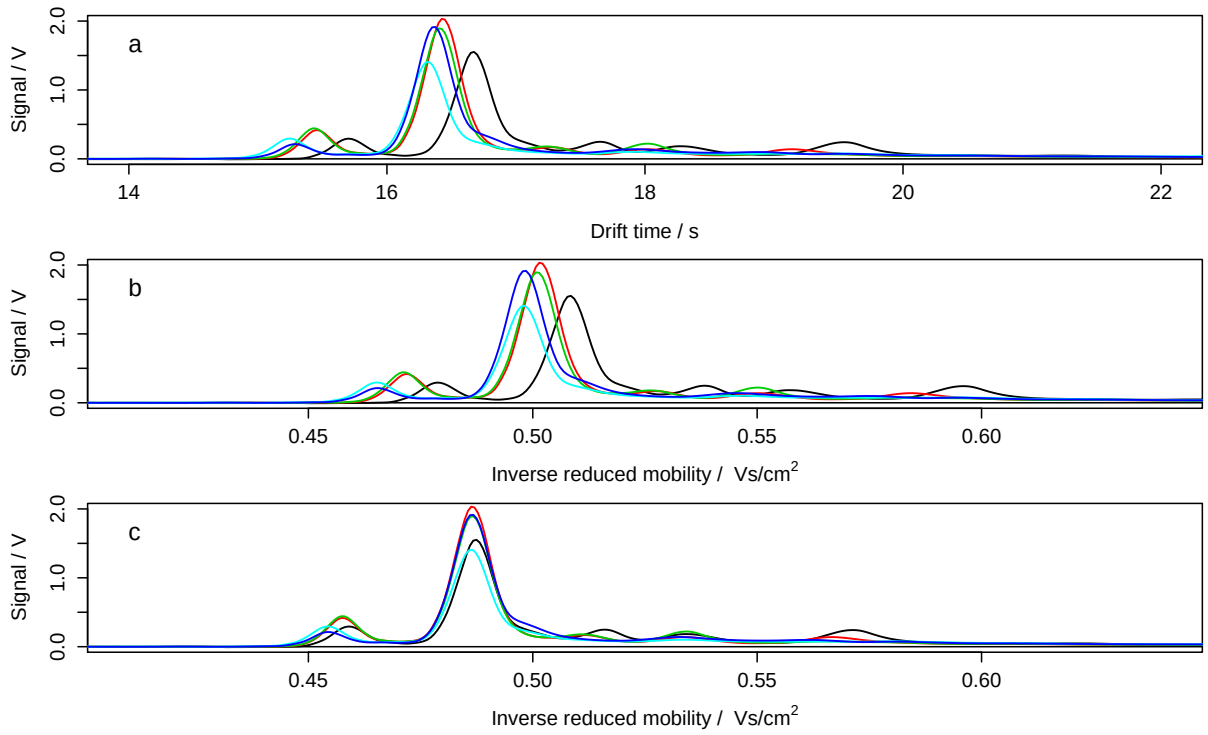

Figure 5.2: Spectra alignment, shown for instance IMS breath spectra with different measurement conditions, for (a) original drift time, (b) inverse reduced mobility, and (c) reproducible inverse reduced mobility: the reproducible inverse reduced mobility shows a considerable better alignment than the two other transformations.

all the IMS spectra with high intensities, and was thus easy to identify. Due to the linear relationship, the reproducible inverse reduced mobility could be determined by a factor of the actual drift times determined by the fraction of the "true" inverse reduced mobility of the RIP (0.49 Vs/cm^2) and the position x^*_{RIP} of the RIP maximum in the drift time:

$$\frac{1}{K_0^r} = \frac{0.49}{x^*_{RIP}} x^*.$$

This term resulted in a more precise alignment of IMS spectra compared to original drift times or the common (inverse) reduced mobility (Fig. 5.2).

The benefit of this method can be explained by the absence of measurement errors and possibly the indirect inclusion of unknown influencing variables, as all the factors affecting the drift time of sample analytes should be inherent in the position of the RIP and can thus be transferred to the rest of the spectra by the determined factor.

In addition to the advantages of the improved alignment using the reproducible inverse reduced mobility, time and effort in daily laboratory work could be reduced, since the detailed reporting of measurement conditions is now no longer necessary. Therefore, the reproducible inverse reduced mobility is used now as a standard at the laboratories at ISAS - Institute for Analytical Sciences, Dortmund.

In the following, the notation $\mathbf{x}$ will be used to denote the vector of the reproducible inverse reduced mobilities for easier notation and clarity.

5.1.2 Corrected retention time

The retention time of an analyte is influenced by the temperature of the MCC. This temperature was kept constant at 30 °C for most measurements considered in this work; in some cases, however, this optimal condition could not be established.

With increased column temperatures, sample analytes pass faster through the MCC, reaching the IMS earlier, and peaks thus appear at lower retention times compared to standard conditions. To eliminate this influence and allow for a better comparability of measurements in the retention time dimension, these peaks were artificially shifted backwards.

Since there is no basing point, such as the RIP maximum position in drift time, for the dimension of retention time, a general formula adjusting deviations from the optimal column temperature of the standard value of 30 °C was determined on the basis of analyte measurements for four substances, at three different column temperatures.

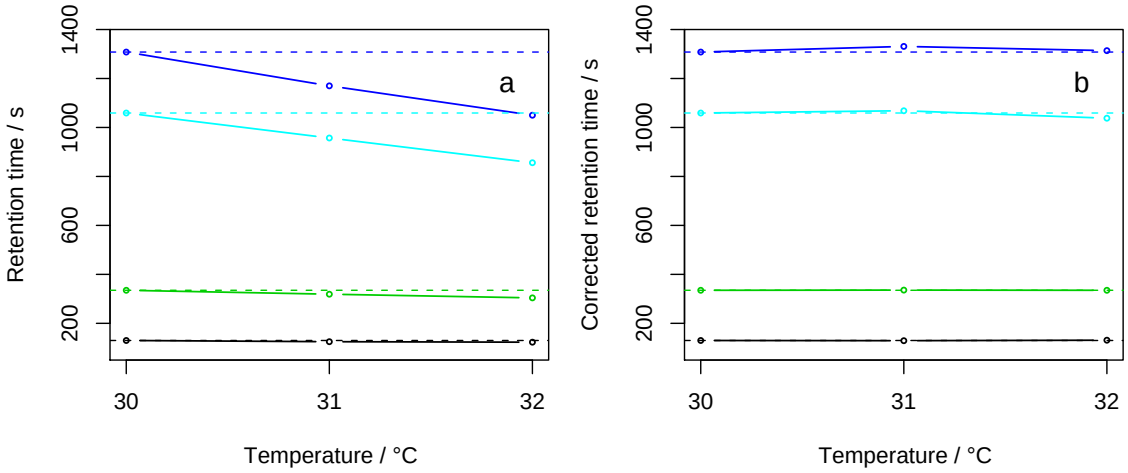

Figure 5.3: Effect of changing column temperature on the retention time (a) before and (b) after adjustment: the corrected retention time shows more constant values for all four characteristic analyte peak maxima at three different column temperatures, especially in the relevant retention time range of up to 600 s.

As the deviation between measured retention time y and expected standard retention time y^{30} grows not only with increasing temperature, but also with increasing characteristic position of an analyte, the behaviour of the position shift of a peak for growing column temperature could not be described in a linear way. Instead, a quadratic function of retention time y was chosen to correct this quantity in dependency of the deviation in column temperature ΔT from 30 °C described by the equation

$$y^{30} = \Delta T \; k_q \; y^2 + (1 + k_T \; \Delta T) \; y.$$

By minimising the absolute deviations between corrected and expected retention time at standard column temperature (30 °C), the optimal solution was found for the factors $k_T = 0.02$ and $k_q = 10^{-4}$.

The optimised transformation yielded improvements for all analytes and temperatures (Fig. 5.3), especially for the two analytes in the retention time range of up to 600 s considered in the current investigations. This result encouraged the application of the determined transformation in all subsequent analyses, where the notation $\mathbf{y}$ will refer to the vector of corrected retention times.

5.1.3 Baseline correction

In the third dimension of MCC/IMS measurements, the signal intensity values of measurement parts containing no analyte peaks but pure noise should vary around zero. This

is, however, not the case for the majority of IMS measurements; instead IMS spectra series are mainly shifted in total by a varying constant in the dimension of signal intensity.

This baseline shift had to be adjusted to allow comparisons of peak heights between different measurements, and was estimated by a constant a_{BSL}, calculated as the mean intensity in a measurement part of pure noise as

$$a_{BSL} = \sum_{i=i_1}^{i_2} \sum_{j=j_1}^{j_2} \frac{z_{ij}}{(i_2 - i_1 + 1)\ (j_2 - j_1 + 1)}. \tag{5.1}$$

For the dimension of the inverse reduced mobility, the two indices defining the noise part were chosen as $i_1 = \lfloor 0.1 i_{RIP} + 0.5 \rfloor$ and $i_2 = \lfloor 0.8 i_{RIP} + 0.5 \rfloor$ with $x_{i_{RIP}} = 0.49$, as no peaks were observed in this area. In the retention time dimension, only the first few spectra were excluded, as these were defective in many cases, resulting in the indices $j_1 = 10$ and $j_2 = n_R$.

The baseline correction was then performed by subtracting the absolute value a_{BSL} from the total data matrix of each measurement to yield the corrected matrix $\mathbf{S}^*$ as

$$\mathbf{S}^* = \mathbf{S} - a_{BSL}\ \mathbf{I}_{n_D, n_R}\ ,$$

where $\mathbf{I}_{n_D, n_R}$ is the $n_D \times n_R$ matrix with all elements equal to 1.

Although the constant a_{BSL} was determined only in the initial part of the spectra, this absolute baseline correction worked well in all parts of the measurements (Fig. 5.4) and was, therefore, subsequently applied as a standard, where the notations $\mathbf{s}$ and $\mathbf{S}$ will be used to denote the baseline corrected spectra and spectra series, respectively.

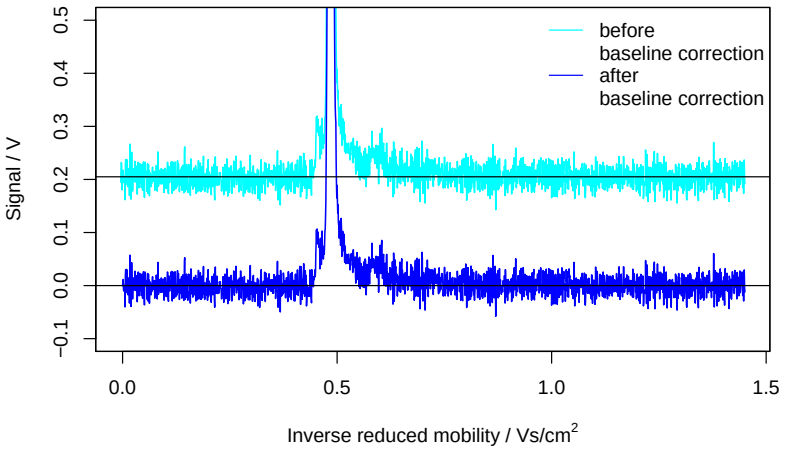

Figure 5.4: Effect of baseline correction for a single spectrum: although the baseline shift was estimated based only on the initial measurement part, the correction achieves variation around zero in noise areas for the whole spectrum.

5.2 Adjustment of the reactant ion peak tailing

Variations in the ion velocity, due to random ion-molecule reactions occurring in the drift tube, cause a strong tailing of the RIP and varying heights for peaks in different parts of the drift time axis (Section 2.2). This results in the need for clarification of IMS measurements by fitting a detailing function which describes the tailing (Subsection 5.2.1) and is subsequently subtracted from each spectrum of the entire spectra series (Subsection 5.2.2).

5.2.1 Detailing function

The detailing function, fitted to the tailing of a considered IMS spectrum $\mathbf{s}$, was chosen as a modified lognormal function of the form

$$L(x_i) = \frac{a_d}{(x_i - \theta)\ \sigma_d\ \sqrt{2\pi}} \cdot \exp\left[-\frac{\left[\log\left(\frac{x_i-\theta}{m}\right)\right]^2}{2\sigma_d^2} \right],$$

where x_i, $i = 1, \ldots, n_D$, denotes the elements of the vector $\mathbf{x}$ of reproducible inverse reduced mobility values. The lognormal function met well the general assumption of Gaussian peaks as well as the right-screwed shape of the RIP tailing and is often used to describe physical processes that are limited in one direction. The variables θ, σ_d, and m are parameters of location, shape, and scale, respectively, whereas the factor a_d induces a shrinkage to the actual intensity magnitude.

As a reasonable side condition it was claimed that the maximum positions $x_{\max}$ of the spectrum and $x_{\max}^{log}$ of the detailing function $L(\mathbf{x})$ should coincide. Since the maximum position of a lognormal function is known to be $x_{\max}^{log} = \theta + \frac{m}{\exp(\sigma_d^2)}$, the position parameter θ could be determined by

$$x_{\max} \stackrel{!}{=} \theta + \frac{m}{\exp(\sigma_d^2)} \quad \Leftrightarrow \quad \theta \stackrel{!}{=} x_{\max} - \frac{m}{\exp(\sigma_d^2)}.$$

Accordingly, only the three parameters of shape σ_d, scale m, and shrinkage a_d had to be optimised, achieved by the minimisation of the developed penalty term P,

$$P = \sum_{i=1}^{n_D} \left[I_{[\Delta_i < -r_p]} p_{abs} + I_{[\Delta_i > r_p]} \min(\Delta_i, b_p) \right],$$

for $\Delta_i = s_i - L(x_i)$. Whilst the scalar r_p, dependent on the standard deviation in noise areas, allows for little variation of the adjusted function around the considered spectrum, the constant p_{abs} assigns an absolute penalty for parts of the detailing function lying

over the spectrum **s**, to yield the detailing function nestling to the data from below, irrespective of occurring peaks. Downwards deviation was penalised by the actual deviation, constrained by the threshold b_p giving a cut-off point for diminishing the influence of the RIP height in so much as deviations greater than this value were punished only by b_p.

For minimisation of this penalty term, yielding the optimal parameter set, a limited-memory modification of a quasi-Newton method was used, allowing the choice of box constraints for each variable (Byrd et al., 1995).

5.2.2 Application of the detailing function

To ensure the comparability between spectra of the same measurement after detailing, and to reduce the computational cost by the performance of only a single fitting step, the adjustment of the detailing function was performed for a representative spectrum. The median spectrum $\mathbf{s}^{med}$ with

$$\mathbf{s}^{med} = (\mathrm{med}(\mathbf{s}_{1.}), \mathrm{med}(\mathbf{s}_{2.}), \ldots, \mathrm{med}(\mathbf{s}_{n_D.}))^T ,$$

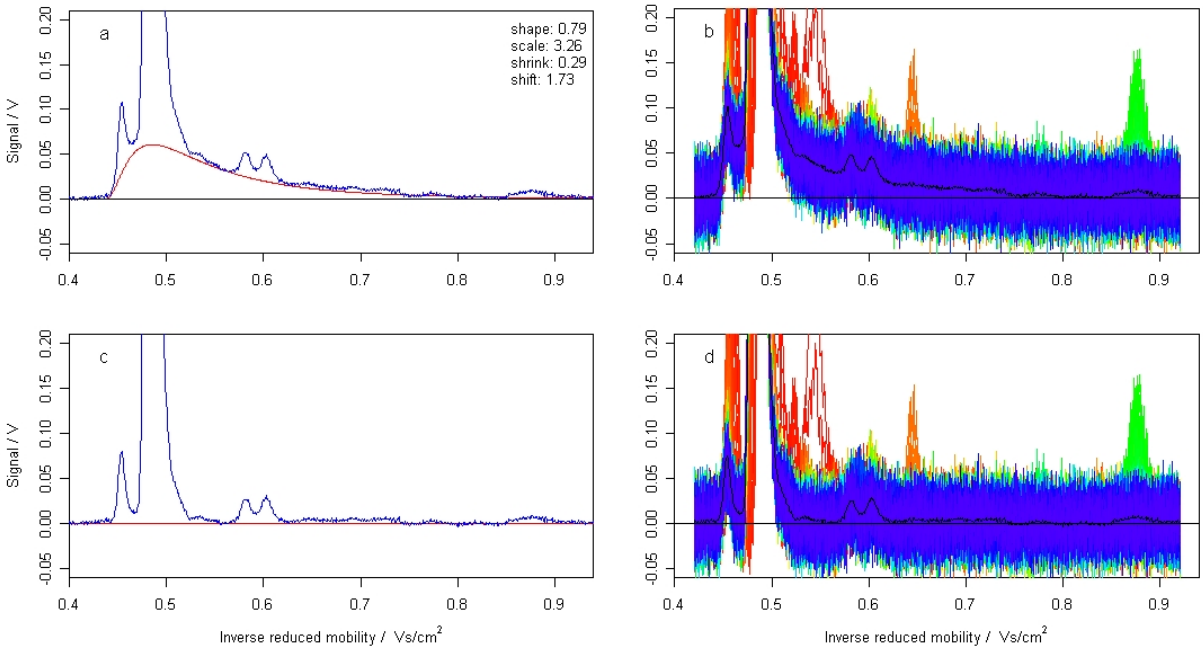

Figure 5.5: Effect of the RIP detailing for (a) the characteristic median spectrum with log-normal function and optimised parameters before and (c) after subtracting of the detailing function, and (b) spectra series with median spectrum before and (d) after subtracting of the detailing function: After RIP detailing the peaks grow from a common base level for the characteristic median spectrum as well as for the whole spectra series.

where $\mathbf{s}_{i\cdot}$ designates the ith row vector of the data matrix $\mathbf{S}$, was found to meet the requirements of characteristic feature conservation for the whole spectra series, a low variance around zero in noise areas, and high robustness very well, and was, therefore, chosen as suitable candidate for the adjustment of the detailing function.

By fitting the lognormal detailing function to the representative median spectrum of a measurement, and using the developed penalty term P for parameter optimisation, the RIP tailing was described and set down by subtraction of the resulting curve. The accurate detailing result after subtraction of the lognormal function from the median spectrum is shown in Fig. 5.5 a and c, which also gives the optimised parameters of the adjusted function. This beneficial effect could be transferred to the entire spectra series smoothly (Fig. 5.5 b, d).

5.3 Data reduction and signal-to-noise ratio increase

After the introduced preprocessing steps, IMS data still consists of a large number of data points with high levels of noise and redundancy, which interferes with most data analytical methods. To solve this problem, the methods of smoothing (Subsection 5.3.1) and denoising (Subsection 5.3.2) by means of wavelets (Chapter 3) were linked to combine the beneficial effects of both reduced dimensionality and a higher signal-to-noise ratio. Although wavelets have been applied for smoothing or denoising of IMS data before (Urbas and Harrington, 2001; Cai and Harrington, 1998), the two methods have not been used in a joint manner for this application before.

Applying the RIP detailing prior to smoothing and denoising by the wavelet transform, peaks were shown to share a common base level afterwards, while the advantageous effects of the wavelet operations were retained (Subsection 5.3.3).

5.3.1 Wavelet smoothing

To diminish the amount of redundancy as well as the computational cost, the next aim of the spectra preprocessing strategy was data compression. By the exclusion of whole levels of the wavelet decomposition before back-transforming, the wavelet transform was used for data smoothing, removing high-frequency components of the signal regardless of their amplitude. The achieved compression rate depended on the number of excepted scales.

To apply this method to the analysed spectra series, the data matrices had to be extended to new dimensions of dyadic length, $n_D^* = 2^{\lceil \log_2 n_D \rceil}$ and $n_R^* = 2^{\lceil \log_2 n_R \rceil}$, which was achieved by concatenating the series with a reflection, thus a reverse ordering of itself.

Executing a one-dimensional wavelet smoothing with a single compression level, initially for all single spectra of the extended data matrix, then for each point of the original drift time dimension across the spectra, a reduction to one quarter of the data points was achieved. The resulting data still contained the relevant information, while peak heights increased and the noise variation remained unchanged, leading to an improved signal-to-noise ratio (Table 5.1, Fig. 5.6 c, d).

After reduction of the data, the resulting matrix was truncated to new dimensions related to those of the original matrix, but with respect to the performed compression, which were determined by

$$n_D^c = \left\lfloor \frac{n_D^*}{c_D + 1} - \frac{n_D^* - n_D}{2^{\log_2(n_D^*) - c_D}} + 0.5 \right\rfloor, \; n_R^c = \left\lfloor \frac{n_R^*}{c_R + 1} - \frac{n_R^* - n_R}{2^{\log_2(n_R^*) - c_R}} + 0.5 \right\rfloor,$$

where the scalars c_D and c_R denominate the compression level in drift and retention dimension respectively.

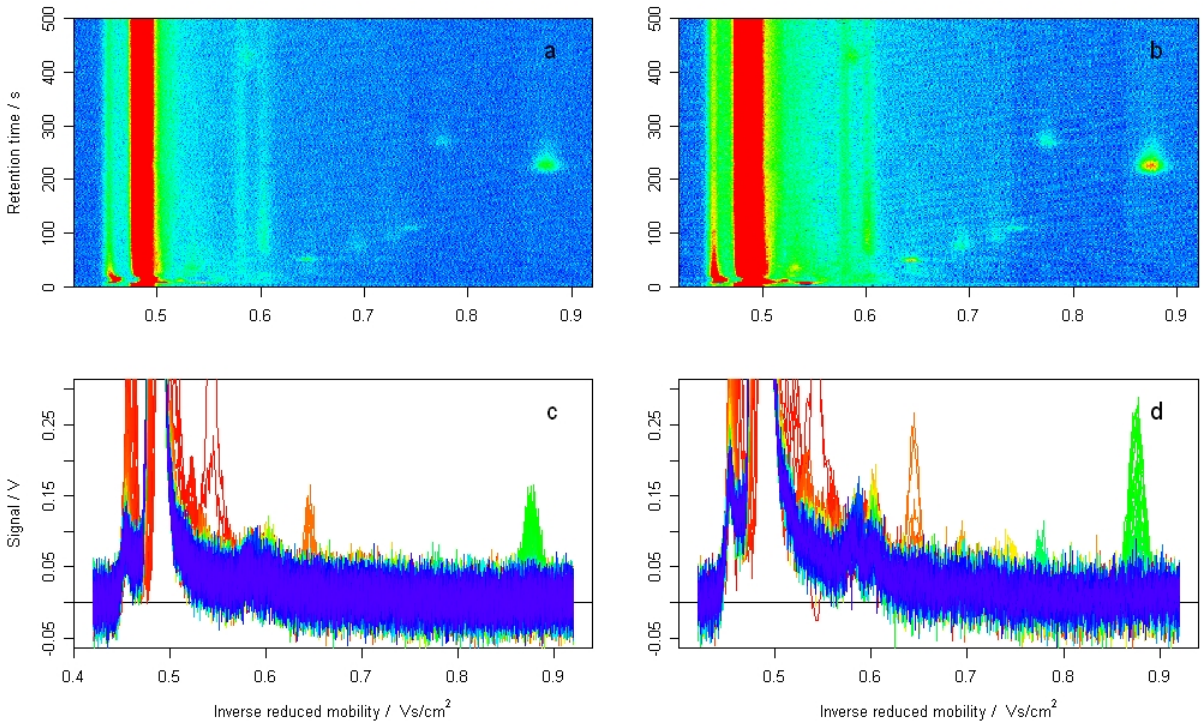

Figure 5.6: Heatmaps of (a) raw and (b) smoothed data show the achieved information conservation concurrently with the reduction to a quarter of data points, while the spectra series in a sideview of (c) raw and (d) smoothed data illustrate an improved signal-to-noise ratio, but also a not desired amplification of the RIP tailing.

Table 5.1: Quantification of height and signal-to-noise ratio (SNR) after different processing steps for the three instance peaks with the drift, retention time position pairs of (20.175 ms, 429 s) for peak A, (22.15 ms, 49 s) for peak B, and (30.05 ms, 225 s) for peak C.

	Peak A		**Peak B**		**Peak C**	
	Height	**SNR**	**Height**	**SNR**	**Height**	**SNR**
axes transformed	0.062	59	0.069	65	0.114	108
& detailed	0.040	38	0.057	54	0.113	107
& smoothed	0.128	82	0.245	157	0.246	157
& denoised – soft	0.059	193	0.142	665	0.224	733
– hard	0.074	252	0.204	696	0.254	866

Besides the beneficial effect of data compression and improved signal-to-noise ratio, the data, however, became more grainy and peaks were covered up to some degree by the now even amplified RIP tailing (Fig. 5.6 a, b), which strengthened the need for additional preprocessing steps.

5.3.2 Wavelet denoising

To yield spectra with an even smoother, less noisy appearance, the wavelet transform was furthermore used for the denoising of signals by removing small-amplitude components via thresholding of the wavelet coefficients of a two-dimensional MODWT regardless of frequency. Therefore, wavelet coefficients under a specified threshold were set to zero before back-transforming.

It can be differentiated between hard and soft thresholding (Cai and Harrington, 1998). While hard thresholding retains the original coefficients $\tilde{W}_{j,t_1,t_2}$ above the threshold λ unchanged, leading to the new coefficients $\tilde{W}^h_{j,t_1,t_2}$ with

$$\tilde{W}^h_{j,t_1,t_2} = \begin{cases} 0 & \text{if } \left|\tilde{W}_{j,t_1,t_2}\right| \leq \lambda \\ \tilde{W}_{j,t_1,t_2} & \text{if } \left|\tilde{W}_{j,t_1,t_2}\right| > \lambda, \end{cases}$$

the value of the threshold is subtracted from the coefficients $\tilde{W}_{j,t_1,t_2} > \lambda$ for soft thresholding, described by

$$\tilde{W}^s_{j,t_1,t_2} = \begin{cases} 0 & \text{if } \left|\tilde{W}_{j,t_1,t_2}\right| \leq \lambda \\ \text{sign}(\tilde{W}_{j,t_1,t_2})(|x_i| - \lambda) & \text{if } \left|\tilde{W}_{j,t_1,t_2}\right| > \lambda. \end{cases}$$

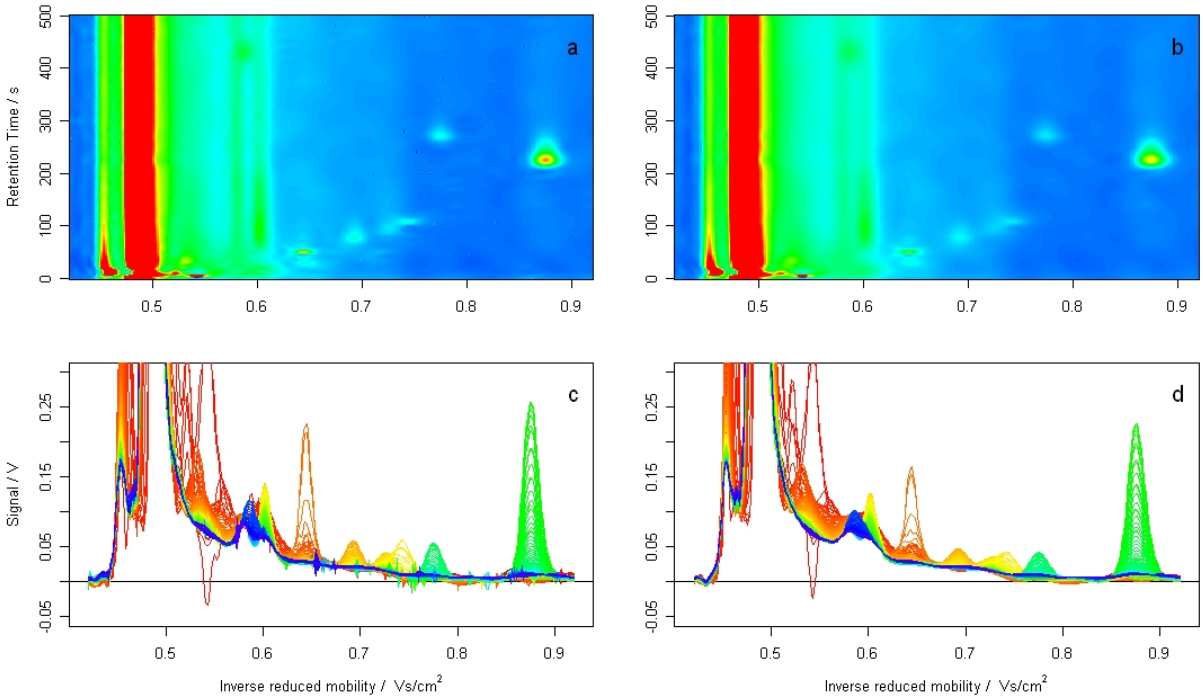

Figure 5.7: Heatmaps of (a) hard and (b) soft denoised data show the sharp peak shapes resulting from denoising via hard thresholding compared to broader peaks after usage of the method of soft thresholding, while the spectra series in a sideview of (c) hard and (d) soft denoised data illustrate the appearance of spiky artefacts for the result of hard thresholding, which were not obtained with soft thresholding.

With both strategies, a significant improvement of the signal-to-noise ratio could be achieved by denoising using a two-dimensional wavelet transform and Donoho's universal threshold $\lambda = \sigma\sqrt{2\log N}$, where σ signifies the standard deviation in noise parts and N the total number of data points (Donoho and Johnstone, 1995).

Although both methods yielded an improvement (Table 5.1), their results, however, differed considerably: hard thresholding lead to sharp peak shapes (Fig. 5.7 a), with a number of spiky artefacts inserted (Fig. 5.7 c). Resulting spectra were smoother in soft thresholding on the other hand (Fig. 5.7 d), but peaks tended to be relatively broad (Fig. 5.7 b). The optimal choice of the thresholding strategy is, therefore, dependent on the application.

5.3.3 Combined application with the detailing function

By applying the wavelet transform for smoothing and denoising of the spectra series, unfortunately, a severe amplification of the influence of the RIP tailing appeared concurrently with the beneficial effect of data compression and noise reduction (Fig. 5.8 a). This effect could be circumvented by using of the RIP detailing prior to the introduced

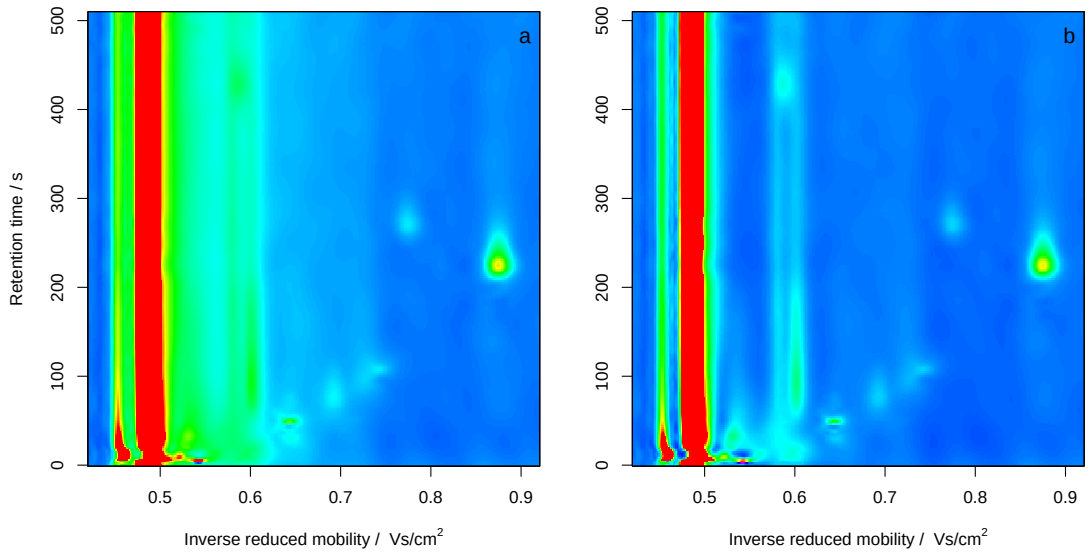

Figure 5.8: Heatmaps of smoothed and denoised data (a) without and (b) with detailing: peaks in the spectra part after the RIP appeared more clear when the RIP detailing was applied, while the beneficial effects of the wavelet operations were retained.

combination of smoothing and denoising by means of the wavelet transform (Bader et al., accepted in 2008). In doing so, peaks became clearer in the spectra parts after the RIP, while the advantageous effects of the wavelet operations were retained (Fig. 5.8 b). The varying impact of detailing on peaks in the different spectra parts can be quantified by

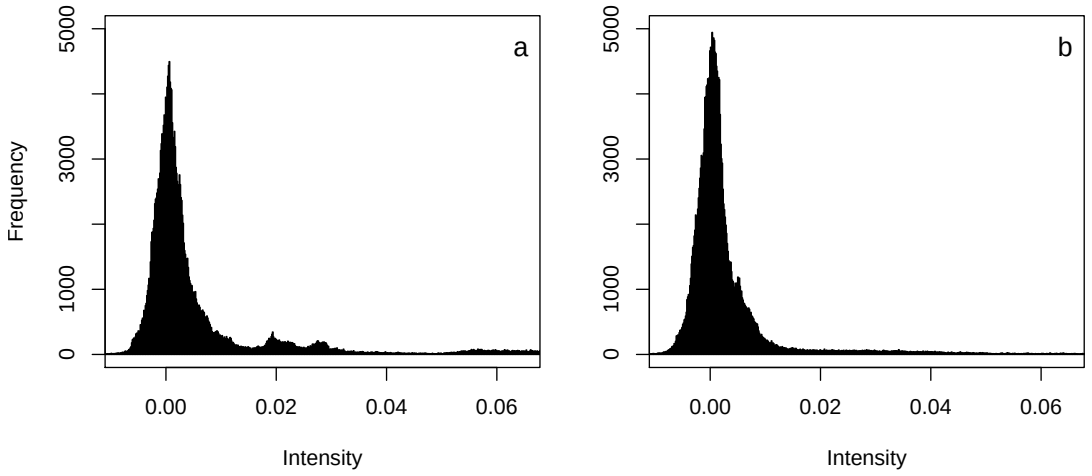

Figure 5.9: Histogram of intensity values of the entire data matrix after wavelet smoothing and denoising (a) without and (b) with detailing: whilst without detailing the data contained several hills, indicating different intensity categories according to peaks lying on different levels of the RIP tailing, only a single hill was left with denoising, as all peaks were set to the same level.

comparing their peak heights before and after transformation, showing only little influence on peaks in the latter spectra parts (Table 5.1).

After RIP detailing all peaks share a common base level, which can be pointed out considering histograms of all intensities values in the entire data matrix. The histogram for smoothed and denoised raw data contained several hills, possibly indicating different intensity categories according to peaks lying on different levels of the RIP tailing (Fig. 5.9 a). In the histogram of the intensity values after subtracting the adjusted lognormal function and performing the combined wavelet procedure, only one hill was left aside from the big noise part around zero, as all peaks were set to the same level by the method of RIP detailing (cp. Fig. 5.9 b).

Recapitulatory, combining the different axis transformations with the developed detailing function and the usage of wavelet transforms for smoothing and denoising, a powerful preprocessing was achieved.

Chapter 6

Peak detection in ion mobility spectra

Ion mobility spectrometry can offer short analysing times, however, the interpretation of the resulting data is often complex and time-consuming. This work aims, therefore, the automatised identification of characteristic IMS features by a peak detection procedure developed in several consecutive steps.

Firstly, a method based on a single threshold defined in a k-means clustering is developed (Section 6.1), which is then enhanced to a stagewise procedure, allowing for the identification of multiple peaks (Section 6.2). This algorithm consequently provides the basis for the more powerful, wavelet-based method developed in a last step, which even enables the detection of peaks without independent maxima (Section 6.3).

6.1 Merged peak cluster localisation

The first approach for peak detection sets an intensity threshold to group the measurement data into two sets of peak and non-peak points in an initial step. After the discrimination of different peaks, their locations are calculated to characterise measurements.

This method of merged peak cluster localisation (MPCL) works on the basis of only partially preprocessed data (Subsection 6.1.1); the main part of the algorithm is based on a k-means clustering to discriminate noise from peak areas and a merging region algorithm used for the discrimination of different peaks (Subsection 6.1.2). Although some limitations remain inherent in the method (Subsection 6.1.3), it has yielded some promising initial results.

6.1.1 Limited preprocessing

Only a confined preprocessing procedure was necessary using MPCL, since the method was constructed for the exposure to noisy data. Instead of the baseline correction introduced in subsection 5.1.3, page 71, a LOWESS algorithm (Cleveland, 1979) was used to shift data parts of noise down to zero and concurrently circumvent the aggravation of the RIP tailing. LOWESS is a robust locally weighted regression method which is implemented in the software package R (R Development Core Team, 2007). The effect of this step can be seen in Fig. 6.1 a: while peaks in all measurements areas were retained, the RIP tailing was corrected towards zero. Additionally, drift time values were converted to reduced mobility (Subsection 5.1.1, page 68), and retention times corrected with respect to the temperature of the column, increasing the comparability of the measurements (Subsection 5.1.2, page 70). Wavelet smoothing and denoising methods were not applied here.

In contrast to the preprocessing strategy introduced in Chapter 5, the spectra were truncated after the RIP in an additional step, as all other present peaks seemed negligible compared to the height of this special peak. This data cut-off was possible because no important differences were observed before the end of the RIP area. To define the end of the RIP area, the standard deviation in a pure noise area,

$$\sigma_{noise} = \sqrt{\frac{1}{(i_2 - i_1 + 1)\,(j_2 - j_1 + 1)} \sum_{i=i_1}^{i_2} \sum_{j=j_1}^{j_2} z_{ij}^2},$$

was used as a reference value to determine the intensity threshold $t_{RIP} = 2\sigma_{noise}$ with i_1, i_2, j_1, and j_2 defined as for formula 5.1 on page 72. The first position x_t after the RIP

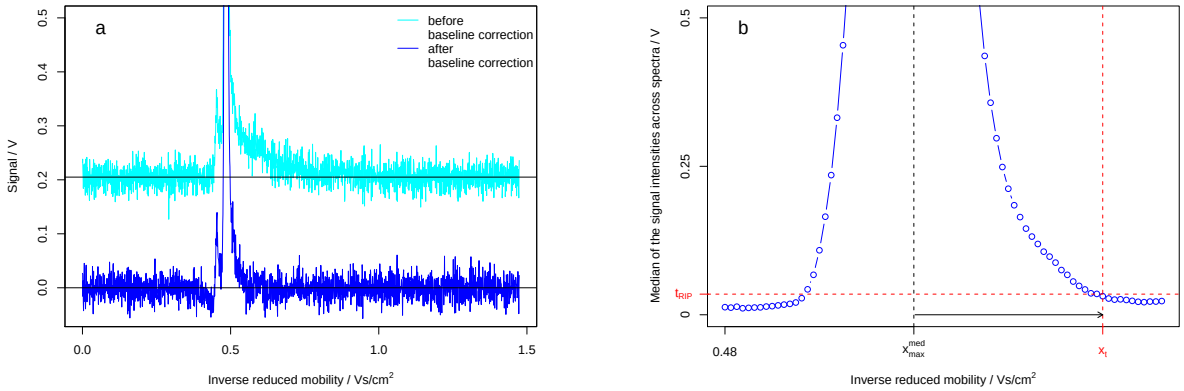

Figure 6.1: Illustration of the preprocessing steps for the merged peak cluster localisation showing (a) the baseline correction with LOWESS for a single spectrum, resulting in the desired effect of variation of intensity values around zero in noise areas as well as an adjustment of the RIP tailing, and (b) the choice of the RIP end cut-off point as the first position after the RIP maximum, where the median spectrum falls below a noise-defined threshold.

maximum s_{med}^{max}, where the median spectrum $\mathbf{s}_{med}$ falls below this value t_{RIP} was chosen as a cut-off point for the data matrix in the dimension of inverse reduced mobility (Fig. 6.1 b), defined by

$$x_t = x_i \text{ with } i = \min\left\{j | x_j > s_{max}^{med}, s_j^{med} < t_{RIP}\right\},\ j = 1, \ldots, n_D.$$

The resulting data gave the basis for proceeding further (Fig. 6.2 a).

6.1.2 Functionality of merged peak cluster localisation

Initiating the peak detection algorithm after the described preprocessing procedure, the values of the truncated data matrix were converted to a binary data set, separating data points belonging to peak structures and areas of noise. The function begins by splitting the measured signal intensities into two clusters – peak and non-peak – by the partition cluster method k-means, choosing the number of clusters equaling $k = 2$. Furthermore, starting values for the cluster means could be given as an input, allowing the integration of the side condition that the mean of the noise cluster should be zero. In addition, this

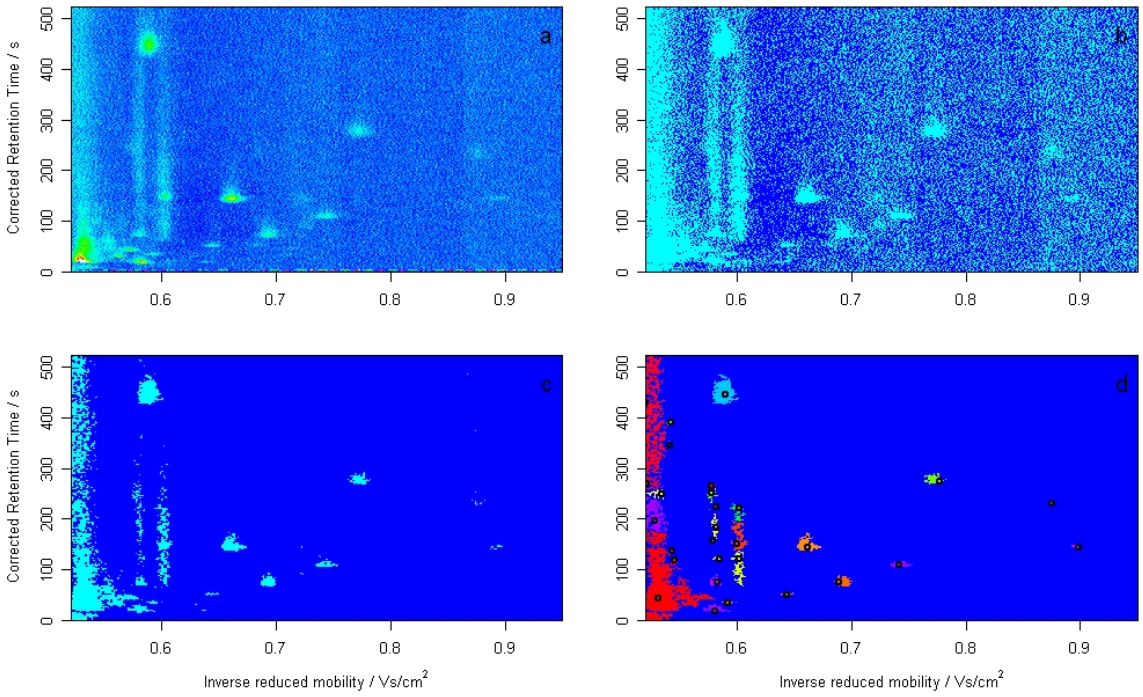

Figure 6.2: Steps of the merged peak cluster localisation for an instance breath measurement showing the result of (a) the truncation of the data after the RIP, (b) the splitting of data into peak and non-peak, (c) the deletion of noise artefacts, and (d) the separation of peaks via the merging regions algorithm along with the defined peak positions.

procedure was advantageous because different measurements are not treated by a uniform threshold, but by a flexible allocation into high and low values (Fig. 6.2 b).

Since this clustering step also assigned extreme values of noise to the peak cluster, a subsequent noise artefact deletion step was required. A point x_{ij} belonging to the peak cluster was, therefore, shifted to the noise cluster, if at least one of its eight-point-neighbours x_{pq}, $p \in \{i-1, i, i+1\}$, $q \in \{j-1, j, j+1\}$, belonged to this cluster. In doing so, border areas of real peaks were switched to the noise cluster, which was acceptable, because peaks were subsequently reduced to single points and the expansion of the detected peak areas was not the decisive factor of this algorithm. After this step peak areas were clearly separated from noise (Fig. 6.2 c) and data well prepared for further calculations.

Merging regions algorithm

Although it was feasible to separate noise and peak areas, it was still not possible to decide to which peak a single peak point belonged from its value. The next step, therefore, was to divide the data values in the peak cluster in a way that distinguishes between different peaks. For this, a merging region algorithm was reimplemented following the procedure of Bruce et al. (2000), constructed for image segmentation for football-playing AIBO[1] robots, which was found to be suitable for peak separation during the course of this research.

The basic functionality of the algorithm is described schematically in Fig. 6.3: starting from the original image of binary data points (Fig. 6.3 a), each row is divided into segments of identical values via run length encoding (Fig. 6.3 b), while starting and end points are stored. Next, adjacent segments containing identical values are merged into multi-row regions (Fig. 6.3 c).

Since the second step is decisive for this algorithm, it was more detailedly illustrated in Fig. 6.4: coming from a fully disjoint forest of segments (Fig. 6.4 a), adjacent lines were

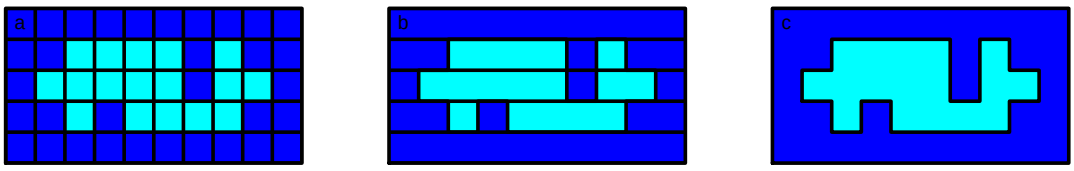

Figure 6.3: Scheme of the merging region algorithm: (a) Original image data points are divided into (b) segments of identical values and lastly merged into (c) multi-row regions.

[1] AIBO = Artificial Intelligence roBOt, homonymous with "companion" in Japanese.

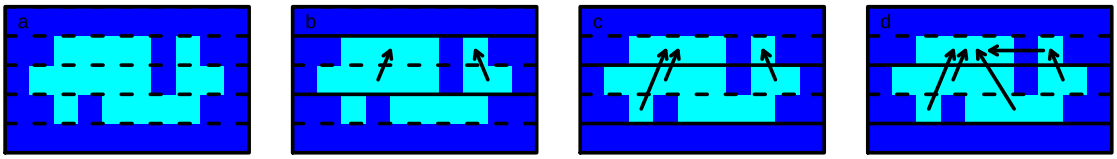

Figure 6.4: Detailed illustration of the merging region step: (a) Fully disjoint segments are (b) merged if neighbours contain the same values, where (c) new assignments are made to the furthest parent, and (d) the latter parent is updated for occuring overlaps.

compared and neighbour segments containing the same values merged (Fig. 6.4 b). New assignments were made to the furthest parent (Fig. 6.4 c) and if an overlap with more than one segment occured, the latter parent was updated (Fig. 6.4 d).

Afterwards data were not only divided into non-peak and peak any longer, but into peak A, B, C, ..., where the assignment of points to one of the peaks was now not only based on the intensity, but also on its position (Fig. 6.2 d).

Peak characterisation

Having identified the points belonging to the different peaks, the next step was the characterisation of peak regions by their peak center locations by computing the mode of the coordinates of all points belonging to the same peak region as

$$\begin{aligned} x_M^A &= x_{i^*}, \quad \text{with} \quad i^* = \max_i \left\{ \#x_{ij} \in \mathbf{s}_{i\bullet} | x_{ij} \in \mathcal{A} \right\}, \text{ and} \\ y_M^A &= y_{j^*}, \quad \text{with} \quad j^* = \max_j \left\{ \#x_{ij} \in \mathbf{s}_{\bullet j} | x_{ij} \in \mathcal{A} \right\}, \end{aligned}$$

where $\mathcal{A}$ is the set of points belonging to the region constituting peak A. The resulting peak position corresponded with the spot, where the peak possessed maximum width and length (Fig. 6.2 d).

6.1.3 Results and limitations of merged peak cluster localisation

The MPCL method yielded a data reduction from a million data point matrix to two vectors of peak coordinates with its length corresponding to the number of peaks detected in a measurement. Additionally the height at each peak position was stored as an index of peak intensity. In this, the MPCL allowed an efficient reduction of the dimensionality of the data, resulting in a reasonable peak representation (Fig. 6.5 a).

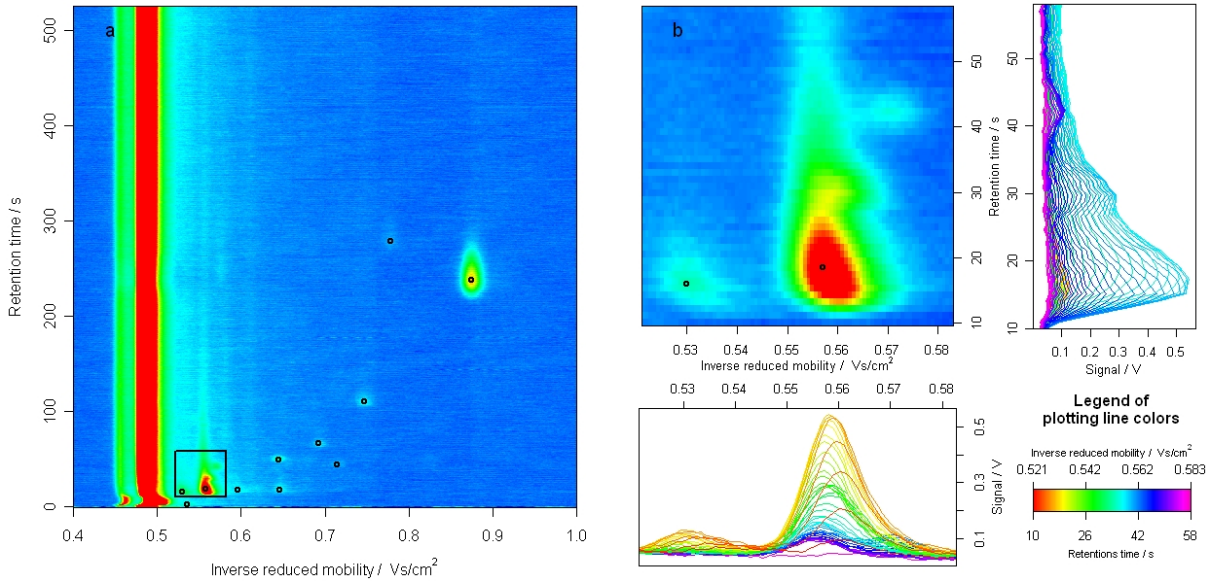

Figure 6.5: Result of the merged peak cluster localisation for (a) the whole relevant part of a measurement in a heatmap with detected peak positions, and (b) a small measurement part, marked in a, considered in a heatmap and two side-views, showing limitations of the method for shoulder peaks and peaks that are not baseline separated.

Nevertheless, some limitations were inherent in this procedure, as it was not possible to distinguish peaks that are not baseline-separated, if their connection lies over the threshold determined in the k-means clustering procedure. Overlapping peaks were, therefore, detected as a single peak, leading to partially inaccurate results (Fig. 6.5 b). Furthermore, the characterisation of peaks was insufficient, since it was only performed on the base of the position and optionally the peak height. It would be more meaningful to additionally give an idea of the peak magnitude for better characterisation purposes.

Still, the method of MPCL could be successfully used for the analysis of IMS data in a way that would not have been previously possible (Section 7.1).

6.2 Growing interval merging

To overcome the limitations of the MPCL method, another approach for peak detection was established, also allowing for the detection of multiple peaks and giving a better characterisation of peak shape and size.

The underlying idea of this peak detection algorithm is to create an intuitive, natural approach proceeding in a similar way as the human visual perception of peaks: viewing an IMS measurement from the topview perspective, outstanding intensity points are

identified with respect to their surrounding area. When transferred to an algorithm, this results in the developed method of growing interval merging (GIM) working in a stage-wise manner, starting from the top of the intensity range. The basis of this algorithm is the IMS data preprocessing procedure, previously described in Chapter 5.

The program flow of the GIM algorithm can be broken down into three main parts: Firstly, a sequence of intensity thresholds is defined, corresponding to growing intensity intervals used at the different stages of the algorithm (Subsection 6.2.1). Secondly, peak detection is processed at each stage in a manner closely related to the method of MPCL, and peaks are described by the adjustment of ellipses (Subsection 6.2.2). The last and most decisive step of the algorithm is the connection of the stages (Subsection 6.2.3), allowing the resolution of twin- and multiple peaks, thus solving one of the largest challenges in peak detection problems.

After the presentation of some final steps for the refinement of peak characterisation, the results of GIM are illustrated and the remaining limitations of this peak detection method are discussed (Subsection 6.2.4).

6.2.1 Sequence of intervals

The initial step of the GIM method was to define a sequence of thresholds characterising the growing intensity intervals, giving a foundation for the stages of the GIM algorithm.

Firstly, the lowest intensity threshold, dividing the noise from peak areas, was determined to set an end point for the algorithm, based on a histogram of all the intensity values, irrespective of their position in the data matrix. As the histogram of data only baseline-corrected was non-informative because of the large extent of noise (Fig. 6.6 a), a comparable histogram was considered for the fully preprocessed data (Fig. 6.6 b), containing indications for different intensity categories, which could be nominated as (I) noise, (II) peak and (III) RIP intensity values. It was, therefore, possible to define a noise threshold at the position of the local minimum between noise and peak intensity marks. The three-fold noise threshold was chosen to be the minimum level, t_n, for peak detection.

Next, the remaining thresholds defining the sequence of intervals corresponding to the stages of the algorithm were constructed in a way that each interval, $\mathbf{I}_k$, included the one from the previous stage,

$$\mathbf{I}_k = [i_k, \max(z_{ij})] \quad \subset \quad [i_{k+1}, \max(z_{ij})] = \mathbf{I}_{k+1}, \qquad \forall\ k \in [1, \ldots, n_s - 1],$$

where n_s denominates the number of stages.

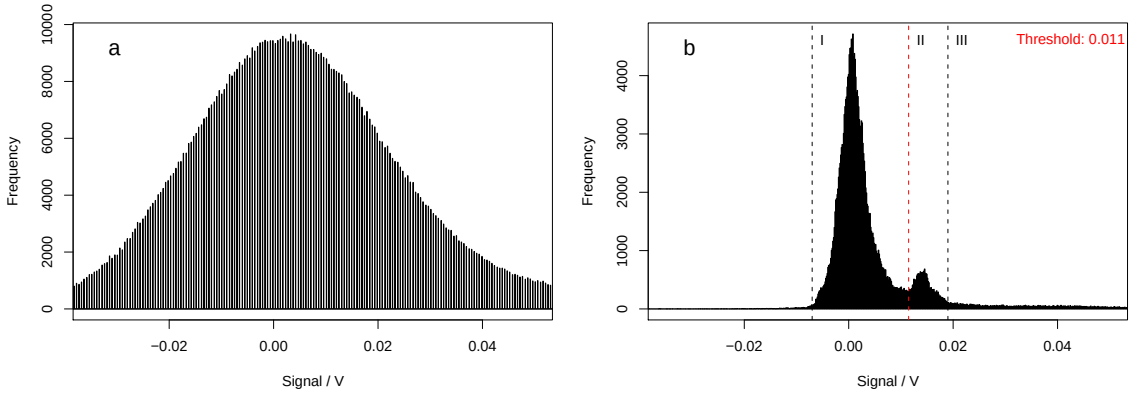

Figure 6.6: Histogram of intensities for (a) raw data and (b) fully preprocessed data: while the histogram of the raw data is non-informative due to the high level of noise, the preprocessed data show indications for different intensity categories, nominated as (I) noise, (II) peak, and (III) RIP, which allows the definition of a noise threshold.

The initial idea was to construct a sequence with a constant growth of the interval range from half RIP height $i_1 = \max(z_{ij}/2)$ down to the minimum level t_n. This yielded equidistant lower interval boundaries i_k with

$$\Delta i = i_{k+1} - i_k \; = \frac{i_1 - t_n}{n_s - 1} \qquad \forall\; k \in [1, \ldots, n_s - 1].$$

This sequence, however, was not well adapted to the data, as the upper intensity area, represented only by the RIP, was considered in the same way as the lower range, containing all other peaks.

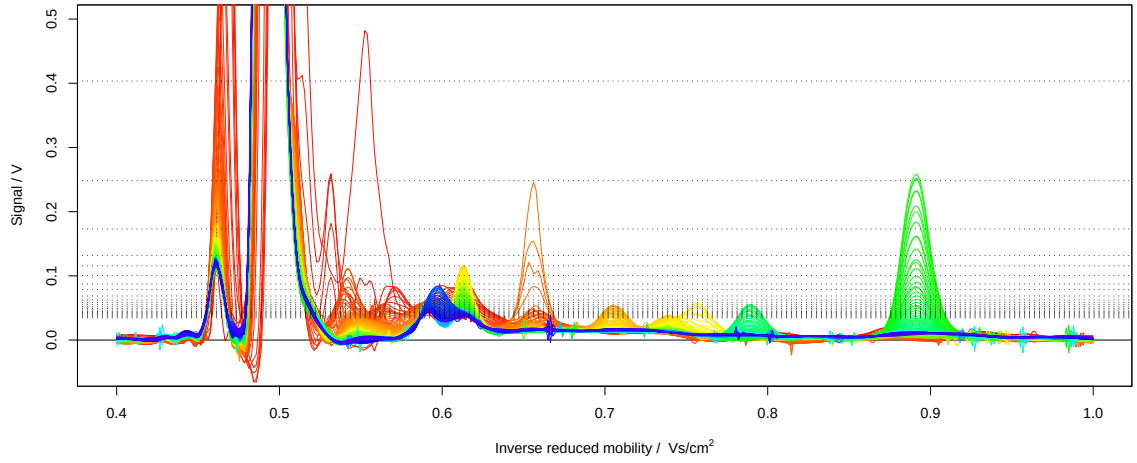

Figure 6.7: Thresholds defining the sequence of intervals in the lower section of the intensity range: signal intensity areas containing the majority of the peaks are scanned more detailed than the upper parts represented basically by the RIP.

For this reason, an alternative strategy was developed, choosing the sequence of intervals in a way that the interval occupancy N_k was growing constantly by ΔN defined as

$$\Delta N = N_{k+1} - N_k = \#\left\{z_{ij}|\ z_{ij} \in \mathbf{I}_{k+1}\right\} - \#\left\{z_{ij}|\ z_{ij} \in \mathbf{I}_k\right\}, \qquad \forall\ k \in [1, \ldots, n_s - 1].$$

To ensure the total number of data points between the first and last threshold allowed the division into n_s intervals containing the same number of data points, the first threshold i_1 was not chosen as the exact RIP half, but as

$$\begin{aligned} i_1 &= z_{(p+q)}, \\ p &= \left\lfloor \frac{\#\left\{z_{ij}|\ z_{ij} \in \left[t_n, \frac{\max(z_{ij})}{2}\right]\right\}}{n_s - 1} + 0.5 \right\rfloor (n_s - 1), \\ q &= \#\left\{z_{ij}|\ z_{ij} < t_n\right\}. \end{aligned}$$

This resulted in a much better adaption to the data, since areas with high peak density were scanned more detailedly than parts only represented by the RIP (Fig. 6.7).

6.2.2 Stagewise procedure

Having defined the intervals corresponding to the different stages of the GIM algorithm, a peak detection procedure closely related to the MPCL method was applied at each stage. The only difference was the criteria for splitting data into peak and non-peak, which was no longer based on a k-means clustering, but determined by the interval corresponding to the current stage.

For this, the points of the data matrix $\mathbf{S}$ were nominated as peak where $z_{ij} \in \mathbf{I}_k$, and as non-peak where $z_{ij} \notin \mathbf{I}_k$, resulting in a matrix of ones and zeros. It was still not possible to tell to which peak a point in the binary data matrix belonged (Fig. 6.8 a), thus, the next step of the procedure separated peaks from one another by the merging region algorithm introduced for the MPCL method (Subsection 6.1.2, page 84). Afterwards, data could not only be divided into peak and non-peak, but different peaks regions could also be distinguished (Fig. 6.8 b).

Due to the applied preprocessing procedures, the shape of the resulting peak areas was less irregular than in the MPCL method. To allow for the storage of peak information sparsely, the defined peak regions had to be characterised adequately with few parameters. This was achieved by the adjustment of ellipses to respect the typical oval peak shape,

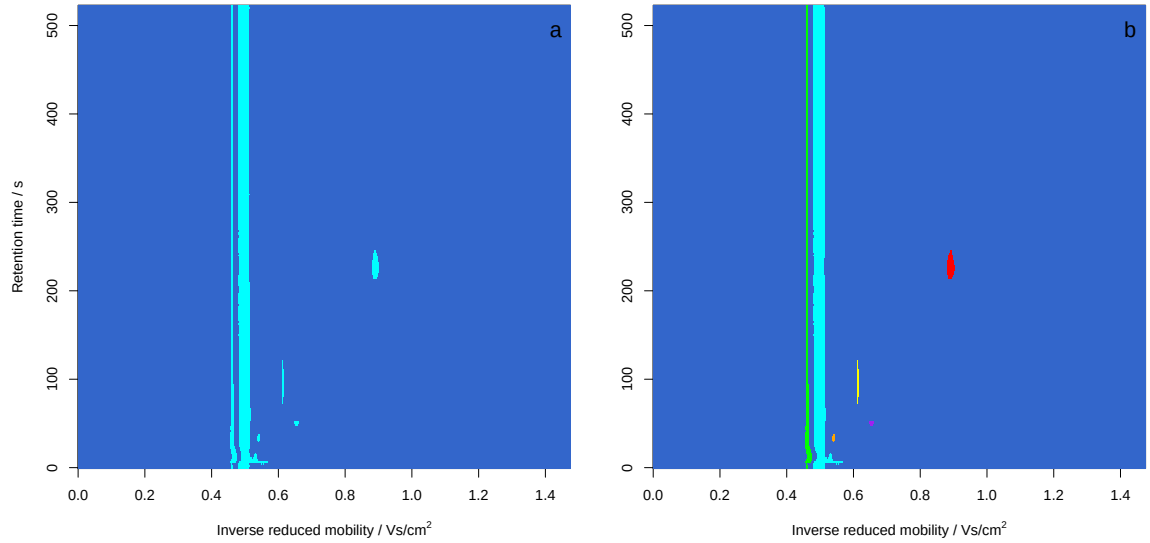

Figure 6.8: Spectra series of an instance measurement (a) after division into peak and non-peak and (b) after distinguishing peaks via the merging regions algorithm: the step of merging regions allows to decide not only whether a data point belongs to a peak or the noise area, but also to which peak it belongs.

claiming the side condition of ellipse axes being parallel to the coordinate axes, which can be expressed as

$$\frac{(x-x_0)^2}{a^2} + \frac{(y-y_0)^2}{b^2} = 1$$

with the four parameters x_0, y_0 for position and a, b for extent (Fig. 6.9).

The position (x_0, y_0) of a peak was specified as the point with the maximum intensity value in the peak region, and could be defined by

$$\begin{aligned} x_0 &= x_{i'} \,, \quad i' = \max_i \{z_{ij} |\ z_{ij} \in \mathcal{A}\} \\ y_0 &= y_{j'} \,, \quad j' = \max_j \{z_{ij} |\ z_{ij} \in \mathcal{A}\} \,, \end{aligned}$$

where $\mathcal{A}$ is the set of data points belonging to an instance peak A.

To determine the extent a, the first and last points of peak A were matched in the dimension of inverse reduced mobility as

$$\begin{aligned} a_1 &= x_{i'} \,, \quad i' = \min\left(i \mid \exists\ z_{ij} \in \mathcal{A}\right) , \\ a_2 &= x_{i''} \,, \quad i'' = \max\left(i \mid \exists\ z_{ij} \in \mathcal{A}\right) , \end{aligned}$$

and the minimum deviation between both points and the position x_0 was set to a:

$$a = \min(x_0 - a_1, a_2 - x_0).$$

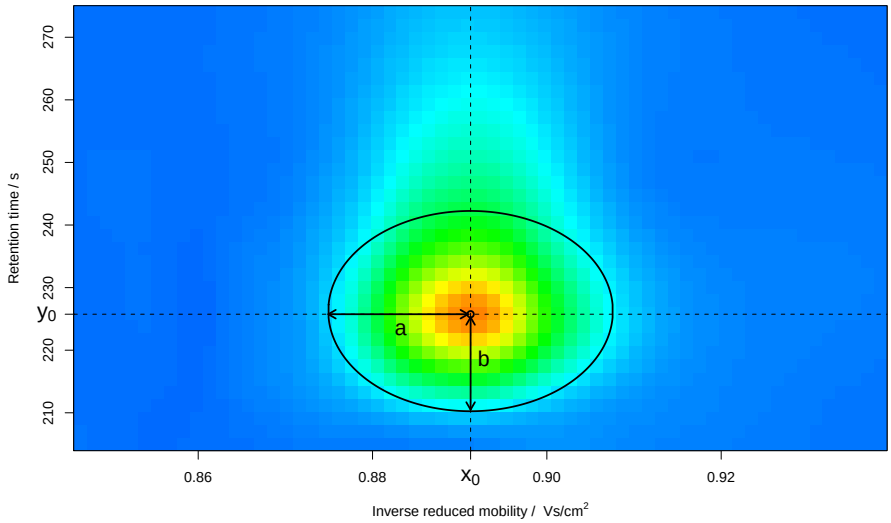

Figure 6.9: Illustration of the ellipse adjustment showing the parameters of position (x_0, y_0) and extent (a, b) for an instance peak.

For the specification of the extent b, the first spectrum in retention time, containing a point belonging to the peak region was searched and its deviation to the peak position y_0 chosen as b by

$$b = y_0 - y_{j'} \ , \quad \text{with } j' = \min\left(j \mid \exists\ z_{ij} \in \mathcal{A}\right).$$

The extent in the other direction was not incorporated for the determination of b, as it was influenced by peak tailing, while the main interest focused on the cores of the peaks.

In addition to the ellipse parameters, the maximum height was stored giving a further indication of peak intensity, resulting in a peak list, $\mathbf{P}_K$, constituted by five values per peak on each stage.

6.2.3 Connection of stages

To enable the detection of twin and multiple peaks, the GIM algorithm was structured in a stagewise manner, where the connection of stages is the decisive step of the method.

A measurement part containing multiple peaks was used as an example to illustrate the challenge of multiple peak detection (Fig. 6.10 a): the higher peaks could not be separated using the highest threshold allowing to find both of the lower peaks (Fig. 6.10 b). Conversely, the lowest peak was not detected at all using the lowest threshold discriminating between the two higher peaks. A single threshold was, therefore, not capable to yield a correct characterisation of this measurement part. To circumvent this problem, a

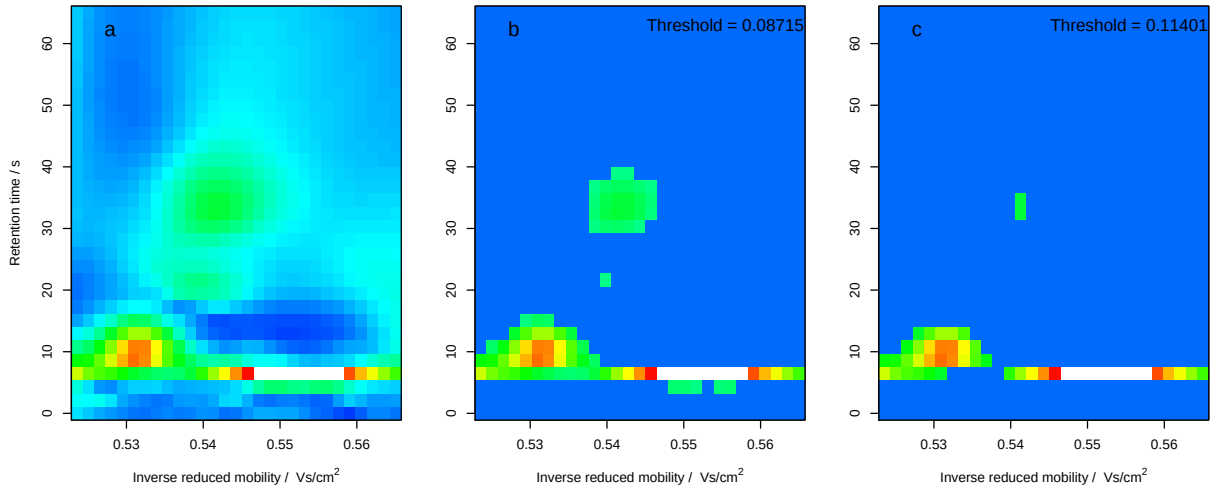

Figure 6.10: Heatmaps of (a) a twin-peak example, where simple thresholds will either detect (b) the two higher peak as one, or (c) will miss at least one of the others at all.

stagewise procedure was applied in the GIM algorithm, connecting the different stages by a comparison of the resulting peak list, $\mathbf{P}_k$, of each stage with the peak list, $\mathbf{P}_{k-1}$, of the previous stage.

A multiple peak was indicated if one of the peak regions of the current stage contained more than one peak position (x_0^{k-1}, y_0^{k-1}) from the previous stage. Subsequently, lines in $\mathbf{P}^k$ belonging to the new peak region found to contain several peak positions of the stage before were rejected and substituted by the corresponding lines of $\mathbf{P}^{k-1}$. Peak regions of the current stage containing no or only one peak position of the previous stage, were kept unchanged. Consequently, the peak list of the last stage always had to be independently retained, but the procedure was very effective for the desired detection of multiple peaks.

As an additional refinement in peak characterisation, some final steps were included in the GIM algorithm to calculate the ellipse area $A^\circ = \pi ab$ as an additional measure for peak intensity. Furthermore, the peak heights were recalculated by matching the detected peak positions in data that were only axes-transformed and denoised, since the data processed

Table 6.1: 5 point summary with mean after different preprocessing steps

Mode of data	Min.	1. Quart.	Median	Mean	3. Quart.	Max.
axes transformed	-0.094	-0.00862	0.00457	**0.029**	0.0197	**3.26**
+ denoising	-0.023	-0.00037	0.00089	**0.029**	0.0042	**3.19**
+ RIP clearing	-0.024	-0.00048	0.00064	**0.024**	0.0027	**3.13**
+ smoothing	-0.074	-0.00073	0.00116	**0.048**	0.0051	**6.35**

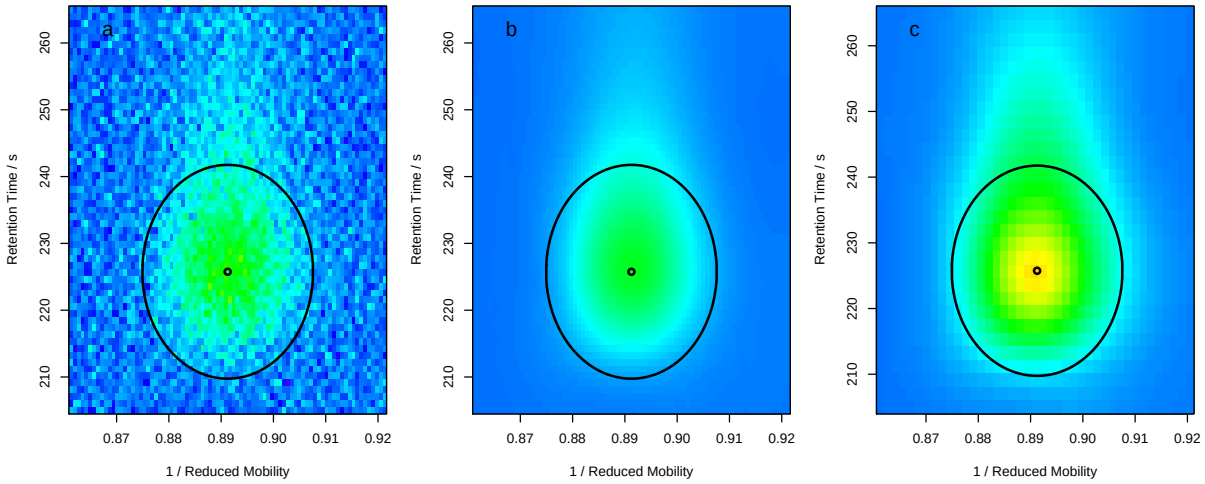

Figure 6.11: Ellipse adaption for (a) raw, (b) denoised, and (c) fully preprocessed data: the height of the the denoised data gives the most appropriate value of the true peak height in the raw data, where the height is biased by noise, while the fully preprocessed data results in an overestimation of the intensity.

in this way was the most similar to the raw data, whilst peak heights were not biased by noise.

This was apparent from the 5 point summaries of the intensity values after the different preprocessing steps, especially for the mean and the maximum values, since the other quartiles were mainly influenced by the noise area (Table 6.1). Although the properties of the fully preprocessed data (Fig. 6.11 c), such as data compression and signal-to-noise ratio increase were useful in other algorithm parts, the denoised data (Fig. 6.11 b) were more appropriate for stating the actual peak height (Fig. 6.11 a), and were thus a reasonable choice for its refinement.

After these final steps, an appropriate peak characterisation was achieved, giving an end point for the algorithm.

6.2.4 Results and limitations of growing interval merging

The result of the GIM algorithm was a data reduction to only six values per peak, instead of one million data points, characterising the measurement data sparsely but with little information loss.

Besides the peak area and height, giving an idea of the peak intensity, the six values contained ellipse parameters allowing for a visualisation of the data in an ellipse representation, enabling a direct comparison of the peak detection results with the original spectra

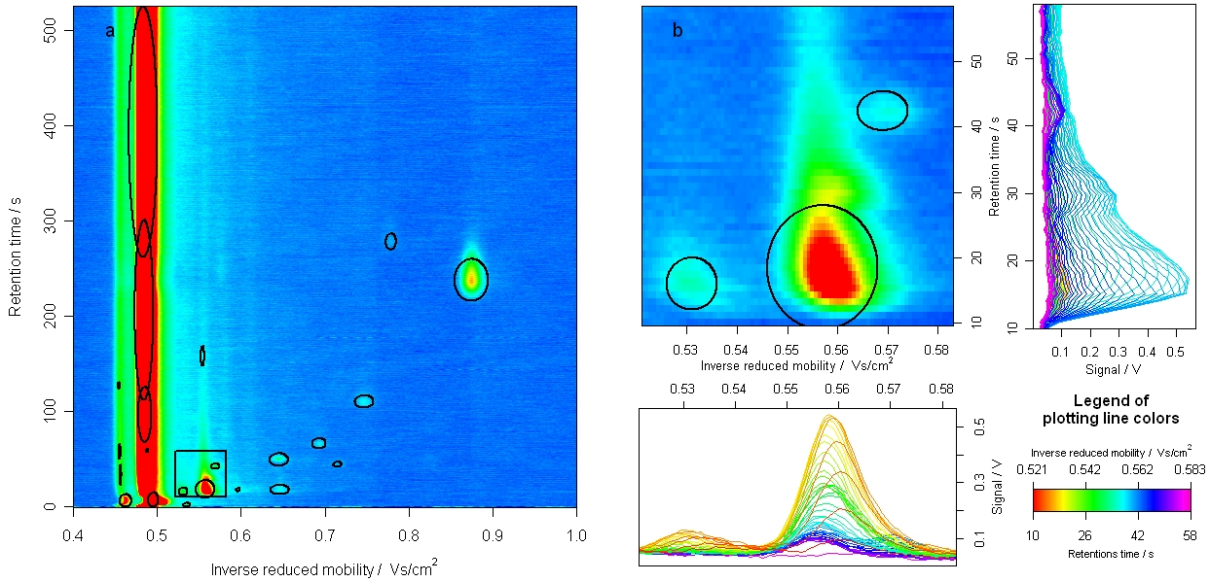

Figure 6.12: Result of the growing interval merging for (a) the entire relevant part of a measurement in a heatmap with detected peak ellipses, and (b) a small measurement part, marked in a, considered in a heatmap and two side-views, showing the limitation of the method for the detection of shoulder peaks.

series, and showing that the aim of an efficient peak detection method was achieved, neglecting tailings and impurities (Fig. 6.12 a).

In conclusion, the GIM method enabled the reduction of spectrometric data to reasonable peak variables by a stagewise pattern recognition and feature extraction, and could resolve multiple peaks by an appropriate strategy for the connection of stages. In doing so, it conserves the essential information and gives a valuable starting point for arbitrary continuative analyses.

A comparison with the MPCL algorithm (Section 6.1) indicates the advantage of an improved resolution of multiple peaks and the enhanced characterisation of peaks by ellipse parameters. Limitations of the GIM peak detection approach, however, include the detection of shoulder peaks, i.e. peaks that do not possess independent maxima can not be found using this algorithm (Fig. 6.12 b).

6.3 Wavelet-based multiscale peak detection

The limitations of the GIM algorithm motivated the development of a more powerful wavelet-based peak detection method, allowing the recovery of peaks disguised by other peaks in their surrounding area and thus not possessing independent maxima.

Enhancing the approach of Randolph and Yasui (2006) for the multiscale processing of single spectra (Subsection 6.3.1) to the use with three-dimensional spectra series (Subsection 6.3.3) and connecting several resolution levels (Subsection 6.3.4), the benefits of a better quantification with less preprocessing could be provided concurrently with an improved peak detection sensitivity. Besides the introduction of the work flow of the developed method, challenges in the transfer to the application to spectra series (Subsection 6.3.2) and the resulting outcomes (Subsection 6.3.5) are described.

6.3.1 Multiscale processing of single spectra

The multiscale processing of single spectra, introduced by Randolph and Yasui (2006), is based on an MRA by means of the MODWT, as this wavelet-based method is translation-invariant and applicable to data of arbitrary dimensions (Chapter 3).

After decomposition of a spectrum into a series of details $\mathcal{D}_j$, the appropriate level containing the relevant frequency information is selected and used as the basis for peak detection. Starting from the identified detail, a MODWT is applied, using a wavelet with one, and then two vanishing moments, such as the Haar and Daubechies $D(4)$ wavelet. The two sets of coefficients of the same level as previously selected in the MRA are then used as a substitute for the first and second derivative to detect peaks as local maxima in

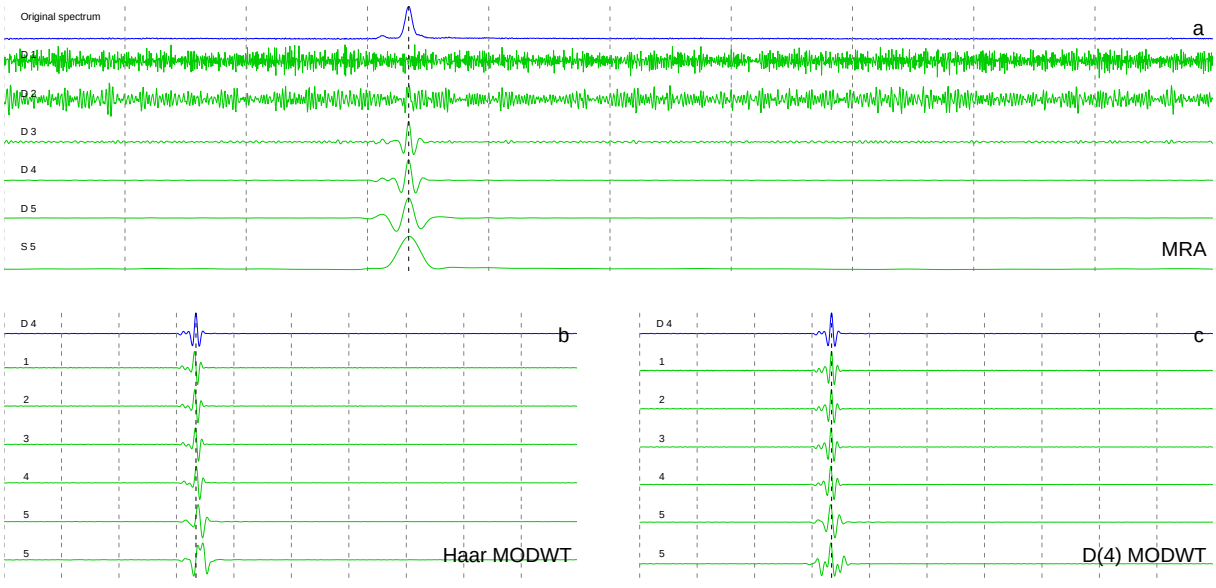

Figure 6.13: Plots of (a) the MRA details and smooth of a single spectrum and the MODWT decomposition of the MRA detail 4 based on (b) the Haar and (c) the D(4) wavelet after alignment with the original spectrum: the wavelet transforms allow to detect the potential hidden peak covered by the right side of the RIP.

the MRA detail of choice. To allow for proper interpretation of the MODWT coefficients, it is important to align them with the underlying details (Section 3.3). This concept is shortly illustrated for a single spectrum. Fig. 6.13 a shows the original spectrum in the top and five details as well as the smooth below. As detail 4 seems to be most promising in grasping the information content inherent to the considered spectrum, this detail was analysed using a MODWT on the basis of the Haar and the $D(4)$ wavelet. The two resulting sets of coefficients are plotted in Fig. 6.13 b and c after the necessary alignment with the underlying detail. The determined coefficients coincide with peak positions of the original spectrum, and even the shoulder on the right side of the RIP can be identified as a potential feature here.

6.3.2 Challenges in the enhancement to three-dimensional data

The positive performance of the multiscale processing for single spectra encouraged the enhancement of this method to the application to three-dimensional spectra series. In doing this, however, a couple of challenges appeared, due to the fact that an MRA now yields a complex system of three matrices per level instead of a sequence of singular detail vectors (Section 3.5, Fig. 3.9, p. 35).

The first question, therefore, was if it was possible to base the algorithm on just one matrix of the optimal level of the MRA instead of working with the HH, LH and HL matrices as a combination, and which of the matrices would be the optimal choice for this procedure.

For the data in question, the LH matrices of the MRA yielded optimal results, containing the main part of the information required for peak localisation aspects. The reason for this was the symmetrical and sharp peak shape in the original drift time dimension, which allowed a satisfying resolution even with the wavelet smooth in this direction, while the broad peaks with tailing effects in the retention time dimension required to be narrowed by the determination of the wavelet detail (Fig. 6.14 c). The HL matrices on the other hand, did not show the same beneficial outcome, as peaks were additionally stretched in the dimension of retention time and, therefore, only a slight benefit was gained by the more precise localisation in the drift dimension (Fig. 6.14 b). Furthermore, the HH matrices, originally favoured due to the fact that the single spectra method is based on the details, contained interfering artefacts, which could be explained by a similar behaviour as for also complex to interpret two-dimensional derivations (Fig. 6.14 d).

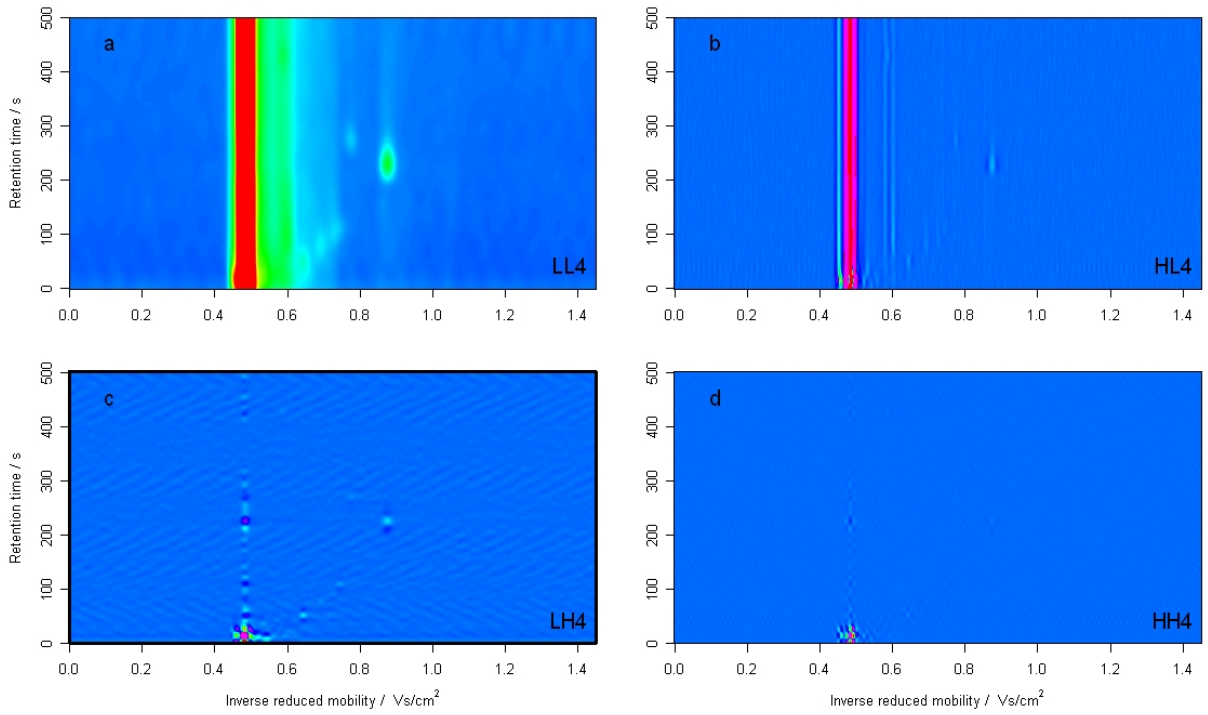

Figure 6.14: Heatmaps of the (a) LL4, (b) HL4, (c) LH4, and (d) HH4 matrices of a two-dimensional MRA decomposition of a spectra series using the MODWT method: the LH4 matrix offers the best representation of the original data for peak detection aspects, as peak positions appear most clearly and the RIP and peak tailings are not the focus of the peak detection method.

After finding a single matrix per level to be appropriate for proceeding further, the next problem encountered was a flood of data, produced by the two complex matrix systems of the Haar and $D(4)$ MODWT decompositions determined for the localisation of local maxima in the chosen detail matrix. Additionally, even if this challenge could be solved efficiently, a potential problem arose if the optimal levels differed between the two time dimensions, as the peak width and thus the relevant scale of peaks in the both directions was not necessarily the same. In this case a combination of several levels could be relevant in the MRA, which would even amplify the problem of a massive set of matrices to deal with.

To solve this problem, the original thought of transferring the method of Randolph and Yasui (2006) analogously to the case of spectra series was abandoned and the GIM algorithm was used for peak detection in the chosen MRA LH detail matrices instead. Additional advantages of this procedure besides a reduction in the amount of data to be processed, included the avoidance of the necessary alignment shifts for the MODWT matrices, and the fact that the choice of the relevant matrix levels had to be done only for the MRA.

Combining the GIM algorithm with the MRA, the method could be used without other complex preprocessing steps, and consequently better GIM results were enabled as the MRA clarified measurement features as shoulder peaks.

6.3.3 Wavelet-based peak detection in spectra series

The procedure of the developed wavelet-based peak detection resulted in a reduced preprocessing effort, as only the axes transformations to inverse reduced mobility and corrected retention time were necessary for alignment aspects. A baseline correction, or a detailing of the RIP, were not required as the details $\mathcal{D}_j$ are based only on local changes, not on the trend of the original spectra, whilst the effect of denoising was achieved concurrently with the MRA.

After data preprocessing, two basic choices had to be made for the performance of the MRA, considering the optimal wavelet function, as well as the most appropriate levels for further processing. The use of the Haar wavelet was most beneficial, since wavelets of greater width than this function yielded a less localised MRA and caused overshooting artefacts, which can be seen best for the spectra series part around the position of the RIP (0.49 Vs/cm^2) in the early retention time range (Fig. 6.15 a-c). Comparing the different LH matrices of the resulting decomposition that were identified as the sufficient basis for further processing (Subsection 6.3.2), level 4 was optimal, as it covers the frequency range most related to that of the peaks, and thus peaks found by visual inspection were present also in the LH4 matrix for the considered instance data. Level 3, on the other hand, was blind for broader peaks in the upper retention time region, while level 5 did not resolve the small peaks, mostly existing in the measurement part with low retention times (Fig. 6.15 d-f).

The LH4 matrix of the Haar wavelet-based MRA was, therefore, the basis of the GIM algorithm for peak detection, which was used as previously introduced (Section 6.2), but with 50 stages and a different noise threshold t_n, now calculated as the four-fold standard deviation in a noise area.

The resulting peak list was subject to some optimisation steps to exclude artefact peaks in the last five spectra, which appeared due to the margin artefacts of the wavelet transform, and to join the peak positions in the RIP area. The latter was performed by merging all peaks in the interval $(x_{RIP} - 0.005, x_{RIP} + 0.01)$ around the RIP position $x_{RIP} = 0.49$ to a single RIP peak, described by the position x_{RIP} in the drift dimension, the mid position of all spectra in the retention time direction, and the maximum height observed in the

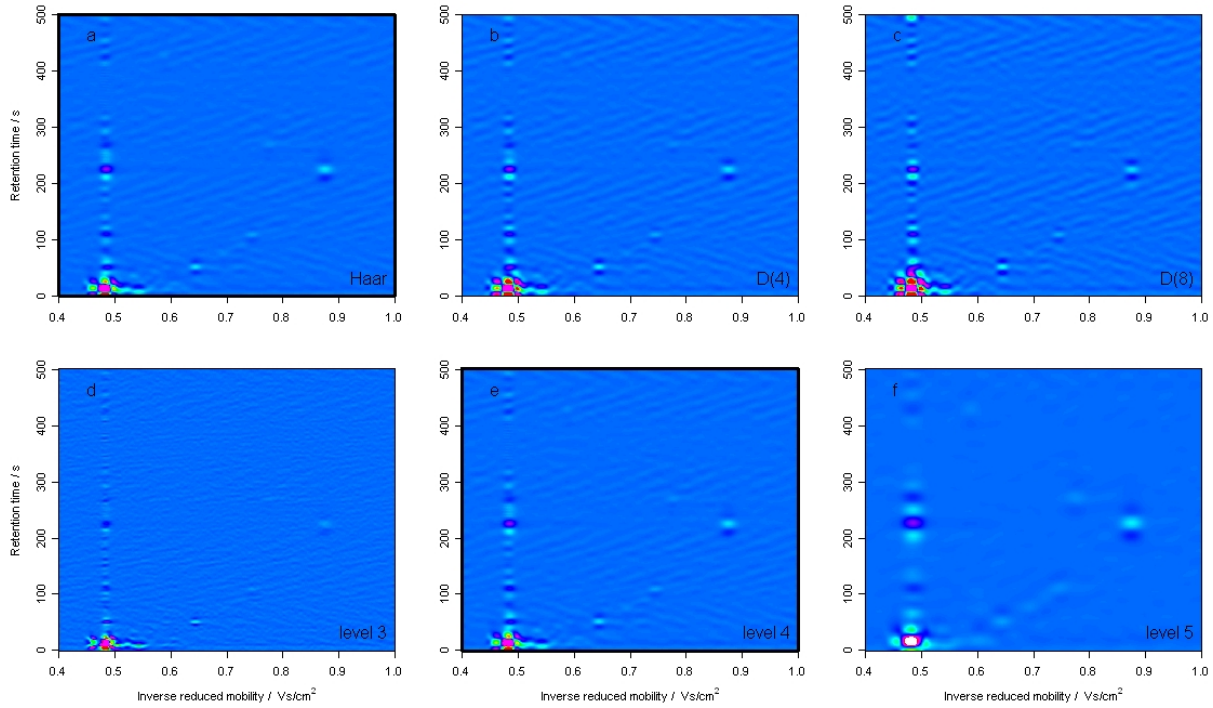

Figure 6.15: Heatmaps of the LH matrices of a two-dimensional MRA decomposition of a spectra series by means of the MODWT method using the different wavelet functions of (a) Haar, (b) D(4), and (c) D(8) on level 4 (top row), and furthermore comparing the different decomposition levels (d) 3, (e) 4, and (f) 5 for the usage of the Haar wavelet (bottom row): best performance for peak detection aspects is obtained using the Haar wavelet, since it does not cause as many overshooting artefacts as other functions, which is especially obvious in the early retention time range around the position of the RIP (0.49Vs/cm^2); in addition, the combination with the consideration of level 4, which covers the frequency range most related to that of the peaks, a good representation of the spectra series is reached.

whole measurement, while the ellipse extents were chosen as $a = 0.01$ and b as the half maximum retention time.

As the values of the MRA matrices were not comparable with the original signal intensities, the matrix had to be normalised by a height factor f_h^{MRA} to allow for a proper quantification of peak heights. All peaks with an inverse reduced mobility position beyond 0.7 Vs/cm^2 were used when determining this factor, since no influence by the RIP tailing was expected in the original data for their height. For each of these peaks, a 64×64 matrix around the according peak position was considered in the data matrix that was base of the MRA; for peaks at the margins of the measurement space, the matrix was chosen smaller where the margins were reached. This matrix was then denoised to obtain a realistic value for the original peak height and the resulting values were divided by the intensity observed for the according peak in the MRA matrix. The third quartile of all

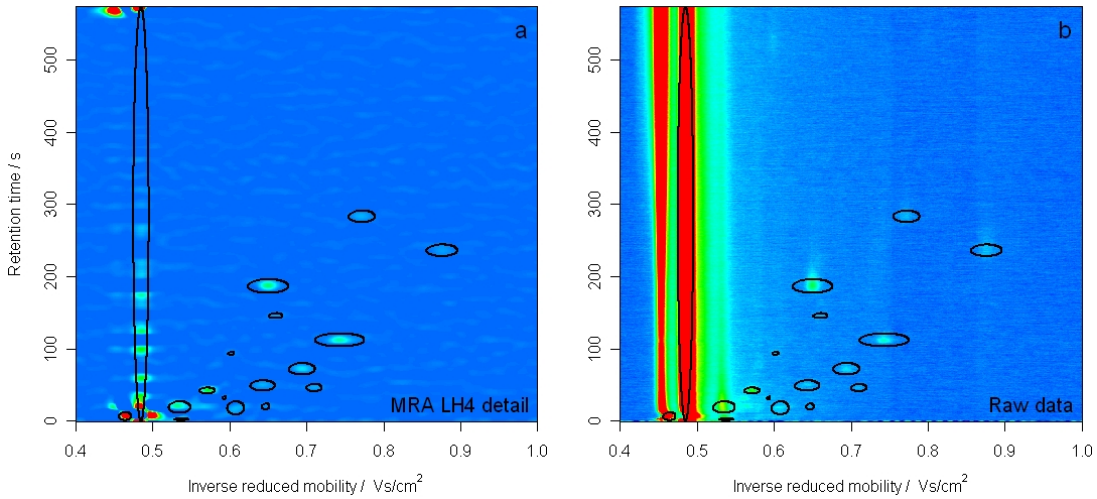

Figure 6.16: Detected peak ellipses in a heatmap of (a) the MODWT MRA LH4 matrix using the Haar wavelet after height transform and adjustment of overshooting artefacts, and (b) the raw data, showing a good agreement of the resulting peak list with the apparent peaks.

these ratios was chosen as the factor f_h^{MRA} and the MRA matrix as well as the heights in the peak list were updated by the multiplication with this value.

After this transformation, the MRA matrices could be plotted with the same choice of axis ranges as the original data along with the detected peak positions, where setting negative values in the MRA matrices to zero furthermore avoided a confusion about overshooting artefacts (Fig. 6.16).

6.3.4 Combination of multiple resolution levels

Depending on the measurement conditions influencing the peak width in the two time dimensions, a single level of the MRA sometimes did not contain all relevant information of a spectra series. While the higher level details of an MRA containing the low frequency information were insensitive to small peaks at the beginning of the retention time range, the lower level details were blind for larger peaks. It was, therefore, reasonable to consider a combination of several LH matrices and their result of the GIM algorithm to achieve a more sufficient overall peak list (Fig. 6.17).

To achieve a peak detection result based on the GIM outcome for several MRA details, starting from the lowest frequency level containing information about the broadest peaks,

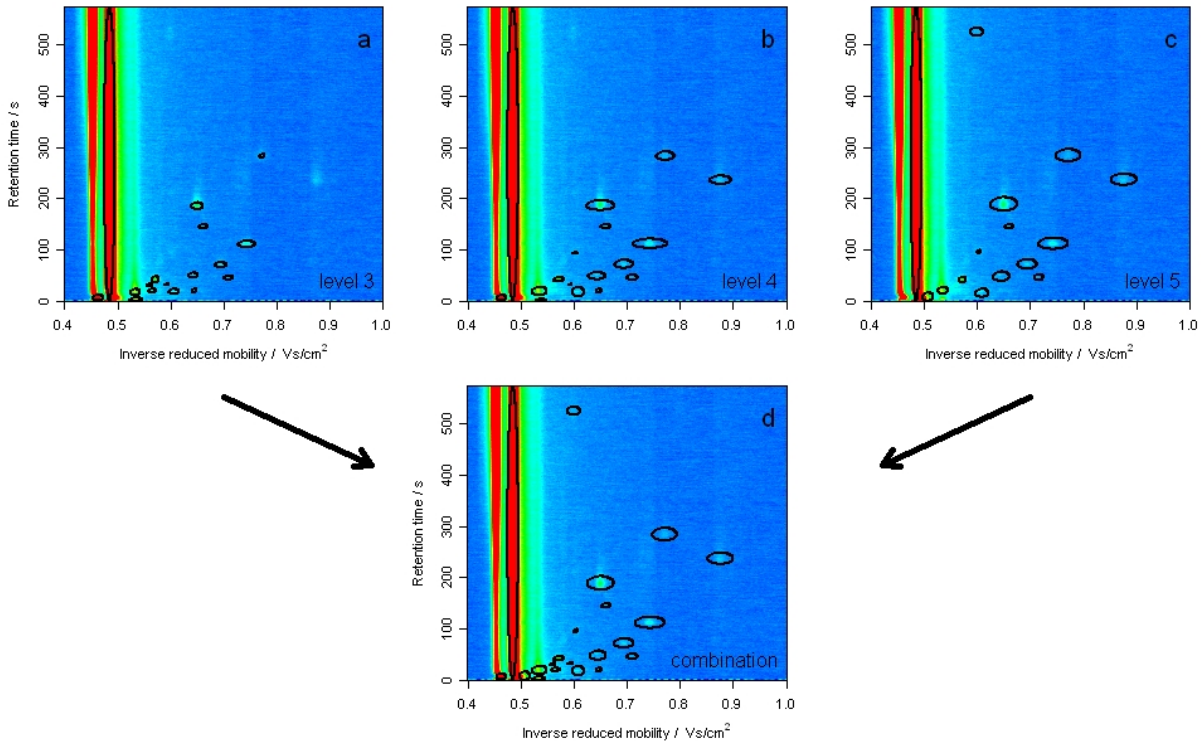

Figure 6.17: Heatmaps of the raw data with the peak ellipses detected in the MODWT MRA LH matrix of level (a) 3, (b) 4, and (c) 5 respectively, and (d) showing the peak ellipses resulting from the combined list of all three levels: while level 3 is blind for larger peaks, level 4 and 5 miss to detect some of the smaller peaks in the beginning of the retention time range – the combination, however, gives a sufficient characterisation of the entire measurement.

the peak list of every level was compared with the one for the subsequent level. The comparison

$$\frac{(x_0^L - x_0^H)^2}{(\sqrt{\pi}a^L)^2} + \frac{(y_0^L - y_0^H)^2}{(\sqrt{\pi}b^L)^2} < 1$$

was, therefore, used to identify, which peak positions, (x_0^H, y_0^H), of a higher frequency level fall into the ellipses of the previous stage, defined by the ellipse parameters (x_0^L, y_0^L, a^L, b^L). The ellipse extents a^L and b^L were extended by the factor $\sqrt{\pi}$ to respect minor variations of peak positions between the MRA matrices.

Connecting the different peak lists, ellipses of the lower frequency level containing no new positions were unchanged in the updated peak list. If an ellipse contained one peak position from the peak list of the next level, the old ellipse was substituted if the new ellipse corresponding to the included peak position constituted a larger area. In the case that one ellipse contained more than one peak position of the comparative peak list, its row in the peak list was substituted by all the rows of the new peak list corresponding to

these peak positions. Peak positions of the higher frequency level that were not located in any of the ellipses from the level before were joined with the updated peak list.

In this way, a new peak list containing the peak information of several frequency levels was created, yielding a better detection of peaks, if the peak width varied considerably within a spectra series.

6.3.5 Results

The combination of an MRA and the GIM method, applied to raw data after transformation of the two time axes for alignment aspects, yielded improved results compared to an application of the GIM algorithm to data that were preprocessed with a combination of axis transformations, RIP detailing, as well as wavelet smoothing and denoising. In contrast to the procedure previously introduced (Section 6.2), shoulder peaks without independent maxima can now be detected.

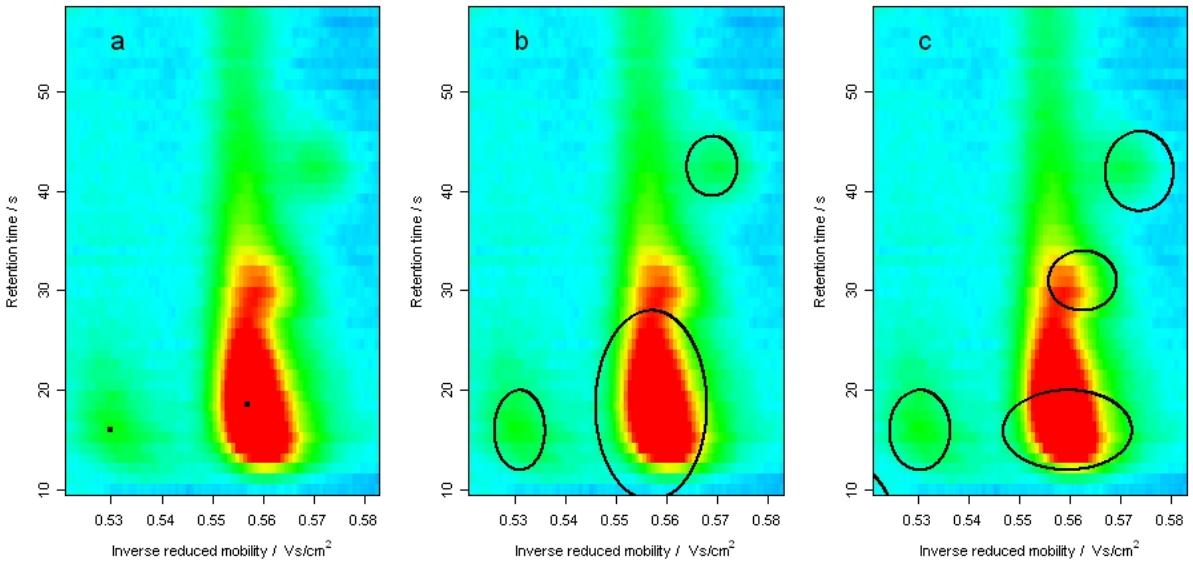

Figure 6.18: Comparison of the peak detection results of the three methods of (a) merged peak cluster localisation, (b) growing interval merging, and (c) the wavelet-based peak detection based on multiple resolution levels: while the first method provides no information about the expanse of a peak and is not able to resolve peaks that are not baseline-separated, the second approach gives a better peak characterisation by ellipse parameters and also allows for a more sensitive peak detection, but still does not recognise shoulder peaks, which is achieved by the third method giving a sufficient peak detection result for the entire measurement.

Comparing the peak detection results of the three methods of MPCL (Section 6.1), GIM (Section 6.2), and the wavelet-based peak detection based on multiple resolution levels, the first noticeable benefit of the two latter methods compared to the MPCL procedure is the better characterisation of detected peaks, which was enhanced from two position and an optional height parameter, providing no information about the expanse of a peak, to the additional values of ellipse extents and area (Fig. 6.18). In addition, the MPCL algorithm could not resolve peaks that are not baseline-separated (Fig. 6.18 a), whilst the GIM approach allowed for a more sensitive peak detection, although still not for the recognition of shoulder peaks (Fig. 6.18 b) which was achieved by the wavelet-based method giving a sufficient peak detection result for the entire measurement, especially if several frequency levels are combined (Fig. 6.18 c).

In conclusion, the developed peak detection approach lead to the generation of competent results. As a high sensitivity was seen as the most important factor for the considered applications, the appearance of some false positive peaks lowering the specificity of the method was acceptable. Additional artefact peaks were filtered out in continuative analyses by the clustering of peak positions to general peak areas, and screening for peak areas that were differentially expressed between groups (Chapter 7).

Data analysis

Chapter 7

Determination of general peak areas and further analyses

The introduced peak detection methods sufficiently characterise single measurements by peak positions and ellipses, respectively, however, the peak positions of the same analyte vary between measurements, as well as the set of peaks appearing in different measurements is changing. This complicates the comparison of different measurements and requires the definition of general peak areas allowing to generate a set of application-related peak variables for further analyses.

This goal is achieved by joining the entirety of peak characterisations by means of cluster analysis. A first approach is based on the peak positions resulting from the MPCL method, which are clustered and summarised to rectangular general peak areas (Section 7.1). This procedure is then enhanced to a method that yields ellipsoid peak areas on the basis of the ellipse parameters obtained by the GIM algorithm (Section 7.2).

Both methods can be used to construct new peak variables, calculated as the mean intensity values in the determined general peak areas. These variables were input for further analyses in two instance applications.

7.1 Peak regions based on peak position clusters

The MPCL algorithm (Section 6.1) characterises single spectra series by a list of peak positions and heights, but a set of variables comparable between the measurements of a study is required to use these values in a continuative analysis. This goal was achieved by clustering the entirety of peak positions belonging to the measurements of an investigation,

in this example data from a breath analysis study constituted of measurements of exhaled air from lung cancer patients and healthy control subjects (Subsection 7.1.1). The resulting clusters gave the basis for the constitution of general peak areas, which were used to define a set of peak variables, allowing for the discriminative analysis of the data (Subsection 7.1.2).

7.1.1 Procedure

To create a raster of general peak areas, the entirety of peak positions detected in all measurements of the study was examined simultaneously in a cluster analysis (Bader et al., 2006). Because the position values in the dimensions of inverse reduced mobility and retention time possessed considerably different magnitudes, an empirical standardisation was executed, ensuring the equal influence of both variables on the constituted cluster solution (Equ. 4.1, page 40). Furthermore, several choices had to be made to find an appropriate distance measure and clustering method for this application, as well as the optimal cluster number.

The distance measure, required for the comparison of classification objects which coincided with the detected peak positions, was chosen as the Euclidean distance (Equ. 4.3, page 41), as this metric measure offers the advantages of translation invariance, independence from the choice of origin, and invariance considering orthogonal transforms. Furthermore, this decision allowed the direct comparison of all clustering methods previously introduced, as its use is implied for the k-means and Ward's method.

Subsequently, a cluster procedure had to be chosen, defining the distance measure between the clusters (Subsection 4.1.3 and 4.1.4, page 42 and 46). Several techniques were compared, since it was not obvious which method performed best. The method of choice had to be appropriate for the specific data which consisted of 90 breath measurements from lung cancer patients and control subjects, yielding a total of 3341 peak positions, and thus constituting a large set of classification objects.

The k-means procedure is generally recommended for large data sets in an object-oriented analysis on the basis of a data matrix. Amongst the partition cluster techniques, it is more appropriately used with quantitative data than the k-mediods method, which is more robust, but results in a considerably higher computational cost (Kaufman and Rousseeuw, 1990). One problem arising in the partition cluster method, was the dependency on the starting values of the algorithm, as the resulting cluster structure notably varied according to the initial partition, which is often randomly chosen from the range of cluster objects.

To circumvent this inconsistency the k-means procedure was combined with the results of hierarchical clustering methods, giving an input for the starting values by the calculation of the mean position of each of the created clusters on the stage corresponding to the cluster number k. For the hierarchical cluster procedures, the average, weighted average, and Ward's method were potential choices for this specific problem, as they are more robust against outliers than, for example, the complete and single linkage methods, which are only appropriate for small sets of classification objects (Punj and Stewart, 1983).

The three hierarchical methods of average linkage, weighted average linkage and Ward's method, as well as the partition clustering method k-means combined with the average linkage and Ward's method were included in the method comparison. As quality criteria, allowing for a decision on the best performing method, the variance ratio criteria and the average silhouette width (Subsection 4.1.5) were computed for cluster solutions consisting of up to 500 clusters. Concurrently, these figures of merit served as a measure for finding the optimal number of clusters in the resulting partition (Fig. 7.1).

Comparing the different methods, the weighted average linkage scored worst for both criteria, which was the cause for not taking its results as initial values for the k-means algorithm. The combined procedures of k-means with average linkage and Ward's method respectively outperformed the use of the sole application of the hierarchical methods and

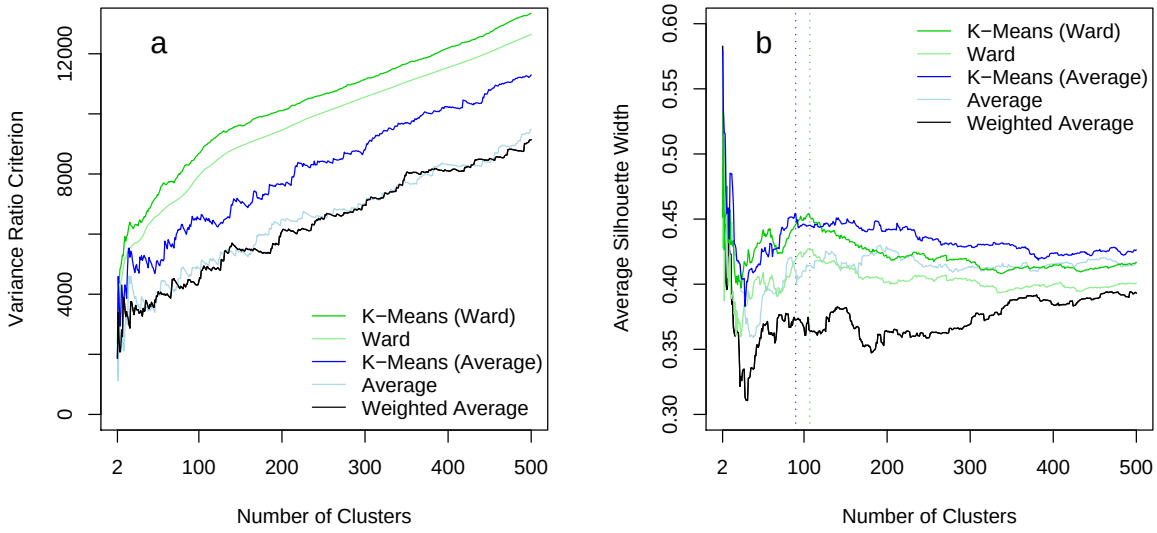

Figure 7.1: Plot of the values of (a) the variance ratio criterion (VRC) and (b) the average silhouette width $\bar{s}$ for solutions consisting of up to 500 clusters using five different cluster procedures: the combined methods of the k-means procedure with the average linkage and Ward's method show the best results; while a similar optimal performance (marked by dotted lines) for these two procedures is obtained for $\bar{s}$, the VRC clearly evaluates the combination with Ward's method best.

were, therefore, closer examined. For the variance ratio criteria the combined application of k-means and Ward's method was clearly better than the combination with the average linkage; a conclusion concerning the optimal number of clusters, however, was harder to draw on the base of this index (Fig. 7.1 a). The average silhouette width, on the other hand, yielded the best values for the first few cluster solutions for all methods, but as these were poor in adapting the data structure for the considered problem, and because of extremely low values of the variance ratio criteria, these clusterings were excluded. Factoring out the beginning of the range of cluster numbers, the maxima found were $\bar{s}_a = 0.454212$ for the method with initial values from the average linkage with $k = 89$, and $\bar{s}_w = 0.454197$ for the combination with Ward's method and $k = 106$ (Fig. 7.1 b). Since the optima of the two joined methods were almost equal, the k-means procedure with starting values of Ward's method and $k = 106$ clusters was chosen to be optimal due to better performance regarding the variance ratio criteria (Fig. 7.2 a).

According to this optimal cluster solution, rectangular areas were composed containing all restandardised peak positions belonging to a cluster, to construct general peak variables that are comparable between different measurements. For the determination of the limits of these areas the minima and maxima of all peak positions in a cluster were considered for

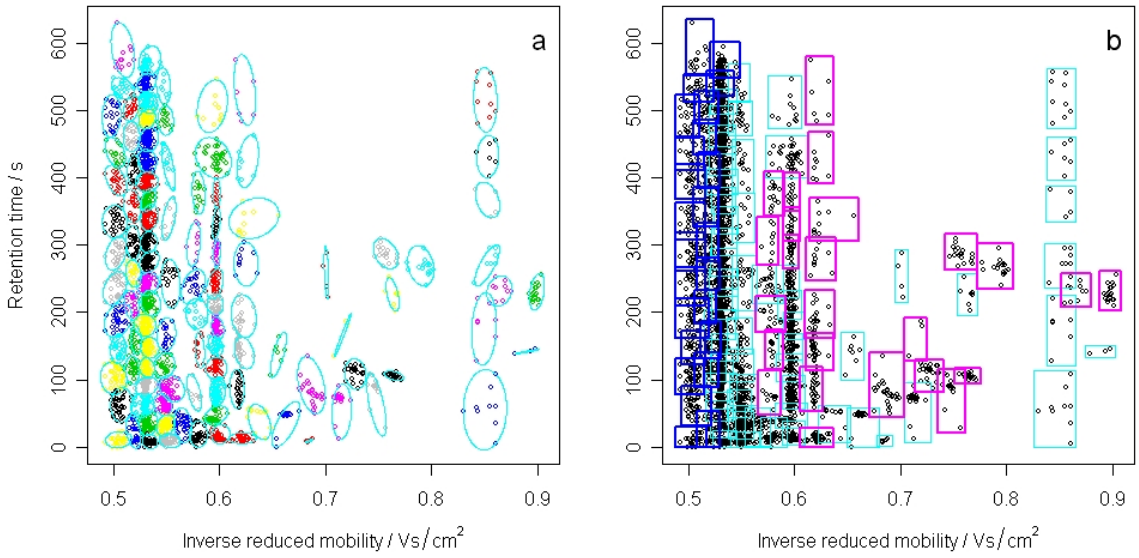

Figure 7.2: Scatter plots of all peak positions showing (a) the optimal cluster solution using the outcome of Ward's method as initial values for the k-means method with $k = 106$, where points belonging to the same cluster are drawn in the same colour and surrounded by an ellipsoid hull, and (b) the raster of general peak areas adjusted to the optimal cluster solution, where areas in the lower inverse reduced mobility range that were excluded during the descriptive analysis are marked in blue and those that correspond to peak variables that are differentially expressed between the groups are drawn in magenta.

the two dimensions. A marginal widening was applied to ensure that peaks belonging to positions at the border of a cluster were almost completely covered by the related general peak area, although the clustered peak positions were determined as peak centroids before. Respecting the different scales of both dimensions, a margin of 0.04 Vs/cm^2 was, therefore, added for the inverse reduced mobility, and the limits of the retention time were enhanced by 6 s. The resulting raster of general peak areas shows overlaps of different rectangulars, which was accepted, since these also appeared among real peaks, building the basis of this clustering (Fig. 7.2 b).

After the raster of rectangular peak areas was obtained, general peak variables could be computed as the mean intensity values in the defined limits of each peak area. Furthermore, the mean intensity for all measurement parts not belonging to any of the peak areas was calculated, yielding an additional variable. Doing this for each measurement, a base for further analyses was given by a vector of length $k + 1 = 107$ that was comparable for the entire study.

7.1.2 Discrimination of lung cancer patients and control persons

The newly created general peak variables served as an input for the analysis of the human breath study on lung cancer patients and control persons by means of a linear discriminant analysis (Bader et al., 2005; Baumbach et al., submitted in 2007; Westhoff et al., submitted in 2008). As the sample size of $n = 90$ measurements was smaller than the number of $k + 1 = 107$ peak variables, a variable selection was necessary, achieved by a descriptive analysis of the data concentrating on the appearance of missing values, and a multiple test procedure screening those variables which showed a significant difference between the breath measurements of the two groups. Subsequently, a concluding discriminant analysis could be arranged on the basis of the differentially expressed variables.

In the descriptive analysis of the data, 1417 missing values were found, obtained if the limits of a peak area were lying outside the observed measurement range. The two reasons causing this effect were firstly the varying cut-off point after the RIP, yielding different starting points of the inverse reduced mobility axis of the data matrix after preprocessing, and secondly the varying duration of the generation of 501 spectra for different measurements by unknown reasons influencing the end of the retention time range. In the area after the RIP, missing values appeared mainly in the control group, while altogether only 20% belonged to the lung cancer samples. Exploring the reasons behind this, the 5 point summary of the extreme values of the RIP (Table 7.1) showed that the peak tended to

Table 7.1: 5 point summary of the extreme values of the reactant ion peak considered for the group of lung cancer patients and control persons, as well as the room air at the two measuring places: the RIP height tended to be generally higher for the measurements generated in Dortmund than for those taken in Hemer.

	Minimum	1. Quartile	Median	3. Quartile	Maximum
Control group (Dortmund)	2.71	2.93	3.01	3.07	3.38
Lung cancer (Hemer)	2.01	2.65	2.7	2.91	3.02
Room air (Dortmund)	2.9	2.94	3.03	3.14	3.21
Room air (Hemer)	1.76	2.11	2.77	2.83	3.15

be higher in the control group than for the lung cancer samples, causing higher values of the cut-off point after the RIP.[1] As the same observation was made comparing the room air measurements at the ISAS in Dortmund, where the control measurements were generated, with those taken at the lung hospital Hemer, a correlation with lung cancer was disregarded for this systematical difference between the measurements. Although the inclusion of the entire measurement range for the calculation of the peak variables could have avoided the appearance of these missing values, the procedure was not adjusted to this issue, because a result not influenced by the RIP was aspired. In the area of late retention times, on the other hand, where two peak variables were found to contain a high amount of missing values, no compensation was possible, as the data after the last spectra were not available.

Excluding 25 variables in the area after the RIP and in the latter retention time dimension (marked by blue corresponding rectangulars in Fig. 7.2 b), no missing values were left for the control group, and only 61 remained for the lung cancer samples, which were spread over just three measurements. Two of these measurements were interrupted after 401 and 430 spectra, leading to the appearance of 8 and 2 missing values respectively. The third measurement contained an extreme shift of the RIP to the right, causing 62% of the peak variables to be missing; for this reason this measurement was excluded from the rest of the analysis.

After the descriptive analysis, the data consisted of 82 peak variables observed in 54 control and 35 lung cancer subjects. This data was subsequently analysed in a multiple test procedure to identify differentially expressed variables between the groups. For the comparison of two groups, the two most common test methods are the t-test and Wilcoxon test. Unequal variances were assumed, according to the distribution of the peak variables

[1] This effect was not found for the remaining areas of the measurements.

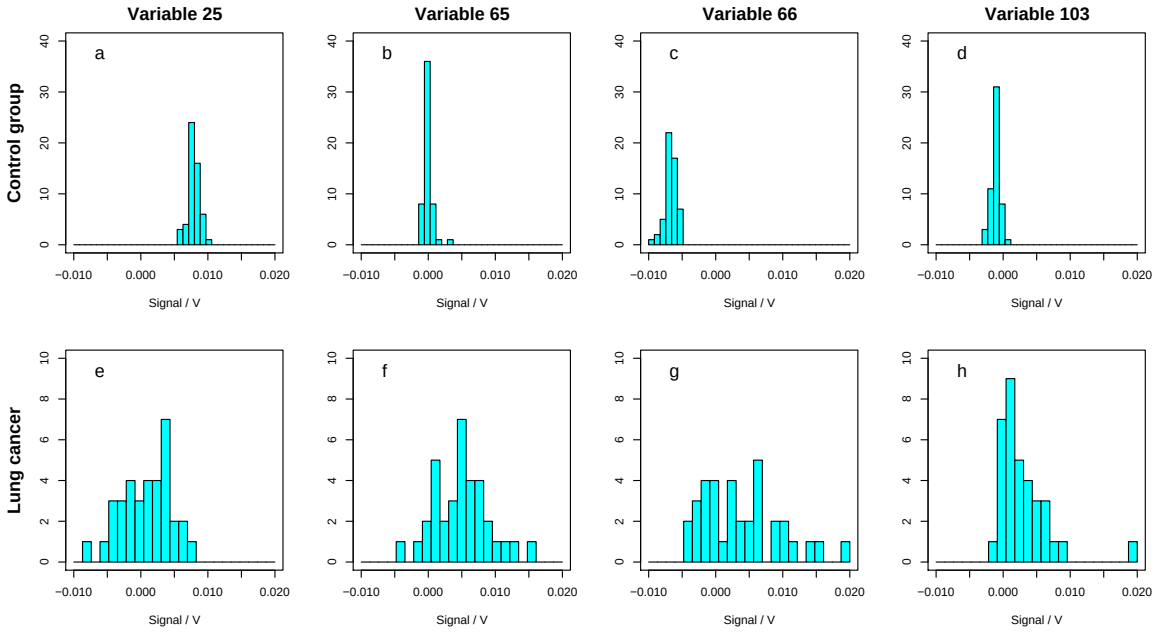

Figure 7.3: Histograms of the observed values for four instance variables (a-d) in the control group and (e-h) for lung cancer patients: the distributions appear to be not generally symmetrical and the variances differ considerably between the groups.

in histograms, where a symmetrical behaviour required by the non-parametric Wilcoxon test was not observed (Fig. 7.3). At the same time, this was in contrast to the postulate of a normal distribution implied by the t-test, but as the sample size was sufficiently large, an asymptotic normal distribution could be assumed because of the central limit theorem. The level α could, therefore, be controlled if the $(1 - \alpha)$ quantile of the t-distribution was substituted by the one of the standard normal distribution. However, because of the dependency of the different tests, and the fact that all variables were measured on the same samples, a multiple test problem had to be considered (Section 4.2.2, page 52). Using the Bonferroni-Holm method (Holm, 1979) with the multiple test level $\alpha_k = 0.001$, 25 variables were found to be differentially expressed between the groups (Table A.1, page 147), marked by magenta corresponding rectangulars in Fig. 7.2 b. For these peak variables only one missing value was observed in one lung cancer patient, whose value was substituted by the mean value of this variable in his group, to allow a further analysis of the measurement.

Having enhanced the data reduction from 107 to 25 differentially expressed variables, a discriminant analysis could be examined. As the two groups had considerably different a priori probabilities in the underlying population, the Bayes decision rule was applied which is also optimal regarding the overall error rate (Def. 4.2, page 56). As lung cancer is a rare and severe disease, it was reasonable to use the inversely proportional cost function

to increase the cost for a wrong assignment of objects from classes with low a priori probabilities (Equ. 4.10, page 57). Combining the Bayes rule with this cost function, the a priori probabilities π_g, $g = 1, 2$ as well as the factor C appearing in the cost function C_p were canceled out, yielding the Bayes rule with equal a priori probabilities, coinciding with the maximum likelihood rule and, therefore, possessing the property of cost optimality (Def. 4.3, page 56). The resulting decision rule, assuming equal variance for both groups, coincides with the classification by means of Fisher's linear discriminant function, which has much weaker assumptions, for example on variances and distributions, than the decision theoretical approach. The equivalence of both rules, however, signifies a broad applicability of the linear discriminant analysis. The assumption of equal variances for both groups could, therefore, be neglected for the analysis of the considered data. Estimating the parameters of the distributions of the variables, the usual unbiased estimators of groupwise mean $\bar{\mathbf{x}}_g$, $g = 1, 2$, and the pooled variance $\mathbf{S}_p^2$ were determined by means of the method of moments.

Using these estimators with the maximum likelihood decision rule, a discriminant function was formed on the basis of the 25 selected variables, whereas the application of the leave-one-out method for the combination of the t-test procedure and the LDA resulted in an error rate estimation of only 0.011. To further increase the goodness of the LDA by reducing the complexity of the model, a stepwise variable selection was performed, evaluating the benefits of excluding the different variables by the resulting leave-one-

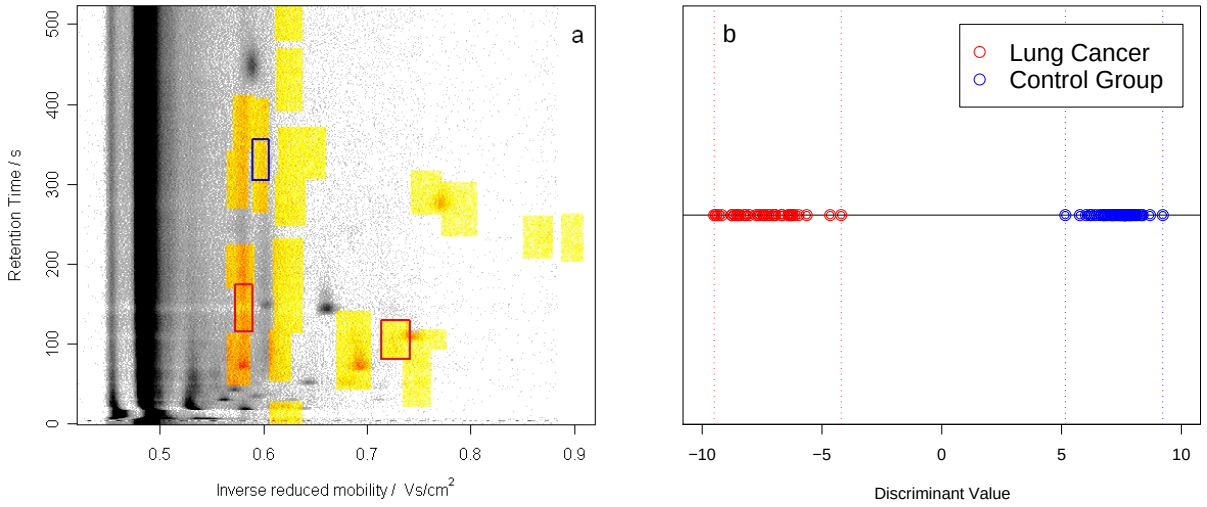

Figure 7.4: Picture of (a) a heatmap highlighting the measurement parts that constitute the peak variables giving the base for the discriminant analysis, where the three areas with the highest influence on the decision rules are marked, and (b) a plot of the resulting discriminant values for the entire study: the two groups of lung cancer patients and control persons could be distinguished perfectly.

out error rate estimation and the values of the discriminant coefficients standardised with the pooled standard deviation, giving an idea of the influence of a variable on the discrimination. For three of the variables a perfect discrimination between the groups was observed after just one step of the downwards selection (Table A.2, page 151). As the variable V_{103} possessed the highest p-value, giving a univariate check of the discriminative properties, and furthermore the lowest standardised discriminant coefficient indicating a low influence of the variable on the resulting classification, this variable was chosen to be excluded in the final discriminant analysis. Although the estimated error rate of zero was also observed with a much lower number of variables, the stepwise procedure was stopped, since as many influencing variables as possible were aimed to be kept in the decision rule, to allow for later biological interpretation after identification of analytes corresponding to the relevant measurement parts (Fig. 7.4 a). The final discriminant function (Table A.3, page 152) resulted in a perfect classification of all persons in the considered data set (Fig. 7.4 b).

The discriminant coefficients were considered to evaluate which variables have the highest influence on the determined decision rule. As their magnitude is dependent on scaling effects, a standardisation by multiplication with the corresponding pooled standard deviation was required. For the considered data, 3 of the standardised discriminant coefficients were larger than 1 (Table A.2, page 151), and, therefore, had a major influence on the result of the LDA. The corresponding peak areas were marked by coloured frames in Fig. 7.4 a, where high intensity values for the variable marked in blue (V_{25}) are relevant for the assignment to the control group, while features in the peak areas marked in red (V_{65}, V_{66}) are associated with lung cancer.

Comparing these results with the outcome of parallel studies on the emission of bacteria, it turned out that the peak variables V_{65} and V_{66}, found to be related with the assignment of classification objects to the group of lung cancer patients, correspond with areas containing analytes found in the measurements of *Escherichia coli* cultures. As the immune system of cancer patients is weakened significantly, it is more likely that they are affected by bacteria, which can yield a bias of the result of human breath measurements by the emission of lung bacteria. A frequent occurrence of bacteria could be expected for the lung cancer patients especially in the data analysed in this study, as the lung hospital Hemer treats severe cases of bronchial carcinomas and hospitalisation is often related with bacterial infection.

The possibility that at least part of the discriminatory information is related to bacterial infections rather than to the carcinoma itself can, therefore, not be discarded. Indepen-

dently of this potential restriction of the determined decision rule, the introduced processing strategy for spectra series enabled the successful characterisation of differences between the two groups of samples.

7.2 Peak regions based on ellipse parameters

Although the introduced approach for the definition of general peak areas and corresponding variables already allows continuative analyses of the entirety of determined peak lists of whole studies, a further enhancement of this method was aspired to. Based on wavelet-derived MRA details the ellipse representations resulting from the GIM method allowed the construction of ellipsoid general peak areas, giving a more appropriate characterisation of typical measurement features.

This new procedure was practically introduced for the breath measurements of patients of various lung diseases at the lung hospital Hemer to determine a collection of general peak areas for breath analysis applications, used to create new peak variables as mean intensity values in these areas (Subsection 7.2.1). The resulting variables were then used to discriminate between different forms of lung cancer, namely circular focuses, endobronchial tumors, and other types of lung carcinomas (Subsection 7.2.2).

7.2.1 Procedure enhancement

The construction of general peak areas was initiated by clustering of the entirety of peak positions equivalently to the introduced method based on the MPCL (Subsection 7.1.1), but now using the peak lists of an MRA-based application of the GIM algorithm. The findings for the choice of measures and methods could, therefore, be adapted from this approach, yielding the use of the k-means procedure with starting values derived from Ward's method, and calculating the average silhouette width for the evaluation of the optimal cluster number.

The input peak positions were still standardised, the standardised retention times now, however, weighted by the factor 0.5 to adjust for the higher variability in this direction. Furthermore, peak positions containing less than two points in their surrounding area, defined by ranges of 0.005 Vs/cm^2 for the inverse reduced mobility and 10 s for retention time around the considered position were excluded before cluster analysis. The optimal solution resulting from this procedure consisted of 148 clusters.

The optimal solution was subsequently characterised by ellipses to yield general peak areas from these clusters (Fig. 7.5). Here, two position parameters x_0^G and y_0^G were chosen as the mean of the quartiles of all peak point positions belonging to a specific cluster in the two dimensions, given by

$$x_0^G = \frac{x_{0_{0.25}} + x_{0_{0.75}}}{2} \quad \text{and} \quad y_0^G = \frac{y_{0_{0.25}} + y_{0_{0.75}}}{2}.$$

The extent parameters were determined as the mean deviation of the two position quartiles from the defined area positions x_0^G and y_0^G, summed with the median ellipse extents for the peaks belonging to a specific cluster, resulting in

$$a^G = \frac{x_{0_{0.75}} - x_{0_{0.25}}}{2} + \text{med}(a) \quad \text{and} \quad b^G = \frac{y_{0_{0.75}} - y_{0_{0.25}}}{2} + \text{med}(b).$$

These ellipse parameters defined general peak areas (Table A.4, page 154), giving the basis for the calculation of peak variables as the mean intensity for the values within each ellipsoid area, which was for the gth variable computed as

$$V_G = \bar{E} \quad \text{with} \quad E = \left\{ z_{ij} \,\middle|\, (i,j) : \frac{\left(x_i - x_0^G\right)^2}{a^{G^2}} + \frac{\left(y_j - y_0^G\right)^2}{b^{G^2}} < 1 \right\}.$$

Continuative analyses were enabled for breath analysis studies after determining these peak variables for each single measurement.

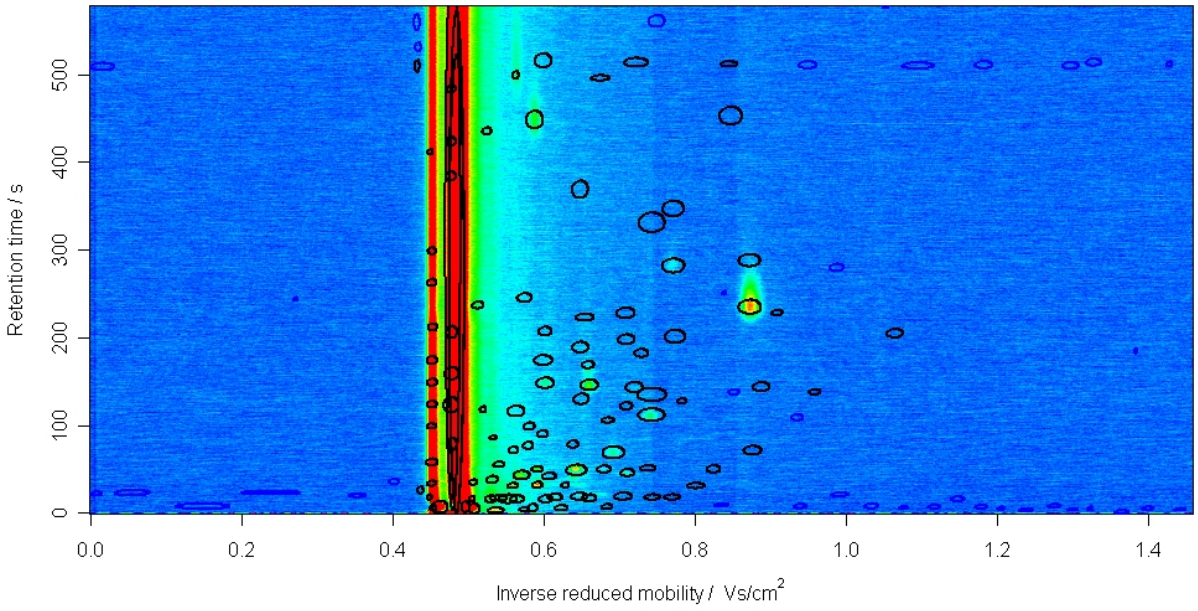

Figure 7.5: Heatmap of a breath measurement with the determined ellipsoid general peak areas: the ellipses drawn in black give the basis for continuative analyses, while the areas marked in blue were excluded in the descriptive analysis because of a large amount of missing values, or a constantly low response over the samples.

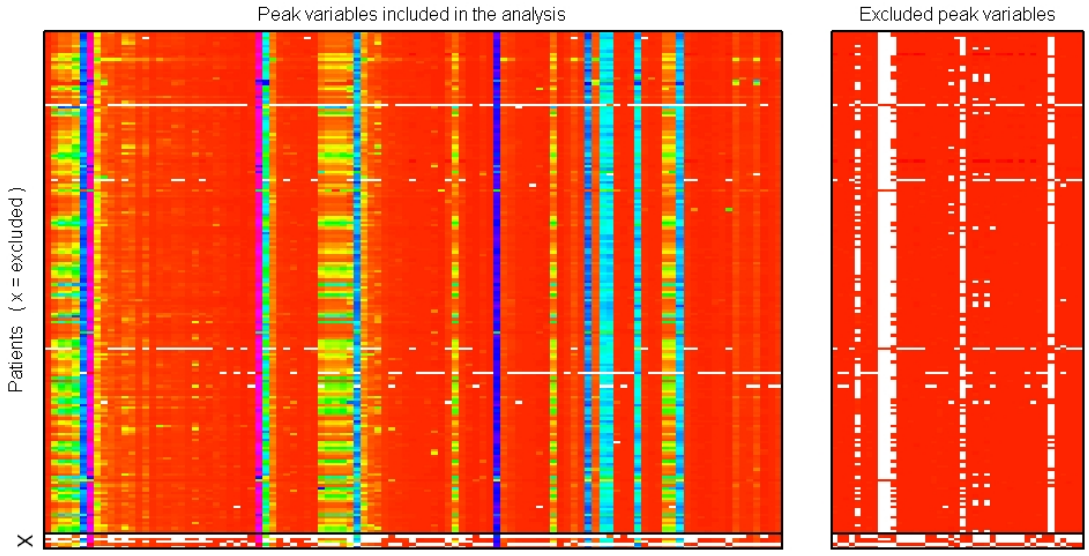

Figure 7.6: Heatmaps illustrating the peak variable values for all patients in the entire study, separated by variables and patients that are included in the analysis and those that were excluded in the descriptive analysis: red values correspond with low responses, higher values are encoded in either yellow, green, blue, up to purple; white spots indicate missing values.

Before the analysis of specific questions, a descriptive analysis was performed considering the peak variable intensity values of the data from the entire study. Concentrating on the identification of variables and patients with high amounts of missing values, 8 peak variables containing more than 20 missing values, as well as 5 patients whose measurements contained between 45 and 115 missing values, were excluded from the further analysis. Afterwards, only 51 out of 1313 missing values were left. Additionally, 35 variables with constantly low responses, therefore constituting peak areas likely to be based on noise artefacts, were also excluded from the continuative analysis (Fig. 7.6). These peak variables were identified by comparing their maximum values with the median of all the peak variable values, remaining in the study after the exclusion of variables and patients with many missing values. The remaining set of 105 peak variables gave a basis for the separate investigation of different questions.

7.2.2 Comparison of different forms of lung cancer

After the separation of lung cancer patients from a control group was previously investigated, the aim was now to compare the profile of patients with different forms of lung cancer. The set of measurements in this study was, therefore, constrained to a subset of three groups, containing 10 patients with circular focuses at an early stage of cancer

development, 21 patients with endobronchial tumors, and 9 patients with other types of carcinoma (Fig. 7.7), analysed in a pairwise manner.

Starting with the matrix of the selected peak variables for each pairwise comparison, a t-test procedure was performed to screen for variables that were differentially expressed between two of the groups (Table A.4, page 154). As the p-values were considerably high, a multiple testing correction could not be integrated, since none of the variables would remain. Using a leave-one-out procedure for the estimation of the error rate, overall significant variables were, therefore, defined as those who possessed a p-value below 0.05 for each test across the patient range. These peak variables gave the basis for the following LDA, where patients possessing missing values for any of these variables were excluded from the analysis.

The LDA applied for the pairwise separation of the different groups was performed using the same settings as in Subsection 7.1.2, estimating the error rate by the leave-one-out method and improving the discrimination between groups in a stepwise selection of variables. Here, at each stage of the procedure the variable yielding the lowest estimated error rate, when left out of the determination of the discriminant function, was excluded from the analysis. If the lowest error rate value appeared for more than one variable, the standardised discriminant coefficients excluded the variable with the lowest value, as this corresponds with little discriminatory information (Table A.5, page 159). This procedure

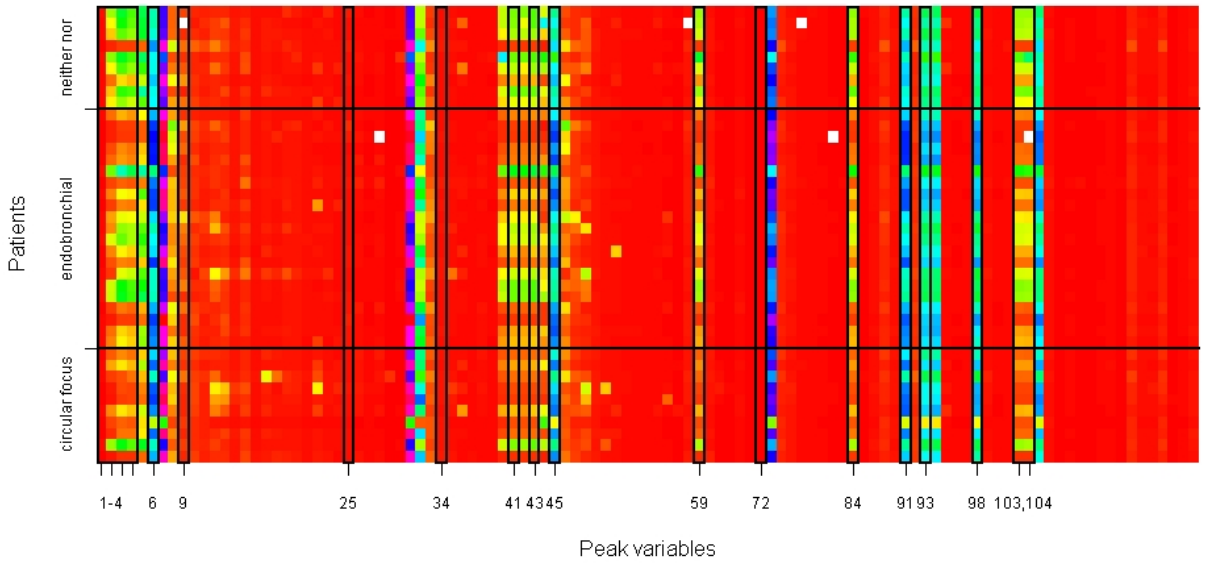

Figure 7.7: Heatmaps illustrating the values of peak variables remaining after the descriptive analysis for the study subset of lung cancer patients: the three groups of circular focuses, endobronchial tumors, and other carcinomas ('neither nor') are indicated on the vertical axis, while relevant variables for this analysis are marked by ticks on the horizontal axis.

was continued until no further improvement was achieved and the remaining variables used as the input for the final discriminant functions. Peak variables were illustrated in density plots comparing the distribution between the two groups, while corresponding areas were additionally marked in a heatmap as shown in Fig. 7.8 for the variables $V9$, $V25$, $V34$, and $V72$ used to separate patients with circular focuses and endobronchial tumors.

When comparing the different groups of cancer patients (Table A.7, page 163), the best differentiation was achieved between patients with circular focuses and those with other forms of carcinoma exclusive of endobronchial tumors (Fig. 7.9 b), where all patients were assigned to the correct group on the basis of the final discriminant function with an estimated error rate of only 5 %. The discrimination between groups suffering of endobronchial tumors, and those suffering from circular focuses and other carcinomas, showed a similar high estimated error rate of about 20 % (Fig. 7.9 a, c) for both comparisons, giving the impression that patients with endobronchial tumors were harder to separate from other groups than the other two sets of patients. This assumption was confirmed in a comparison between patients with circular focuses and a combined group including those with endobronchial tumors and all other forms of carcinoma. The differentiation showed an estimated error rate of 15 % (Fig. 7.9 d), where all four wrongly assigned patients belonged to the group of endobronchial tumors.

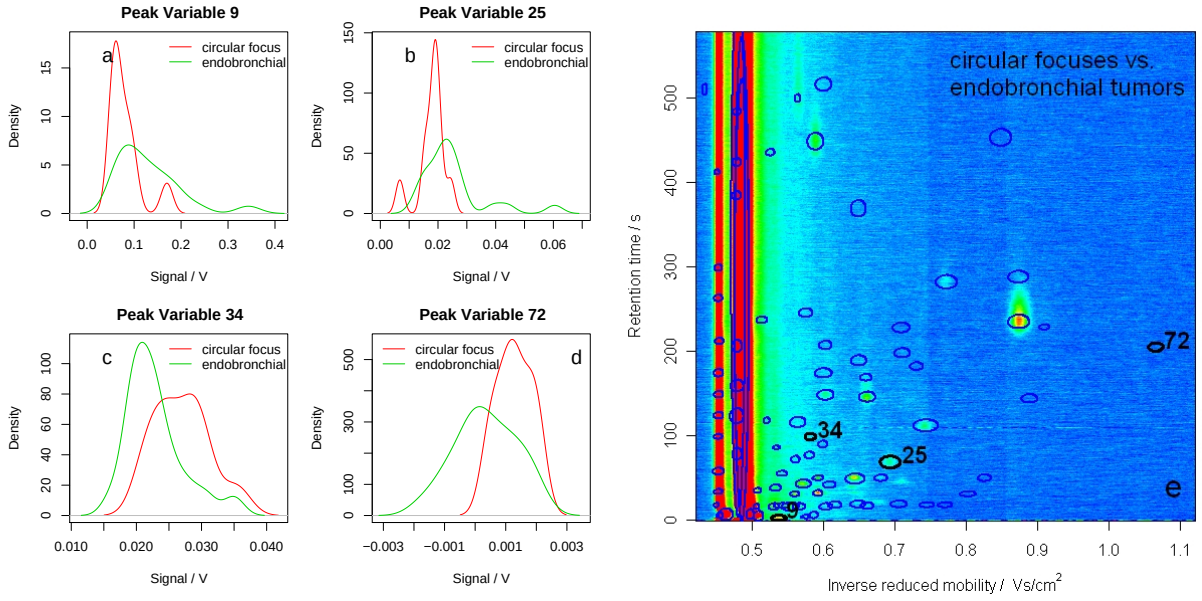

Figure 7.8: Plots overlaying the density curves for the two groups of patients with circular focuses and endobronchial tumors for the relevant peak variables of this comparison (a) $V9$, (b) $V25$, (c) $V34$, and (d) $V72$, which are also marked in (e) the heatmap of an instance breath measurement.

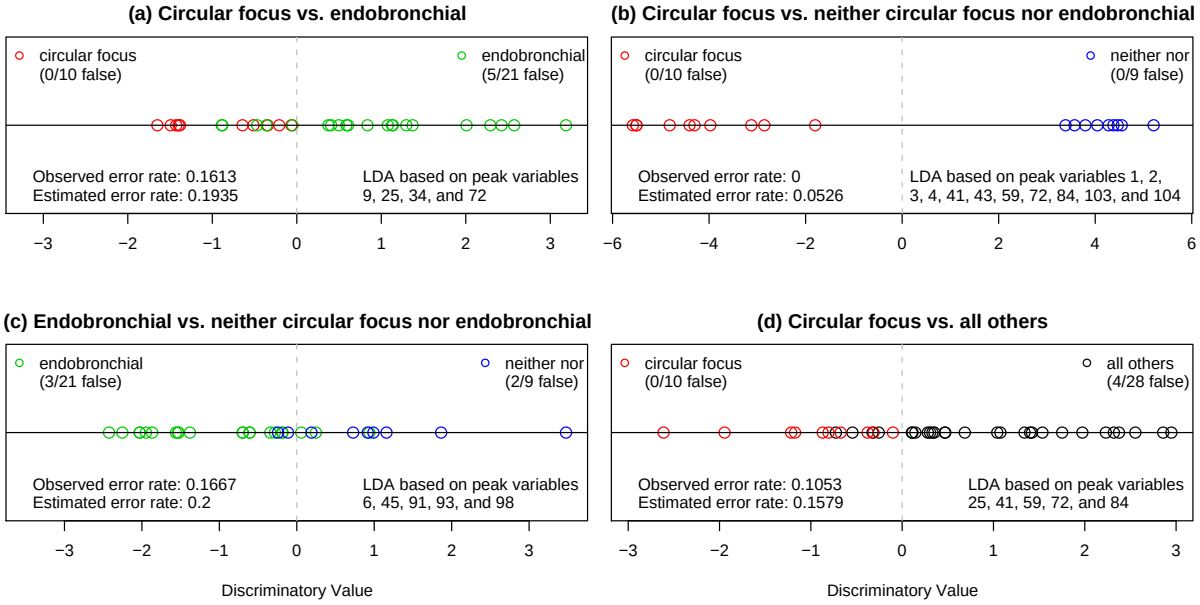

Figure 7.9: Results of the linear discriminant analysis for the four different comparisons of (a) circular focuses with endobronchial tumors, (b) circular focuses with other lung carcinoma exclusive endobronchial tumors, (c) endobronchial tumors with other lung carcinoma exclusive circular focuses, and (d) circular focuses with other lung carcinoma inclusive endobronchial tumors: the absolute values of falsely assigned patients are given as well as the observed and estimated error rates and the peak variables forming the basis of the discriminant function.

It was observed that the LDA for the last comparison was based on the peak variables $V25$, $V41$, $V59$, 72, and $V84$ (Fig. 7.10 b and B.1, page 165), which were also involved in the decision rules for the discrimination of circular focuses against the two groups of endobronchial tumors ($V25$, $V72$) and other carcinoma ($V41$, $V59$, 72, $V84$), separately. It could, therefore, be seen as an amalgamation of the discriminant functions, combining the two relevant measurement parts of the pre-RIP, on the left-hand side of the RIP, and the measurement part behind the RIP in the early retention time range. The first measurement part contains the basic features, influencing the separation of the patients with circular focuses and those with other forms of carcinoma besides endobronchial tumors (Fig. 7.10 a and B.2, page 166); whilst the second measurement part consists of the peak areas giving the input for the discrimination between circular focuses and endobronchial tumors (Fig. 7.8). The peaks yielding the separation between endobronchial tumors and other carcinoma exclusive circular focus on the other hand were all located in the RIP area (Fig. B.3 and B.4, page 167).

In summary, peak regions corresponding to the variables found to be relevant for the comparison of the different forms of tumors could be assigned to three main groups: the

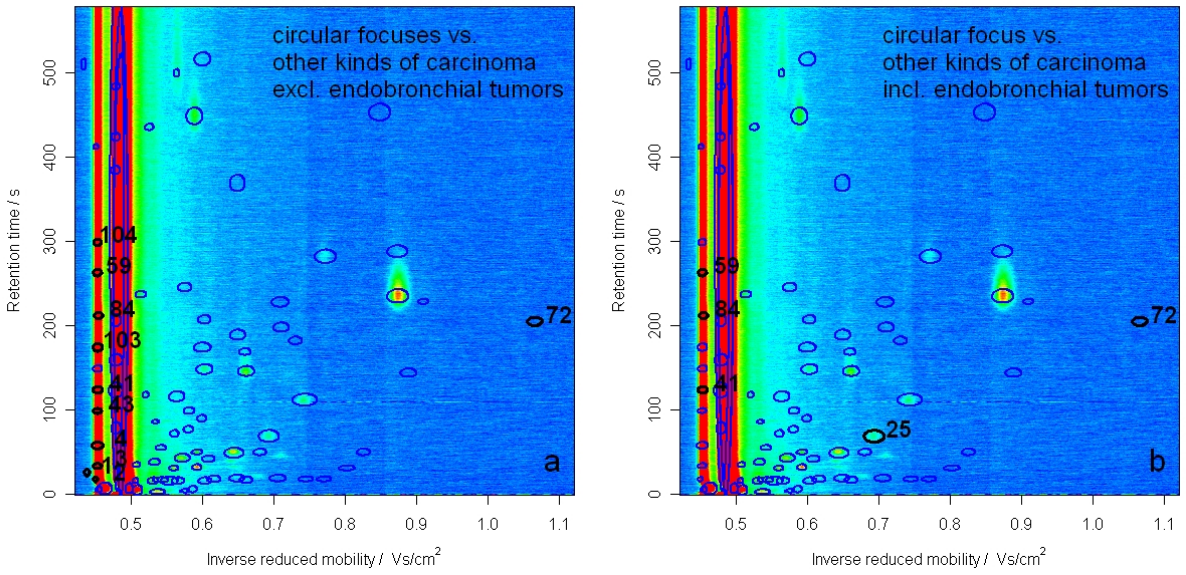

Figure 7.10: Heatmap of an instance breath measurement showing the general peak areas corresponding to the relevant variables for the comparison of (a) circular focuses with other carcinoma excluding endobronchial tumors and (b) circular focuses with all other carcinoma including endobronchial tumors.

group of ellipse areas lying in the pre-RIP, those located in the RIP, and those spread over the measurement area on the right-hand side of the RIP. The three groups of corresponding variables influenced three different comparisons respectively, while the fourth decision rule was based on a mixture of the first and the third group. No similarities were observed between any of the identified peak areas with those that showed to be most important in the comparison of lung cancer patients with control persons.

In conclusion, the main part of the peak variables which give the basis for the separation of groups lies in measurement parts, where relevant features were not expected a priori. Many variables are located in the pre-RIP or the RIP itself, where it was not directly apparent why specific positions of these long stretched peaks should contain discriminatory information. It might, however, be possible that the whole peaks were differentially expressed between the groups, which may influence the result in this way. This hypothesis was strengthened by the fact that more variables lying in the two peaks of RIP and pre-RIP were included in the original LDA, before the stepwise selection resulted in the exclusion of some of them to reduce the complexity of the model and thus improve the actual error rate. As the positive and negative discriminant coefficients, however, were spread over these variables randomly without an obvious pattern, this assumption could not be verified. On the other hand, the variables possessing the highest standardised discriminant coefficients and, therefore, giving the main part of discriminatory information,

were for the comparison of circular focuses with all other carcinoma exclusive ($V1$ and $V72$) and inclusive ($V25$ and $V72$) endobronchial tumors respectively found to be the two variables that did not lie on the straight line with all the peak areas falling into the pre-RIP. This might be a reason why patients with different forms of tumors were still sufficiently separated and only members of the group with endobronchial tumors were hard to assign to the correct group.

These results gave an interesting insight into differences in human air composition, not only between control persons and lung cancer patients, but also in different forms of lung tumors. Without the developed methods for preprocessing, peak detection, and the determination of general peak areas, specifying the relevant measurement parts for the comparison and differentiation of breath monitoring data would have been time-consuming, subjective, and insufficient, if even at all feasible. The introduced approaches, therefore, constitute a valuable contribution for the analysis of spectrometric data.

Chapter 8

Transfer to another spectrometric method

To prove a wider applicability of the developed methods, the introduced algorithms are transferred for use with another spectrometric method, namely the differential mobility spectrometry (DMS), which is introduced in Section 8.1. The analysis of the dependency of two analytical dimensions of a DMS coupling with pyrolysis-GC for measurements on bacteria cultures was used as an example, containing a high number of peaks within a dense assembly. For this, the specification of peak positions is required (Section 8.2), enabled by the application of the GIM algorithm to the three-dimensional spectra data after a preprocessing based on wavelet transforms is performed.

8.1 Differential mobility spectrometry

Differential mobility spectrometry (DMS) is a method for the characterisation of gaseous analytes using differences in mobility K (Equ. 2.1, page 13) between two electric fields. In DMS drift tubes, gaseous ions are formed from a sample through ion chemistry and are evaluated for the dependence of K between two extremes of electric field strength. The required drift tubes offer low costs, are insensitive to mechanical influences, and are available as miniaturised devices. Additionally, DMS offers the simultaneous characterisation of positive and negative ions using a single analyser (Schmidt et al., 2004).

More specifically, the electric field is in contrast to the IMS applied orthogonal to the ion flow with a supporting drift gas between two closely spaced electrodes, while alternating strong and weak electric fields are generated using a high frequency asymmetric field. For

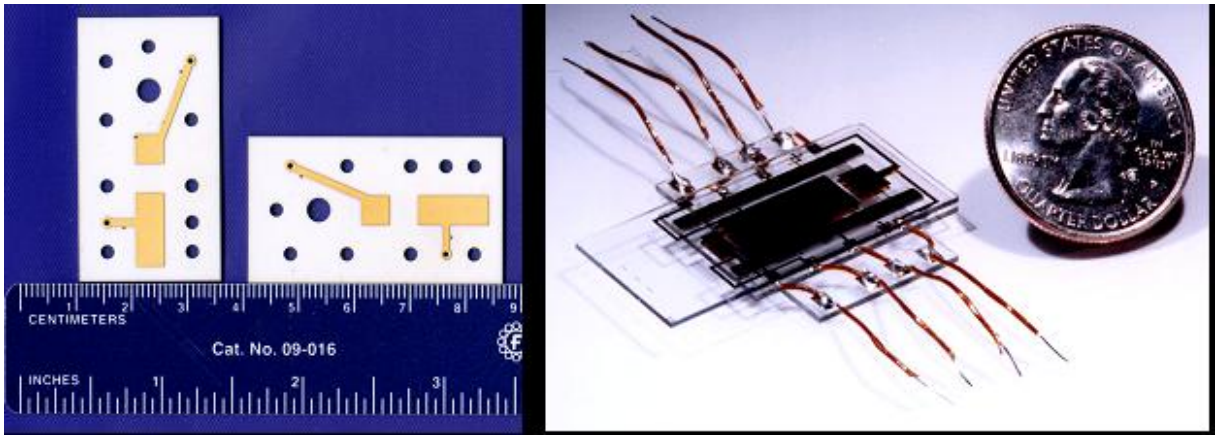

Figure 8.1: Illustration of the size and appearance of the parts of a planar DMS device (Picture provided by Dr. Gary A. Eiceman, New Mexico State University, USA).

DMS devices with planar drift tubes (Fig. 8.1) the ions are carried between two parallel-plate electrodes, where a high frequency electric field is applied to one plate and the other is grounded. Only ions with a certain field dependent mobility pass through the electrodes; others eventually collide with the electrode walls and are no longer detectable. The net migration of the ions can be corrected so that ions pass through the analyser to a detector using a compensation voltage, which is varied over a range of weak dc voltages to scan a differential mobility spectrum. Ions from a substance have a certain compensation voltage that is a measure of the magnitude of field dependence of mobility, giving a direct measure of the difference in ion mobility between the field extremes and is characteristic for fixed experimental parameters. Ions that pass from the drift tube are collected by two Faraday plate detectors floated slightly to a negative or positive potential.

Measurements in DMS analysers provide a low resolution, restricted technically by the aperture dimensions and under certain conditions also by ion-molecule clustering. To compensate the low resolution, a DMS device can be coupled to a gaschromatographic column (GC) to obtain a second separation dimension.

In the investigation of bacteria culture measurements considered here, the additional integration of pyrolysis (py) was beneficial, providing chemical information to detect microorganisms by heating in the absence of oxygen or any other reagents and is well-matched to the instrumentation of GC. The chemical method of py-GC/DMS was, therefore, used to analyse the mixtures of volatile constituents derived from bacteria samples of different strains.

The chemical information from a py-GC/DMS analysis constitutes a three-dimensional data structure of signal intensity, retention time, and compensation voltage from the

mobility scan. For the analysis of orthogonality of the two separation dimensions of py-GC and DMS, the generated data were processed similarly as introduced for the IMS breath measurements before. While axes transformation and RIP detailing were not necessary, data were smoothed and denoised by means of wavelet transform before they were subject to the method of GIM for analyte peak detection.

8.2 Orthogonality calculations for bacteria data

In a recent publication by Prasad et al. (2007), four different bacteria strains were analysed by means of py-GC/DMS. Investigating the influence of different growth temperatures, a partially limited separation between the clusters of 10 replicates per specific temperature was found for all the bacteria.

To determine if an improvement of this separation could be reached by instrumentation modifications, the orthogonality of the dimensions of retention time and compensation voltage was quantified using an approach of Liu et al. (1995) (Subsection 8.2.1) for the example of the two bacteria strains *Escherichia Coli* and *Staphylococcus Warneri* (Subsection 8.2.2). These calculations, performed at the Chemistry and Biochemistry Department at the New Mexico State University, USA, required the specification of peak positions for the entirety of relevant measurements, derived by the GIM algorithm. As the peaks in the py-GC/DMS measurements of bacteria data were situated in a very dense constellation without a general baseline, the beneficial properties of this peak detection method could be proved in this application.

8.2.1 Analytical orthogonality

Analytical orthogonality is a measure of the degree of dependency of analyte peak positions in the two dimensions of a two-dimensional separation. If a strong dependency between the two dimensions is present, such as for the coupling of two identical devices, its value is near zero; if the two dimensions are orthogonal, i.e. the analyte peaks are spread randomly over the entire measurement range, its value is near one. The investigation of analytical orthogonality is of interest for the evaluation of the performance of two-dimensional separations, as dependency between the two dimensions delimits the actual peak capacity of a device.

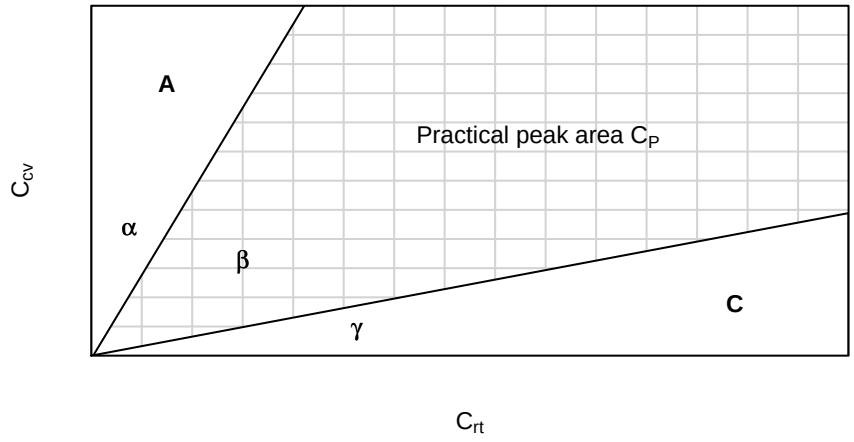

Figure 8.2: Practical peak area of a nonorthogonal two-dimensional retention space: the peak spreading angle is denominated as β, while the two areas that do not contain peaks are indicated by **A** and **C** with corresponding angles α and γ.

While the peak capacity is defined as the ratio of the measurement range and the average peak width for one-dimensional separations, it is generally considered to be the product of the peak capacities of the two separate dimensions in two-dimensional separations, such as C_{rt} and C_{cv} for retention time and compensation voltage of a py-GC/DMS device. This theoretical peak capacity, C_T, however, often overestimates the true peak capacity of coupled devices, as correlations of the two dimensions can restrict the measurement range practically containing peaks (Fig. 8.2).

The practical peak capacity, C_P, can be derived as the difference of the theoretical peak capacity, C_T, and a constant characterising the amount of correlation between the two separation dimensions (Liu et al., 1995). The calculation of this constant is based on the idea of a peak spreading angle, β, (Fig. 8.2) and requires the specification of peak capacities in the two dimensions as well as a list of peak positions of sample measurements that span the whole range of analyte peaks that are relevant for a considered application.

The peak spreading angle, β, can be determined as

$$\beta = cos^{-1} C_{corr},$$

where C_{corr} gives the sample correlation between the two peak position vectors of the separation centered and scaled with dimensionwise mean and standard deviation respectively. Quantifying the areas of the measurement space **A** and **C** where no peaks appear as

$$\mathbf{A} = 0.5 C_{cv}^2 \tan(\alpha) \quad \text{and} \quad \mathbf{C} = 0.5 C_{rt}^2 \tan(\gamma)$$

with α and γ are defined as

$$\alpha = \tan^{-1}(C_{cv}/C_{rt})(1 - 2\beta/\pi) \quad \text{and} \quad \gamma = \pi/2 - \alpha - \beta,$$

the practical peak capacity can be calculated as

$$C_P = C_T - (\mathbf{A} + \mathbf{C}).$$

The discrepancy between practical and theoretical peak capacity, given by $\mathbf{A} + \mathbf{C}$, allows conclusions to be drawn regarding the benefit of coupling different separation techniques.

8.2.2 Analysis of bacteria data

The concept of analytical orthogonality was used to investigate whether class separation of measurements on bacteria cultures grown at different temperatures (Prasad et al., 2007) is limited by the dependency between the two dimensions, and can thus be improved by modifications of the py-GC/DMS instrumentation. For this analysis, besides the peak capacity of the py-GC and DMS dimension, the positions of peaks spanning the entire potential measurement range were required. Using the GIM algorithm for the analysis of

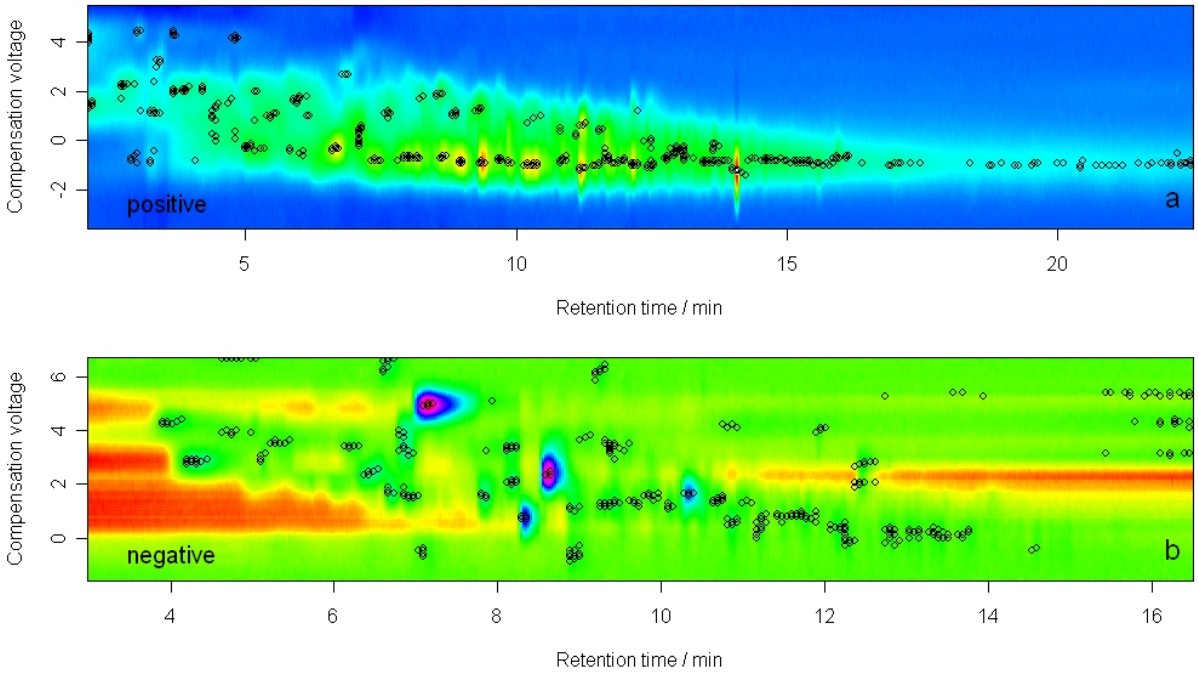

Figure 8.3: Illustration of a py-GC/DMS measurement of bacteria for the example of *S. warneri* displaying (a) positive and (b) negative mode data in heatmaps along with the detected peak positions of ten replicate measurements used as input for the orthogonality calculation.

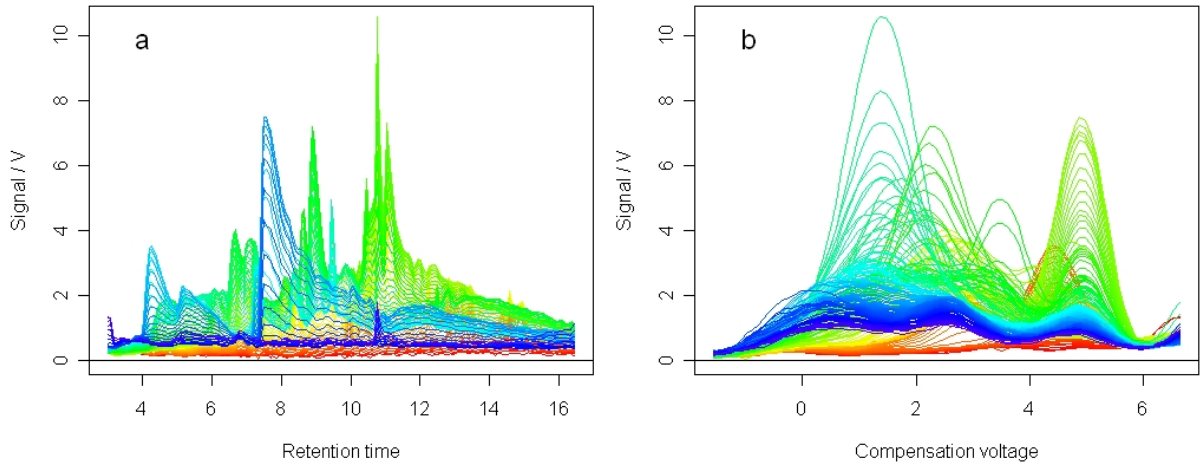

Figure 8.4: Illustration of a py-GC/DMS measurement of bacteria for the example of *S. warneri* showing a measurement taken in the negative mode in a sideview for the dimensions of (a) retention time and (b) compensation voltage to give an idea of the average peak width required as input for the orthogonality calculation.

the py-GC/DMS data of bacteria cultures containing multiple peaks in a dense assembly that do not grow from a general baseline, this application allowed the validation of the developed algorithm, especially regarding the identification of non-baseline separated peaks.

Base of the calculations were positive and negative mode data of ten py-GC/DMS measurements of *E. Coli* and *S. Warneri* grown at a temperature of 37°C. After determination of the peak lists for all replicates, these were linked for both bacteria separately, but also a joint peak list was established for both positive and negative modes (Fig. 8.3). The peak capacity in retention time was specified as 50.97 for the positive and 33.71 for the negative mode; for the compensation voltage it was 4.11 for the positive and 3.75 for the negative mode (Fig. 8.4).

The results of the orthogonality calculations for this data input (Table 8.1) indicated a negative correlation between the two dimensions of py-GC and DMS. Furthermore, the amount of dependency was found to be higher for data in the positive than for negative mode, when used for this special application for *E. coli* and *S. warneri* as well as the joint examination. This effect was evident as the peak positions for the positive mode lay on a straight line, while being more scattered over the measurement range for the negative mode data. In the positive mode, the coupling of the two separation methods yielded a gain in practical peak capacity of around 3 for the dimension of py-GC and about 40 for the dimension of DMS, when compared to the solely application of the two methods. About one quarter of the theoretical peak capacity, however, was lost in the

Table 8.1: Results of the orthogonality calculations on the base of the GIM peak lists of *E. coli*, *S. warneri*, and the joint peak list, respectively, for both data from positive and negative mode.

	E.coli	*S.warneri*	*E.coli* & *S.warneri*
Positive mode			
Correlation	-0.64	-0.66	-0.65
Theoretical peak capacity	209.30	209.30	209.30
Practical peak capacity	156.42	154.07	155.70
Loss	25.26	26.39	25.61
Negative mode			
Correlation	-0.23	-0.38	-0.31
Theoretical peak capacity	126.42	126.42	126.42
Practical peak capacity	115.37	108.34	111.79
Loss	8.74	14.30	11.58

two-dimensional separation, indicating that the discrimination of the considered bacteria sample clusters could be limited by a certain amount by dependencies of the dimensions. If the orthogonality of the py-GC/DMS instrumentation could be improved, closely related samples might, therefore, be classified more reliably. For the negative mode around 90 % of the theoretical peak capacity was maintained, implying that modifications of the instrumentation can here hardly improve the separation of the bacterial samples.

These results could not have been derived without the automated peak detection via the GIM algorithm, allowing the processing of complex spectra series data despite the dense arrangement of the peaks. Furthermore, the fact that the peaks were not lying on a general baseline did not harm the obtained peak detection results, demonstrating the wide applicability of the developed peak detection method.

Concluding remarks

Chapter 9

Conclusions and outlook

The conclusions made during the course of this work cover the two main areas of the processing of spectra series (Section 9.1) as well as the subsequent analysis for three instance applications (Section 9.2). Contributions of new knowledge throughout this thesis and inferences are discussed, and additionally, ideas for potential future research generated while working on this project are proposed (Section 9.3).

9.1 Concluding comments on spectra processing

The main problems encountered throughout this work included high amounts of redundant data generated by spectrometric methods, the limited reproducibility of measurements, high levels of noise, shoulder and overlapping peaks, as well as the phenomenon of peak tailing, especially when using the IMS method. The objectives resulting from these challenges were the development of an efficient preprocessing strategy for data reduction, denoising, and an improvement in the comparability of characteristic measurement features (Subsection 9.1.1), the characterisation of measurements by peak detection and quantification (Subsection 9.1.2), as well as the preparation of the resulting data for continuative analyses (Subsection 9.1.3).

9.1.1 Preprocessing

Before the detection and quantification of spectrometric peaks could be investigated, a preprocessing strategy was necessary taking into account the complex data structure. As

only inadequate methods were available for the challenges arising in the spectra series, some existing methods were extended and several new preprocessing steps developed.

The first issue examined in the range of the preprocessing of raw IMS spectra series was the comparability of measurements (Section 5.1, page 67), achieved by transformations of the drift time, retention time and signal intensity axes, yielding a better alignment for different measurements. The beneficial development of a reproducible version of the (inverse) reduced mobility for the IMS dimension not only yielded reliable results, but also simplified daily laboratory work, as the detailed reporting of measurement conditions such as ambient pressure and temperature was no longer required. A considerable amount of variability, however, remained in the MCC dimension as the adjustment to the column temperature, modeled by a quadratic function, resulted in an alignment improvement, but could not explain some parts of the variability. The search for other affecting factors or the development of a data based alignment method could, therefore, be beneficial.

Furthermore, the influence of the characteristic RIP tailing in the breath measurements was investigated since it interfered with data visualisation and peak detection, causing peaks in different parts of the IMS dimension to grow from different heights. By fitting a lognormal detailing function with a specially created penalty term, the tailing could be suitably described (Section 5.2, page 73). Computational costs were reduced by applying the method to a representative spectrum, while the transfer of the resulting function to the entire spectra series gave satisfying overall results.

The introduced preprocessing strategy was concluded with a combined application of smoothing and denoising using the wavelet transform (Section 5.3, page 75). Coupling Daubechies wavelets for smoothing with one compression level in each dimension and denoising via hard thresholding, an efficient data reduction was accomplished, not only involving less redundancy but also less computational cost for subsequent computations, as well as an increased signal-to-noise ratio. Peaks that previously showed a strong overlap now exhibited a better resolution. Although wavelets have previously been applied in the smoothing or denoising of IMS data, the two methods have not, to date, been linked for this application.

In summary, the developed strategy for the preprocessing of spectra series improved the reproducibility of measurement results by a data alignment using different axes transformations, eliminated a large extent of noise, improved peak clarity, and reduced the amount of data by the combined application of a new detailing function with smoothing and denoising using the wavelet transform.

9.1.2 Peak detection

After the data was prepared for further processing, the actual peak detection, characterisation, and quantification could be investigated. This step could also be seen as an extended stage of data reduction, as only a matrix consisting of a few values per detected peak remained, though still sufficiently describing the relevant information. This aim was achieved by the development of three successively established peak detection algorithms with increasing sensitivity.

The basic method of MPCL is based on a single threshold distinguishing between noise and peak measurement points, which were later merged to peak regions. These regions were subsequently characterised by their centroid positions and the maximum peak height for quantification (Section 6.1, page 81), allowing an efficient data reduction to only three values per detected peak. This procedure yielded reasonably well results shown in an instance application for the perfect discrimination between control patients and lung cancer patients. Nevertheless, the results using this method showed limitations concerning the resolution of shoulder peaks and peaks that were not baseline-separated, and furthermore the characterisation of peaks did not give an idea of their size or magnitude.

The peak detection procedure was, therefore, extended by the GIM algorithm, designed in a stagewise manner to improve sensitivity (Section 6.2, page 86). At each stage of the method, a peak detection step was applied, closely related to the MPCL algorithm, which assigned values in growing intervals of the intensity axes to the peak cluster, and characterised the merged regions by ellipses. The decisive step of this algorithm was the connection of the peak lists from the different stages, allowing the detection of multiple non-baseline separated peaks, solving a common challenge in peak detection. The result of this algorithm was a list of six values per detected peak, characterising each peak by four ellipse parameters and the determined ellipse area in addition to the peak height. The method, therefore, achieved improved peak quantification by ellipse extents and area, thus enabling the visualisation of data in ellipse representations, and showing its wide applicability by the transfer of the method to the application with DMS instrumentation. Although the detection of multiple peaks is now possible, limitations can still be found in the identification of shoulder peaks without independent maxima.

In a final step, this procedure was further improved by the substitution of a number of introduced preprocessing steps by a wavelet-based MRA, resulting in details that partially resolved covered peaks, and are now the basis of the GIM algorithm (Section 6.3, page 94). Here, the LH details of level 3 to 5 obtained using the Haar wavelet showed to be

most beneficial. Combining the results of the GIM algorithm for the three levels enhanced peak detection, allowing the detection of shoulder peaks without independent maxima, in addition to the benefits of the two other introduced peak detection methods, while peaks were still characterised by ellipse parameters and maximum height. The results of this method could be used to analyse the complex comparison of lung cancer patients with different forms of tumors.

All three methods focused on a high sensitivity of peak detection rather than on specificity, as artefact peaks evolving from noise areas, were filtered out in the continuative analyses.

9.1.3 General peak areas

Before the benefit of the three consecutive peak detection methods could be proven in the application to real questions, general peak areas allowing the direct comparison of features between different measurements, and thus further statistical evaluations, had to be created based on the established peak lists (Chapter 7, page 107).

The construction of general peak areas for specific applications was initiated by a cluster analysis of the entirety of peak positions detected in all measurements of a study. The cluster procedure of k-means with starting values derived from Ward's method, yielded solutions that were most appropriate for this kind of data. The cluster solution with the cluster number, optimised according to two performance indices, was subsequently used as the basis for the adjustment of general peak areas.

For peak lists resulting from the MPCL algorithm this was achieved by rectangular areas containing all peak positions belonging to a cluster respectively, where a marginal widening in all directions around a cluster was found to be beneficial, also including main parts of peaks whose position was located at the margins of a cluster (Subsection 7.1.1, page 108). The results of the GIM algorithm, either applied to the data preprocessed in the original manner or on top of the determination of relevant MRA details, were on the other side combined to ellipsoid general peak areas giving a more precise characterisation of important application-related measurement parts (Subsection 7.2.1, page 116).

The proposed general peak areas could subsequently be used in defining general peak variables, which were computed as the mean intensity value of all points located in a general peak area for each measurement of interest respectively. These newly established peak variables constituted a further improved, application-related quantification of peaks and gave the base for statistical evaluations in continuative analyses.

9.2 Concluding comments on different applications

To show the practical benefits of the developed processing steps, three separate studies were examined, each based on the results of one of the three developed peak detection methods. While the MPCL (Subsection 9.2.2) and the wavelet-based GIM algorithm (Subsection 9.2.3) were applied to MCC/IMS breath measurements, constituting the main topic of this work, the GIM method used with an adjusted version of the proposed preprocessing strategy was transferred for use with py-GC/DMS spectra series (Subsection 9.2.1).

9.2.1 Orthogonality of bacteria measurements

A wide applicability of the GIM algorithm by the potential transfer to other spectrometric methods was proven investigating the analytical orthogonality of py-GC/DMS data of different bacteria cultures (Chapter 8, page 125). After a preprocessing strategy involving smoothing and denoising by wavelets, the detection of peaks situated in a very dense constellation without a general baseline, showed the beneficial properties of the GIM method for data resulting from both positive and negative mode.

The detected peak positions gave the input for the orthogonality calculations, analysing if dependencies of the two separation dimensions limited the practical peak capacity. For the negative data 90% of the theoretical peak capacity was maintained, while for the positive data a correlation of -0.65 between the peak positions in the two measurement dimensions caused the loss of one quarter of the theoretical peak capacity. Although the two-dimensional separation still showed a much better peak capacity than observed for the solely application of py-GC or DMS respectively, this result implied that an improvement of the orthogonality of the py-GC/DMS instrumentation could yield better classification results for closely related samples.

9.2.2 Discrimination between lung cancer and control group

By analysing two sets of measurements on lung cancer patients and healthy control persons using the MPCL method, 106 general peak areas with corresponding peak variables were generated based on a cluster analysis of the entirety of detected peak positions of all samples in the study. Based on 25 differentially expressed variables screened in a multiple t-test procedure after the exclusion of variables with a high amount of missing values

in a descriptive analysis, an LDA could be applied yielding a perfect separation of the groups with a leave-one-out error rate estimation of zero after one more variable was excluded in a stepwise selection. This concurrently meant a further data reduction from 106 general peak variables to a single discriminant value, still allowing the two groups of measurements in this application to be distinguished from one another (Subsection 7.1.2, page 111).

In parallel studies on the emission of bacteria, peaks occurring within either one of two specific general peak areas with the highest influence on the assignment of patients to the cancer group, corresponded with areas containing analyte peaks detected in measurements of *E. coli* cultures. This led to the assumption that part of the discriminatory information could be related to a bacterial infection which is more likely to appear in lung cancer patients at the time of a hospitalisation than for healthy control subjects. For an ensured statement on the influence of this factor, further studies concerning this relationship between bacterial emission and the outcome of exhaled air measurements by MCC/IMS would be worthwhile. In any case, the introduced processing strategy around the MPCL method enabled a successful characterisation of differences between two groups of samples and showed its potential as an efficient instrument for the analysis of spectrometric data.

9.2.3 Comparison of different forms of tumor

For the more complex comparison of lung cancer patients with different forms of tumors, the peak detection method of GIM based on wavelet-derived MRA details was applied (Subsection 7.2.2, page 118). By adjusting ellipsoid general peak areas to the optimal cluster solution for the entirety of the resulting GIM peak positions, 148 corresponding mean intensity peak variables were defined, giving the basis for further analysis. After the exclusion of variables and patients with high amounts of missing values, as well as variables based on noise artefacts, a pairwise comparison of circular focuses with endobronchial tumors, other lung carcinoma, and the entirety of all other tumors together, as well as of the groups of patients with endobronchial tumors with other lung carcinoma was investigated. For every comparison, differentially expressed variables were screened in a leave-one-out procedure identifying those variables that showed a significant difference across the entire patient range.

Analysing these comparisons in an LDA with a stepwise selection based on the estimated leave-one-out error rate, and the values of the standardised discriminant coefficients, error rate estimations between 5 and 20% were determined for the different problems. The

best observed prediction rate was found in the comparison between patients with circular focuses and those with other carcinoma exclusive endobronchial tumors. Here, all patients were assigned to the correct group on the basis of the established decision rule. The group of patients with endobronchial tumors, on the other hand, proved to be the most difficult to separate from other groups.

The peak regions corresponding with the variables relevant for the specified decision rules could be assigned to three main groups, located in the pre-RIP, the RIP, and later spectra parts of the IMS dimension. While three decision rules were influenced by one of these peak area groups, the fourth comparison was based on a mixture of the pre-RIP and later spectra parts peak regions. The peak variables containing the main part of discriminatory information for the comparison of patients with circular focuses and all other tumors, exclusive of and inclusive of endobronchial tumors respectively, were found to be those that corresponded with general peak areas not lying on a straight line with the majority of peak areas in the pre-RIP and the RIP itself. This could explain why patients with different forms of tumors were separated rather efficiently, while patients with endobronchial tumors, whose assignment was mainly based on the peaks in the RIP which was a priori not expected to contain discriminatory information, were hard to assign to the correct group. A relationship to the relevant peak areas for the comparison of lung cancer patients with control persons was not apparent.

This application, giving further interesting insights into the differences of human breath composition for lung cancer patients, showed that the developed processing strategy around the wavelet MRA based GIM algorithm constitutes a valuable contribution for the analysis of spectrometric data, avoiding time-consuming, subjective, and likely insufficient manual characterisation of measurements.

9.3 Future work

The investigations on the characterisation of the metabolite composition in exhaled air of different groups of patients by means of MCC/IMS can be both broadened and specialised in different directions.

One potential question of interest could examine the influence of bacterial emission on the outcome of the breath measurements, verifying the relevance of the determined discriminant function between lung cancer patients and the control group (Subsection 7.1.2). For this, one possible approach could characterise the measurements from the emissions

of the most common lung bacteria with a subsequent matching with patients known to be affected by these bacteria. Additionally, isolated lung cancer cells could be analysed to identify their emission products and check for coincidence with parts of the profiles detected in the exhaled air of lung cancer patients.

Another important field is the investigation of other diseases besides lung cancer, enabling their characterisation and comparison in a similar way as introduced in this work for different types of lung tumor. Because of the high number of confounding factors such as smoking, perfume, a recent visit to the swimming pool, freshly brushed teeth and the consumption of food or drinks in the time period before measurement, an analysis like that would require a sufficiently high number of samples. Optimally, all these factors would be surveyed in further studies to allow for the investigation of some main influencing factors which might be characterised and incorporated into further analyses.

An additional application of the breath monitoring via MCC/IMS could also be the control of the effect of drugs on metabolism (Baumbach et al., 2005). By taking several measurements over a specific course of time from patients after the start of treatment with the drug of interest, such measurements could be handled in a similar way as in this work, respecting the special structure of the related time series data.

Furthermore, as already shown using the py-GC/DMS, the transfer of the introduced methods to other two-dimensional separations such as the popular method of GC-MS is possible, probably involving only minor adjustments of the developed algorithms. In doing so, a broader field of potential applications would be addressed by the methods that originated in the range of this work, increasing the achieved benefit for spectrometric analytics.

Methodologically, there are two main aspects that should be respected in future research on this topic, both related to specifics of the MCC/IMS instrumentation (Fig. 9.1).

Firstly, a problematic feature of MCC/IMS measurements is the ion absorption by dominating peaks, constricting peaks with a low proton affinity (Fig. 9.1 a). Consequently, the quantification of the effected peaks is insufficient and can even lead to the detection of multiple peaks for a single analyte dependent on the existence of specific other peaks and is therefore also influencing the peak characterisation of measurements. It would, therefore, be worthwhile to create a strategy dealing with this characteristic behaviour of the MCC/IMS instrumentation to further improve peak characterisation and quantification for these special cases.

A second problem is the partially unsatisfying alignment of the data in the retention time direction, even showing some systematic effects, which resulted in alignment-artefact peak areas with a position in inverse reduced mobility that corresponded with that of the most common general peak areas, but were shifted backwards in the MCC dimension with increasing shift for latter original retention time positions of the area (Fig. 9.1 b). As for some measurements the entirety of peak positions was shifted as indicated for these regions, this likely had an undesirable effect on the values of the determined general peak variables influencing the discriminant value of a measurement and, therefore, the classification result. As different measurements do not contain the same combination of peaks, this problem can not be solved in a straightforward way, but is currently addressed in a diploma thesis at the ISAS - Institute for Analytical Sciences and the Faculty of Statistics at the Technical University Dortmund, in a data based way. Another solution might be an alignment strategy based on the employment of an internal standard optimally consisting of around three analytes causing peaks in diverse areas of the measurement space, added in a defined amount to each measured sample. This proceeding would also allow for an efficient standardisation of peak heights, which might further improve the comparability of measurements in addition to the effect of an optimised alignment of measurements, and thus moreover precise the constitution of general peak areas.

Respecting these additional aspects, a universal method for the analysis of spectrometric data in two-dimensional separations could be achieved, further strengthening the derived

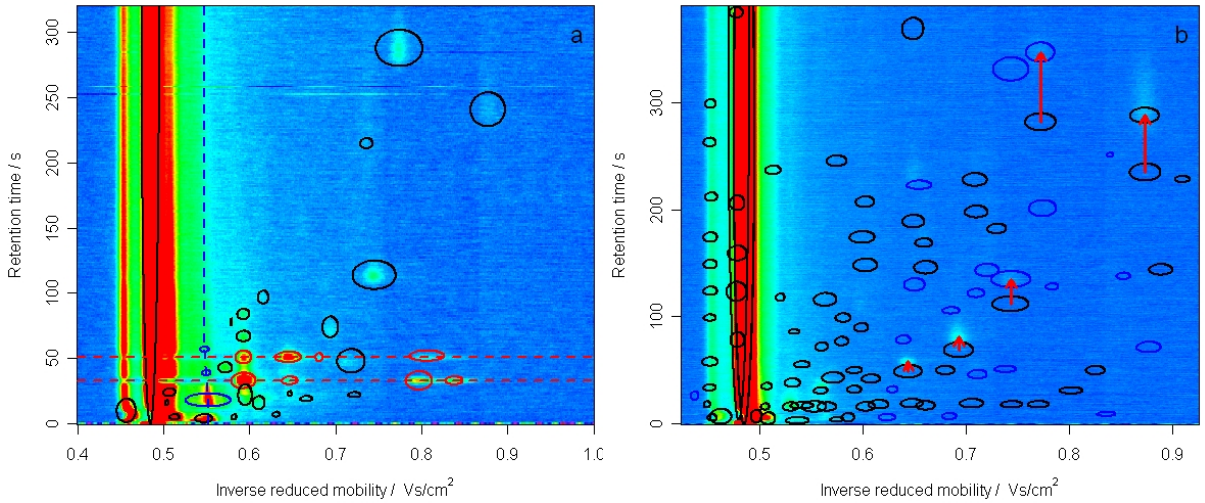

Figure 9.1: Heatmaps illustrating remaining limitations considering (a) analyte quantification by reason of ion absorption by dominating peaks (red horizontal lines) yielding the constriction of peaks with low proton affinity (blue vertical line), and (b) peak characterisation, because of systematic alignment distortions in the direction of retention time resulting in undesired side clusters at the end of three of the red arrows.

benefits offered by the methods developed in this work. On this basis, future prospects for the special application of human breath measurements by MCC/IMS might focus on the comparison of diverse lung diseases, tracing the vision of a new broad screening method for pneumological aspects.

Acknowledgement

I would like to thank Prof. Dr. Wolfgang Urfer for the supervision of this work and for giving me my first insight into statistics in life sciences, encouraging me to get deeper into and finally stay in this field. I also appreciate the willingness of Prof. Dr. Claus Weihs to survey this work and his competent, helpful, and friendly support regarding this project and my career planning. Furthermore, I acknowledge Dr. Jörg Ingo Baumbach for proposing of this PhD thesis, his motivation advise regarding my research, but also for making me care about my future, and always being such a reliable, flexible, and understanding boss.

Importantly, I also want to thank Dr. Wolfgang Vautz, Luzia Seifert, Susanne Krois, and Dr. Vera Ruzsanyi for the vivid exchange of ideas and needs in the generation and analysis of measurement data, and Bertram Bödeker for valuable, diverting discussions. Besides there are many more colleagues and friends at the ISAS – Institute for Analytical Sciences such as Dr. Jürgen Nolte, Rita Fobbe, Rolf Bandur, and of course Kaoru Tachikawa, Sven Tombrink, Marco Becker, and Hendrik Kortmann whose support in the daily madness of our institutes life I will remember gratefully. Also, I want to thank Prof. Dr. Gary Eiceman, Dr. Hartwig Schmitt, Dr. Satendra Prasad, and Jaime Rodriguez from the New Mexico State University for giving me an insight to a different world with new instrumentations and scientific questions as well as for the provision of some of their data for this thesis. In addition, I am thankful to Dr. Michael Westhoff, Dr. Patrick Litterst, and Barbara Oberdrifter from the lung hospital in Hemer for the patience in collecting data and the interdisciplinary discussions originating questions of relevance.

My special thanks goes to Darius Wilson for motivating me through the last couple of months that were crucial for the completion of this thesis, supporting me not only by critically going through each page and giving valuable advise, but also by making me calm down and relax when necessary. Above all, I thank my family and all my friends that gave me an excellent background, supporting me wherever they could, and thus gave me the perfect foundation for the creation of this work.

Appendix A

Tables

This appendix contains tables whose values are not necessary to follow the general proceeding in this work, but were included to allow a deeper insight into the data analysis of the specific applications if desired. All the tables refer to Chapter 7, were the introduced methods for spectra series processing were used and enhanced to the analysis of entire studies. For the two instance applications out of the range of human breath monitoring considering the separation of lung cancer patients and healthy control persons (Section 7.1, Table A.1, A.2, and A.3) as well as the comparison of the profiles of patients with different kinds of bronchial carcinomas (Section 7.2, Table A.4, A.5), and A.7, the analysis was broken down to the most important results here.

Table A.1: Limits of the determined rectangular general peak areas for the comparison of lung cancer patients and a control group (Section 7.1), the number of data points lying within these limits and values of the corresponding peak variables for an instance measurement, as well as p-values of the t-tests: Lines in italic font correspond with variables that were excluded in the descriptive analysis and, therefore, do not contain a p-value; bold lines belong to variables that were differentially expressed in the t-test. In line 107 no area limits are given, as this variable corresponds with all those measurement parts that do not belong to any of the other peak areas.

	Limits				Instance measurement		p-values
	L_{K_0}	U_{K_0}	$L_{t_r^{30}}$	$U_{t_r^{30}}$	observed value	data points	of the t-tests
1	0.52	0.55	0	25.04	0.1030	713	0.0017
2	0.59	0.61	2	38	0.0063	990	0.2575
3	0.58	0.61	28	66	0.0035	1517	0.7461
4	0.55	0.58	1	59	0.0269	1728	0.0780
5	0.59	0.61	56.14	106	0.0153	1392	0.2281
6	0.70	0.73	5.95	96	0.0011	2871	0.8838

	Limits				Instance measurement		p-values
	L_{K_0}	U_{K_0}	$L_{t_r^{30}}$	$U_{t_r^{30}}$	observed value	data points	of the t-tests
7	0.54	0.56	12.42	53	0.0400	1240	0.2050
8	0.52	0.55	69.5	109	0.0409	1292	0.0464
9	**0.75**	**0.78**	**93.5**	**117.39**	**-0.0019**	**782**	$\mathbf{3.8 \cdot 10^{-21}}$
10	0.58	0.61	97	139	0.0156	1353	0.0538
11	0.54	0.56	97	145	0.0155	1692	0.4021
12	0.54	0.56	138	192	0.0085	1716	0.1250
13	0.59	0.61	130	169	0.0203	1140	0.4201
14	0.59	0.61	160	200	0.0202	975	0.0090
15	0.51	0.54	159	204	0.0291	1408	0.8873
16	0.58	0.61	191	233	0.0149	1230	0.4034
17	**0.61**	**0.63**	**53**	**118**	**-0.0073**	**1827**	$\mathbf{7.6 \cdot 10^{-8}}$
18	0.58	0.60	223	271	-0.0097	1363	0.6508
19	**0.89**	**0.91**	**203.5**	**262**	**0.0007**	**1568**	$\mathbf{7.5 \cdot 10^{-6}}$
20	**0.61**	**0.64**	**248**	**311**	**-0.0083**	**2379**	$\mathbf{6.5 \cdot 10^{-20}}$
21	**0.77**	**0.81**	**235**	**302.3**	$\mathbf{3.7 \cdot 10^{-5}}$	**2925**	$\mathbf{5.6 \cdot 10^{-13}}$
22	**0.59**	**0.60**	**265**	**314.5**	**0.0072**	**912**	$\mathbf{6.5 \cdot 10^{-11}}$
23	**0.61**	**0.66**	**307**	**371**	**-0.0082**	**3969**	$\mathbf{1.1 \cdot 10^{-7}}$
24	0.54	0.56	303.58	377	0.0046	1988	0.0600
25	**0.59**	**0.61**	**305.5**	**356**	**0.0049**	**1029**	$\mathbf{2.4 \cdot 10^{-13}}$
26	**0.59**	**0.61**	**350**	**407**	**0.0030**	**1176**	$\mathbf{7.8 \cdot 10^{-15}}$
27	0.57	0.61	400.33	464	0.0232	3233	0.7429
28	0.52	0.54	404	448	0.0241	1462	0.0013
29	**0.61**	**0.64**	**391**	**469**	**0.0093**	**2400**	$\mathbf{-1.1 \cdot 10^{-16}}$
30	**0.61**	**0.64**	**479**	**581**	**-0.0083**	**1462**	$\mathbf{9.5 \cdot 10^{-16}}$
31	0.52	0.54	469	507	0.0253	1110	0.0003
32	0.54	0.57	0	23	0.0249	777	0.0337
33	0.57	0.59	0	44	0.0290	1353	0.0198
34	**0.61**	**0.64**	**0**	**27**	**0.0032**	**1050**	$\mathbf{-2.8 \cdot 10^{-17}}$
35	0.52	0.54	13	55.5	0.1120	1344	0.0001
36	0.65	0.68	0	59	-0.0052	2255	0.4991
37	0.52	0.54	39	81.5	0.0785	1230	0.0004
38	0.54	0.57	62	108	0.0207	1620	0.0581
39	0.52	0.55	128	169.5	0.0277	1360	0.4668
40	**0.61**	**0.64**	**162**	**232.5**	**-0.0107**	**2652**	$\mathbf{5.7 \cdot 10^{-19}}$

	Limits				Instance measurement		p-values
	L_{K_0}	U_{K_0}	$L_{t_r^{30}}$	$U_{t_r^{30}}$	observed value	data points	of the t-tests
41	0.52	0.55	254	295	0.0225	1360	0.1109
42	0.52	0.55	317	358.5	0.0244	1200	0.0273
43	0.52	0.54	347	386	0.0261	1140	0.0046
44	0.52	0.54	436.5	479.17	0.0224	1353	0.0010
45	0.52	0.55	496	533.25	0.0242	891	0.0019
46	0.54	0.56	43.5	85.18	0.0396	1240	0.2379
47	0.52	0.54	98	138.5	0.0362	1209	0.4898
48	**0.61**	**0.64**	**113**	**169.5**	**-0.0156**	**1890**	$\mathbf{2.9 \cdot 10^{-18}}$
49	0.52	0.55	284	326	0.0251	1312	0.0358
50	0.52	0.55	375.5	414	0.0264	1330	0.0156
51	0.52	0.55	190.31	235	0.0243	1462	0.8669
52	*0.52*	*0.54*	*522.5*	*560*	*0.0265*	*31*	–
53	*0.52*	*0.55*	*549*	*602*	–	–	–
54	0.52	0.55	224	265	0.0233	1440	0.2814
55	0.57	0.61	473	552.25	0.0093	2107	0.3585
56	**0.74**	**0.76**	**21**	**114.5**	**0.0015**	**3150**	$\mathbf{8.1 \cdot 10^{-10}}$
57	0.54	0.57	202	284	0.0024	2607	0.0416
58	0.68	0.69	1.15	17	-0.0009	247	0.8356
59	*0.51*	*0.53*	*132*	*186.71*	*0.0321*	*742*	–
60	0.56	0.59	214	270.5	0.0083	2160	$5.6 \cdot 10^{-5}$
61	0.54	0.56	377	457.5	0.0041	2184	0.1997
62	**0.67**	**0.70**	**43.5**	**140**	**0.0017**	**4185**	$\mathbf{3.0 \cdot 10^{-7}}$
63	0.62	0.65	5.5	62	-0.0028	2160	0.2156
64	**0.57**	**0.59**	**48**	**112.77**	**0.0130**	**1736**	$\mathbf{1.2 \cdot 10^{-7}}$
65	**0.71**	**0.74**	**80.5**	**129.5**	**0.0077**	**1786**	$\mathbf{9.3 \cdot 10^{-9}}$
66	**0.57**	**0.59**	**115**	**174.5**	**0.0096**	**1254**	$\mathbf{1.7 \cdot 10^{-12}}$
67	0.64	0.67	100	170	0.0104	1943	0.0614
68	0.87	0.90	133	151	0.0088	684	0.3515
69	**0.56**	**0.59**	**170.5**	**224**	**0.0066**	**1976**	$\mathbf{6.1 \cdot 10^{-10}}$
70	**0.85**	**0.88**	**207.5**	**258.65**	**0.0028**	**1862**	$\mathbf{8.1 \cdot 10^{-9}}$
71	*0.51*	*0.53*	*232.33*	*282*	*0.0311*	*611*	–
72	**0.74**	**0.77**	**262.5**	**316**	**0.0049**	**2080**	$\mathbf{2.3 \cdot 10^{-8}}$
73	*0.51*	*0.53*	*385*	*439*	*0.0337*	*624*	–
74	*0.51*	*0.53*	*481.5*	*530*	*0.0332*	*440*	–

	Limits				Instance measurement		p-values
	L_{K_0}	U_{K_0}	$L_{t_r^{30}}$	$U_{t_r^{30}}$	observed value	data points	of the t-tests
75	*0.51*	*0.53*	*88*	*140*	*0.0323*	*612*	–
76	*0.51*	*0.53*	*0*	*52.12*	*0.0927*	*588*	–
77	*0.49*	*0.52*	*123*	*172*	–	–	–
78	*0.51*	*0.53*	*183*	*233*	*0.0322*	*392*	–
79	0.75	0.77	196	258	-0.0045	1586	0.0330
80	*0.50*	*0.53*	*435*	*487*	*0.0315*	*450*	–
81	*0.51*	*0.53*	*275*	*333*	*0.0341*	*616*	–
82	*0.50*	*0.53*	*328.33*	*387*	*0.0360*	*684*	–
83	**0.57**	**0.59**	**343.5**	**410**	**0.0054**	**1664**	$\mathbf{5.7 \cdot 10^{-9}}$
84	*0.49*	*0.52*	*162.5*	*220*	–	–	–
85	0.54	0.56	513	570.29	0.0071	297	0.8692
86	**0.56**	**0.58**	**270**	**341**	**0.0061**	**1863**	$\mathbf{6.1 \cdot 10^{-10}}$
87	*0.49*	*0.52*	*78.5*	*132.15*	–	–	–
88	*0.49*	*0.52*	*262*	*318.61*	–	–	–
89	*0.49*	*0.52*	*308.5*	*364*	–	–	–
90	*0.49*	*0.52*	*369*	*420.5*	–	–	–
91	*0.49*	*0.52*	*411*	*462*	–	–	–
92	*0.49*	*0.51*	*459*	*523*	–	–	–
93	*0.50*	*0.52*	*514*	*554*	*0.0363*	*50*	–
94	*0.50*	*0.52*	*553*	*636.5*	–	–	–
95	*0.49*	*0.52*	*0*	*29*	–	–	–
96	*0.49*	*0.52*	*209.5*	*267.5*	–	–	–
97	*0.49*	*0.52*	*31.09*	*89.5*	*0.0221*	*110*	–
98	0.69	0.71	216	294	-0.0033	1292	0.0564
99	0.54	0.56	464	514.5	0.0050	1519	0.7429
100	0.83	0.87	0	114.31	-0.0041	5832	0.2998
101	0.84	0.87	237	302	-0.0046	2457	0.0132
102	0.84	0.87	121.2	226	-0.0039	3700	0.0597
103	**0.70**	**0.72**	**128.5**	**192.52**	**0.0020**	**1708**	$\mathbf{6.3 \cdot 10^{-7}}$
104	0.84	0.87	335	388.29	-0.0028	1976	0.1226
105	0.84	0.86	396.5	461.83	-0.0027	2232	0.1646
106	0.84	0.87	474.67	564	-0.0025	1739	0.2093
107	–	–	–	–	-0.0007	527036	0.8488

Table A.2: (Standardised) discriminant coefficients of the discriminant function for the separation of lung cancer patients and control persons (Subsection 7.1.2) based on 25 and 24 variables respectively, as well as the leave-one-out error rate excluding the 25 variables solely: the variable names are oriented on the nominations in table A.1.

	25 variables			24 variables	
Variable name	Discriminant coefficients	Standardised coefficients	Error rate estimation	Discriminant coefficients	Standardised coefficients
V_9	51.9	0.145	0.011	44.2	0.123
V_{17}	-31.4	-0.174	0.011	-22.3	-0.123
V_{19}	-38.1	-0.238	0.011	-36.9	-0.230
V_{20}	-51.8	-0.099	0.011	-68.5	-0.132
V_{21}	52.9	0.082	0.011	54.2	0.084
V_{22}	-18.0	-0.053	0.011	-27.2	-0.081
V_{23}	-514.5	-0.595	0.011	-513.3	-0.593
V_{25}	418.6	1.006	0.011	427.1	1.026
V_{26}	276.2	0.667	0.011	287.5	0.694
V_{29}	228.7	0.516	0.011	246.4	0.556
V_{30}	-155.4	-0.343	0.011	-118.7	-0.262
V_{34}	89.2	0.283	0.011	75.6	0.240
V_{40}	-214.4	-0.507	0.011	-262.8	-0.621
V_{48}	258.6	0.894	0.011	250.2	0.865
V_{56}	178.5	0.235	0.011	144.9	0.191
V_{62}	-209.1	-0.368	0	-208.5	-0.367
V_{64}	-42.9	-0.280	0	-16.9	-0.110
V_{65}	-403.1	-1.032	0.011	-393.1	-1.007
V_{66}	-404.6	-1.504	0.011	-429.0	-1.595
V_{69}	65.5	0.223	0.011	56.9	0.194
V_{70}	-107.0	-0.189	0.011	-84.7	-0.149
V_{72}	-324.4	-0.662	0.011	-296.0	-0.604
V_{83}	196.5	0.641	0.011	187.3	0.611
V_{86}	-205.6	-0.862	0.011	-203.4	-0.853
V_{103}	90.3	0.225	0	*excluded in stepwise selection*	

Table A.3: Discriminant values for the separation of healthy control persons (ko_i, $i = 1, \ldots, 54$) and lung cancer patients (bc_j, $j = 1, \ldots, 35$) of the considered sample set (Subsection 7.1.2) for the discriminant function based on 25 and 24 variables, and using the entire sample set (overall) and the leave-one-out method (loo), respectively.

	25 variables		24 variables			25 variables		24 variables	
Person	overall	loo	overall	loo	Person	overall	loo	overall	loo
ko_1	7.45	7.32	7.44	7.34	bc_1	-8.96	-11.01	-8.75	-10.22
ko_2	8.27	8.58	8.29	8.61	bc_2	-7.16	-6.75	-7.25	-6.96
ko_3	7.93	8.03	7.90	8.01	bc_3	-6.42	-5.86	-6.28	-5.73
ko_4	7.21	7.12	7.29	7.22	bc_4	-7.93	-8.18	-7.75	-7.84
ko_5	7.40	7.33	7.33	7.24	bc_5	-8.30	-8.77	-8.19	-8.59
ko_6	7.20	7.10	7.17	7.08	bc_6	-7.02	-4.93	-6.96	-4.78
ko_7	7.70	7.70	7.65	7.64	bc_7	-6.73	-4.73	-6.70	-4.71
ko_8	7.67	7.67	7.65	7.65	bc_8	-9.36	-10.57	-9.50	-10.64
ko_9	7.62	7.60	7.70	7.70	bc_9	-7.62	-7.61	-7.64	-7.65
ko_{10}	7.78	7.78	7.76	7.77	bc_{10}	-6.19	-5.13	-5.98	-5.02
ko_{11}	7.69	7.68	7.63	7.63	bc_{11}	-9.36	-10.88	-9.33	-10.85
ko_{12}	6.85	6.76	6.76	6.67	bc_{12}	-6.37	-3.43	-6.40	-3.64
ko_{13}	7.82	7.83	7.81	7.83	bc_{13}	-8.70	-9.15	-8.62	-9.03
ko_{14}	7.66	7.64	7.60	7.57	bc_{14}	-6.21	-5.46	-6.15	-5.38
ko_{15}	8.44	8.55	8.38	8.49	bc_{15}	-7.00	-6.77	-7.03	-6.82
ko_{16}	7.11	7.04	7.08	7.01	bc_{16}	-6.28	-6.03	-6.30	-6.07
ko_{17}	8.38	8.88	8.34	8.83	bc_{17}	-7.31	-7.18	-7.43	-7.36
ko_{18}	7.21	7.10	7.19	7.08	bc_{18}	-6.72	-5.95	-6.66	-5.86
ko_{19}	7.95	7.99	7.99	8.04	bc_{19}	-6.17	-5.65	-6.23	-5.76
ko_{20}	6.31	6.11	6.25	6.05	bc_{20}	-6.31	-6.00	-6.37	-6.08
ko_{21}	6.93	6.81	6.97	6.86	bc_{21}	-7.11	2.61	-7.18	-1.61
ko_{22}	6.86	6.74	6.86	6.74	bc_{22}	-8.59	-18.27	-8.49	-16.86
ko_{23}	9.30	9.88	9.23	9.77	bc_{23}	-8.30	-8.81	-8.23	-8.69
ko_{24}	7.63	7.61	7.65	7.63	bc_{24}	-8.43	-12.42	-8.54	-12.45
ko_{25}	8.12	8.19	8.08	8.16	bc_{25}	-9.19	-10.93	-9.18	-10.94
ko_{26}	7.91	8.02	7.85	7.95	bc_{26}	-4.58	-3.02	-4.66	-3.27
ko_{27}	7.98	8.01	7.94	7.97	bc_{27}	-8.65	-9.33	-8.76	-9.45
ko_{28}	7.70	7.68	7.71	7.69	bc_{28}	-8.29	-8.56	-8.42	-8.71
ko_{29}	8.11	8.20	8.11	8.22	bc_{29}	-7.30	-6.55	-7.54	-7.49
ko_{30}	7.74	7.84	7.67	7.73	bc_{30}	-8.34	-8.97	-8.04	-8.26

	25 variables		24 variables			25 variables		24 variables	
Person	overall	loo	overall	loo	Person	overall	loo	overall	loo
ko_{31}	8.21	9.33	8.16	9.25	bc_{31}	-9.62	-13.62	-9.44	-12.63
ko_{32}	5.79	5.53	5.77	5.51	bc_{32}	-6.97	-6.70	-6.97	-6.70
ko_{33}	6.12	5.86	6.02	5.77	bc_{33}	-4.39	-3.04	-4.20	-3.09
ko_{34}	6.74	6.63	6.78	6.68	bc_{34}	-7.53	-7.46	-7.40	-7.25
ko_{35}	7.82	7.84	7.81	7.84	bc_{35}	-5.83	-5.46	-5.64	-5.36
ko_{36}	7.47	7.42	7.49	7.45					
ko_{37}	7.51	7.47	7.55	7.51					
ko_{38}	7.63	7.62	7.50	7.45					
ko_{39}	5.10	4.89	5.16	4.98					
ko_{40}	6.48	6.25	6.41	6.17					
ko_{41}	6.80	6.67	6.77	6.64					
ko_{42}	6.90	6.60	6.87	6.57					
ko_{43}	7.62	7.59	7.61	7.58					
ko_{44}	6.79	6.63	6.68	6.52					
ko_{45}	7.21	7.15	7.19	7.13					
ko_{46}	7.17	7.10	7.11	7.04					
ko_{47}	7.61	7.59	7.59	7.57					
ko_{48}	6.64	6.48	6.58	6.42					
ko_{49}	7.56	7.52	7.42	7.37					
ko_{50}	6.86	6.75	6.75	6.65					
ko_{51}	8.77	9.50	8.69	9.37					
ko_{52}	8.06	8.20	7.89	7.96					
ko_{53}	6.18	5.97	6.18	5.98					
ko_{54}	7.01	6.89	7.10	7.01					

Table A.4: Parameters of the determined ellipsoid general peak areas for the general comparison of breath measurements (Section 7.2) and p-values of the t-tests comparing patients with different kinds of lung cancer (Subsection 7.2.2): the pairwise comparisons are indicated by the headings using the nominations CF for circular focuses, EB for endobronchial carcinomas, and NN for other kinds of carcinoma, as well as AO comprising all tumors exclusive circular focuses. Lines of peak areas that do not contain p-values for the t-test correspond with peak variables that were excluded in the descriptive analysis.

General	**Ellipse parameters**				**p-values**			
peak areas	x_0^g	y_0^g	a^g	b^g	CF - EB	CF - NN	EB - NN	CF - AO
1	0.438	26.00	0.004	4.00	0.321	0.025	0.076	0.101
2	0.449	18.00	0.003	3.00	0.421	0.032	0.057	0.138
3	0.452	33.75	0.006	3.00	0.162	0.007	0.044	0.032
4	0.452	58.00	0.007	4.00	0.129	0.006	0.054	0.024
5	0.463	6.77	0.010	7.00	0.201	0.009	0.047	0.034
6	0.479	159.25	0.009	7.00	0.393	0.570	0.008	0.644
7	0.485	261.00	0.010	262.00	0.508	0.579	0.027	0.760
8	0.509	5.00	0.007	5.53	0.212	0.725	0.168	0.453
9	0.537	3.00	0.011	3.00	0.015	0.141	0.744	0.008
10	0.530	16.00	0.007	4.00	0.506	0.426	0.644	0.352
11	0.541	55.25	0.007	3.00	0.529	0.585	0.868	0.436
12	0.553	16.00	0.012	5.00	0.491	0.334	0.664	0.413
13	0.560	31.00	0.007	3.00	0.476	0.442	0.879	0.454
14	0.560	72.00	0.006	4.00	0.197	0.220	0.909	0.193
15	0.571	43.00	0.011	5.00	0.797	0.808	0.636	0.900
16	0.599	90.00	0.007	4.00	0.178	0.201	0.901	0.161
17	0.592	32.00	0.007	3.52	0.270	0.301	0.619	0.276
18	0.592	50.00	0.007	3.00	0.189	0.210	0.816	0.192
19	0.603	16.00	0.009	4.00	0.825	0.846	0.716	0.899
20	0.616	18.00	0.010	4.00	0.686	0.872	0.841	0.730
21	0.629	32.00	0.005	3.00	0.698	0.849	0.880	0.724
22	0.644	49.00	0.013	6.00	0.206	0.220	0.979	0.200
23	0.661	17.00	0.009	4.00	0.620	0.352	0.574	0.390
24	0.680	50.00	0.009	4.00	0.171	0.735	0.432	0.260
25	0.693	69.00	0.014	7.00	0.020	0.069	0.761	0.008
26	0.743	112.00	0.018	7.00	0.463	0.049	0.067	0.228
27	0.738	51.00	0.010	3.00	0.055	0.100	0.562	0.056

General peak areas	Ellipse parameters x_0^g	y_0^g	a^g	b^g	p-values CF - EB	CF - NN	EB - NN	CF - AO
28	0.772	282.00	0.015	8.00	0.599	0.073	0.110	0.335
29	0.802	31.00	0.011	3.02	0.302	0.332	0.643	0.308
30	0.433	510.00	0.003	7.00	0.167	0.198	0.294	0.081
31	0.485	286.00	0.010	287.00	0.500	0.631	0.036	0.735
32	0.497	7.00	0.005	6.00	0.407	0.097	0.122	0.249
33	0.507	35.00	0.005	3.00	0.070	0.831	0.130	0.190
34	0.581	99.00	0.007	4.00	0.027	0.102	0.916	0.021
35	0.648	19.00	0.010	4.00	0.810	0.203	0.450	0.600
36	0.661	146.00	0.011	6.00	0.277	0.813	0.191	0.512
37	0.686	105.75	0.008	3.00	0.146	0.219	0.627	0.145
38	0.711	46.00	0.009	4.00	0.760	0.116	0.034	0.473
39	0.706	19.00	0.011	4.00	0.057	0.836	0.357	0.250
40	0.450	412.00	0.004	3.00	0.137	0.030	0.101	0.022
41	0.453	124.00	0.007	4.00	0.127	0.006	0.057	0.024
42	0.453	149.00	0.007	4.00	0.109	0.009	0.103	0.022
43	0.452	99.00	0.006	3.00	0.129	0.006	0.058	0.025
44	0.454	5.75	0.004	3.00	0.770	0.201	0.101	0.615
45	0.478	483.75	0.006	4.00	0.324	0.567	0.017	0.594
46	0.506	16.00	0.004	3.00	0.079	0.725	0.261	0.114
47	0.541	17.00	0.011	3.00	0.241	0.844	0.324	0.284
48	0.565	16.00	0.009	4.00	0.516	0.436	0.895	0.465
49	0.580	77.00	0.007	4.00	0.094	0.120	0.707	0.072
50	0.589	448.50	0.011	10.00	0.181	0.190	0.665	0.183
51	0.603	148.25	0.011	6.00	0.605	0.842	0.677	0.587
52	0.639	78.50	0.007	4.00	0.303	0.413	0.846	0.300
53	0.684	7.00	0.007	3.00	0.719	0.575	0.806	0.623
54	0.710	198.00	0.011	5.50	0.071	0.321	0.356	0.099
55	0.720	143.50	0.011	6.00	0.457	0.258	0.107	0.850
56	0.874	235.00	0.015	8.00	0.483	0.319	0.243	0.430
57	0.889	144.00	0.011	5.00	0.272	0.827	0.134	0.413
58	0.575	4.00	0.007	2.00	0.468	0.689	0.338	0.687
59	0.452	263.00	0.006	4.00	0.125	0.009	0.057	0.019
60	0.586	6.00	0.006	4.00	0.824	0.164	0.048	0.534
61	0.825	50.00	0.009	4.00	0.188	0.190	0.885	0.183

General	**Ellipse parameters**				**p-values**			
peak areas	x_0^g	y_0^g	a^g	b^g	CF - EB	CF - NN	EB - NN	CF - AO
62	0.836	9.00	0.009	2.00				
63	0.848	453.00	0.015	10.00	0.822	0.810	0.124	0.928
64	0.950	511.00	0.012	4.00				
65	1.328	514.00	0.010	4.00				
66	0.940	7.75	0.008	2.50				
67	0.649	189.00	0.011	6.00	0.622	0.580	0.951	0.560
68	0.959	138.00	0.008	3.00	0.391	0.273	0.524	0.258
69	0.435	531.50	0.004	5.00				
70	0.272	244.00	0.002	2.00				
71	0.840	251.00	0.003	2.00				
72	1.065	205.00	0.010	5.00	0.004	0.017	0.853	0.001
73	0.480	323.50	0.010	320.00	0.411	0.637	0.019	0.645
74	0.534	86.00	0.005	2.02	0.372	0.890	0.602	0.589
75	0.564	116.00	0.011	6.00	0.087	0.135	0.935	0.080
76	0.655	223.00	0.012	4.00	0.107	0.396	0.826	0.107
77	0.710	122.00	0.008	4.00	0.250	0.860	0.404	0.478
78	0.730	182.25	0.009	4.50	0.403	0.521	0.090	0.682
79	0.873	288.00	0.014	7.00	0.467	0.292	0.539	0.386
80	0.432	631.00	0.004	11.00				
81	0.772	347.00	0.014	9.00	0.704	0.703	0.928	0.680
82	0.803	638.00	0.012	5.00				
83	0.433	560.00	0.004	9.00				
84	0.454	212.00	0.006	4.00	0.106	0.005	0.049	0.015
85	0.877	71.50	0.012	5.00	0.899	0.145	0.104	0.554
86	0.602	207.25	0.009	5.00	0.206	0.752	0.189	0.318
87	1.276	8.00	0.009	2.00				
88	0.532	38.25	0.007	4.00	0.622	0.694	0.893	0.585
89	0.575	245.50	0.010	5.00	0.101	0.138	0.636	0.058
90	1.179	7.00	0.008	2.00				
91	0.479	206.25	0.007	7.00	0.380	0.728	0.014	0.593
92	0.513	237.00	0.007	4.00	0.200	0.489	0.851	0.233
93	0.478	384.25	0.006	5.00	0.320	0.561	0.008	0.588
94	0.478	424.00	0.006	5.00	0.367	0.525	0.010	0.642
95	0.659	169.00	0.008	4.00	0.156	0.573	0.157	0.505

General peak areas	Ellipse parameters x_0^g	y_0^g	a^g	b^g	p-values CF - EB	CF - NN	EB - NN	CF - AO
96	0.723	513.75	0.015	5.00	0.954	0.845	0.738	0.933
97	0.744	135.00	0.019	7.50	0.182	0.551	0.555	0.219
98	0.479	78.50	0.007	7.00	0.535	0.529	0.022	0.788
99	0.608	42.00	0.009	4.00	0.198	0.175	0.759	0.173
100	0.709	228.00	0.012	6.00	0.355	0.551	0.157	0.648
101	0.650	130.00	0.010	6.00	0.100	0.640	0.328	0.519
102	1.148	16.00	0.009	2.50				
103	0.453	174.25	0.007	4.50	0.119	0.005	0.054	0.020
104	0.453	299.00	0.005	4.00	0.088	0.010	0.071	0.011
105	0.478	123.50	0.010	9.00	0.477	0.641	0.023	0.703
106	0.624	6.00	0.008	3.00	0.536	0.049	0.131	0.214
107	1.400	5.00	0.007	2.00				
108	1.118	7.00	0.008	2.00				
109	0.016	509.00	0.016	4.00				
110	0.846	512.00	0.011	3.00	0.087	0.517	0.387	0.138
111	0.744	18.00	0.010	3.00	0.323	0.220	0.721	0.166
112	0.774	201.25	0.013	7.50	0.755	0.236	0.397	0.938
113	0.936	109.00	0.008	3.50				
114	0.988	280.00	0.009	4.00				
115	1.366	1.00	0.010	2.00				
116	0.784	128.00	0.006	3.00	0.911	0.539	0.539	0.775
117	0.403	36.00	0.007	3.00				
118	0.750	561.00	0.010	7.00				
119	1.298	510.00	0.010	4.00				
120	0.744	331.25	0.017	11.00	0.413	0.868	0.402	0.514
121	1.442	4.00	0.008	1.00				
122	0.649	369.00	0.010	10.00	0.179	0.476	0.358	0.230
123	1.324	5.00	0.004	2.00				
124	1.301	4.00	0.007	1.00				
125	1.429	512.00	0.004	3.00				
126	0.149	7.50	0.036	3.00				
127	1.034	8.00	0.009	3.00				
128	1.206	5.00	0.004	2.00				
129	0.770	18.00	0.010	3.00	0.342	0.866	0.443	0.446

General	**Ellipse parameters**				**p-values**			
peak areas	x_0^g	y_0^g	a^g	b^g	CF - EB	CF - NN	EB - NN	CF - AO
130	0.520	118.00	0.004	3.50	0.195	0.776	0.696	0.300
131	0.992	21.00	0.011	2.00				
132	0.600	174.25	0.012	5.50	0.138	0.761	0.393	0.198
133	1.076	6.00	0.006	2.00				
134	1.183	511.00	0.011	4.00				
135	1.238	3.25	0.009	1.50				
136	0.852	138.00	0.007	3.00				
137	0.600	516.00	0.011	8.00	0.198	0.211	0.891	0.189
138	1.096	510.00	0.021	3.50				
139	0.525	435.50	0.006	4.00	0.114	0.773	0.546	0.214
140	0.909	228.50	0.007	3.00	0.515	0.079	0.175	0.285
141	1.053	581.50	0.004	6.00				
142	0.354	20.00	0.011	2.00				
143	0.675	496.00	0.013	3.00	0.577	0.890	0.532	0.694
144	1.384	185.00	0.001	3.00				
145	0.563	499.75	0.004	4.50	0.158	0.210	0.901	0.144
146	0.008	22.00	0.008	2.00				
147	0.056	23.00	0.023	3.00				
148	0.239	23.00	0.039	2.00				

Table A.5: (Standardised) discriminant coefficients of the discriminant functions for the considered sample set of lung cancer patients for different comparisons on the different steps of the stepwise selection (Subsection 7.2.2), and error rates estimated with the leave-one-out method: the pairwise comparisons are indicated by the headings using the nominations CF for circular focuses, EB for endobronchial carcinomas, and NN for other kinds of carcinoma, as well as AO comprising all tumors exclusive circular focuses. The variable names are oriented on Table A.4; lines in italic font correspond to variables that were excluded in the range of the stepwise selection, which was based on the error rate and the standardised discriminant coefficients.

LDA		**Full model**			**Step 1**			**Step2**		
variable		Discr.	Stand.	Error	Discr.	Stand.	Error	Discr.	Stand.	Error
name		coeff.	coeff.	rate	coeff.	coeff.	rate	coeff.	coeff.	rate
CF	V9	18	0.5	0.23						
vs	V25	170	3.3	0.29						
EB	V34	-429	-8.1	0.29						
	V72	-2527	-49.2	0.29						

LDA		**Full model**			**Step 1**			**Step2**		
variable		Discr.	Stand.	Error	Discr.	Stand.	Error	Discr.	Stand.	Error
name		coeff.	coeff.	rate	coeff.	coeff.	rate	coeff.	coeff.	rate
CF	V1	46936	2745.9	0.21	40386	2316.7	0.21	31735	1927.4	0.16
vs	V2	-1791	-104.8	0.21	-1573	-90.2	0.21	-1320	-80.2	0.21
NN	V3	1498	87.7	0.26	1210	69.4	0.21	963	58.5	0.21
	V4	-3047	-178.2	0.21	-2576	-147.7	0.11	-1838	-111.6	0.11
	V5	*-128*	*-7.5*	*0.11*						
	V40	*72*	*3.6*	*0.16*	*14*	*0.6*	*0.11*	*120*	*7.2*	*0.05*
	V41	-7967	-468.0	0.47	-7085	-406.7	0.47	-5379	-326.6	0.47
	V42	*-184*	*-11.0*	*0.21*	*-158*	*-9.1*	*0.05*			
	V43	1360	81.0	0.21	1502	86.6	0.05	1242	75.7	0.05
	V59	-173	-11.0	0.16	-327	-19.0	0.05	-543	-33.2	0.05
	V72	-31231	-1838.2	0.42	-27377	-1573.8	0.32	-20615	-1254.5	0.26
	V84	2089	122.5	0.11	1264	72.5	0.11	920	55.9	0.05
	V103	9674	566.0	0.42	8430	483.3	0.26	6188	375.1	0.16
	V104	-2360	-136.4	0.11	-1605	-91.4	0.05	-1126	-67.9	0.05
		Step 3								
CF	V1	23190	1460.3	0.11						
vs	V2	-1006	-63.3	0.21						
NN	V3	711	44.8	0.21						

	V4	-1198	-75.4	0.11					
	V5								
	V40								
	V41	-4114	-258.3	0.37					
	V42								
	V43	1125	70.6	0.05					
	V59	-627	-39.4	0.00					
	V72	-16929	-1066.2	0.26					
	V84	505	31.9	0.05					
	V103	4376	274.8	0.11					
	V104	-403	-25.3	0.00					
LDA		**Full model**			**Step 1**			**Step2**	
variable		Discr.	Stand.	Error	Discr.	Stand.	Error	Discr.	Stand.	Error
name		coeff.	coeff.	rate	coeff.	coeff.	rate	coeff.	coeff.	rate
EB	V6	-209	-19.7	0.50	-181	-18.2	0.50	-153	-16.2	0.37
vs	*V7*	*23*	*2.2*	*0.43*	*13*	*1.3*	*0.43*	*14*	*1.6*	*0.37*
NN	V45	235	22.1	0.47	176	17.5	0.43	144	14.8	0.37
	V73	*-8*	*-0.8*	*0.47*	*10*	*1.1*	*0.40*			
	V91	249	23.4	0.43	198	19.8	0.43	162	16.9	0.40
	V93	-528	-49.7	0.50	-396	-39.7	0.43	-327	-33.9	0.43
	V94	*27*	*2.5*	*0.40*						
	V98	42	4.0	0.40	35	3.6	0.40	30	3.3	0.40
	V105	*77*	*7.2*	*0.40*	*68*	*6.8*	*0.40*	*64*	*6.6*	*0.37*
		Step 3			**Step 4**					
EB	V6	-120	-7.9	0.40	-77	-5.6	0.33			
vs	*V7*									
NN	V45	117	7.7	0.43	81	5.8	0.27			
	V73									
	V91	122	8.0	0.33	105	7.5	0.37			
	V93	-259	-17.0	0.40	-190	-13.6	0.33			
	V94									
	V98	12	0.8	0.33	37	2.7	0.27			
	V105	*81*	*5.3*	*0.20*						

LDA variable name		Full model Discr. coeff.	Stand. coeff.	Error rate	Step 1 Discr. coeff.	Stand. coeff.	Error rate	Step2 Discr. coeff.	Stand. coeff.	Error rate
CF	*V4*	*123*	*6.8*	*0.37*	*169*	*9.7*	*0.34*	*148*	*8.8*	*0.29*
vs	*V9*	*-68*	*-3.8*	*0.34*	*-34*	*-1.9*	*0.29*	*-28*	*-1.6*	*0.26*
AO	V25	-961	-53.1	0.42	-940	-53.7	0.42	-843	-49.8	0.42
	V34	*865*	*47.8*	*0.42*	*743*	*42.5*	*0.42*	*634*	*37.5*	*0.32*
	V40	*335*	*18.5*	*0.32*						
	V41	428	23.6	0.37	233	13.3	0.32	143	8.5	0.29
	V42	*42*	*2.3*	*0.34*	*4*	*0.2*	*0.29*	*-9*	*-0.6*	*0.24*
	V43	*-190*	*-10.5*	*0.39*	*-85*	*-4.9*	*0.34*	*-140*	*-8.4*	*0.24*
	V59	-616	-33.9	0.37	-116	-6.6	0.29	160	9.4	0.32
	V72	8236	455.5	0.34	7480	427.5	0.37	6844	405.2	0.37
	V84	-181	-10.1	0.42	-547	-31.2	0.26	-405	-24.0	0.26
	V103	*65*	*3.6*	*0.39*	*-8*	*-0.4*	*0.29*	*54*	*3.2*	*0.24*
	V104	*-72*	*-3.9*	*0.37*	*267*	*15.3*	*0.26*			
		Step 3			**Step 4**			**Step 5**		
CF	*V4*	*121*	*7.5*	*0.26*	*102*	*6.6*	*0.21*	*124*	*8.9*	*0.18*
vs	*V9*	*-27*	*-1.7*	*0.24*	*-25*	*-1.6*	*0.21*			
AO	V25	-758	-46.8	0.39	-672	-43.4	0.34	-615	-43.6	0.26
	V34	*553*	*34.2*	*0.26*	*506*	*32.7*	*0.26*	*545*	*38.8*	*0.18*
	V40									
	V41	119	7.4	0.24	88	5.8	0.21	57	4.1	0.21
	V42									
	V43	*-113*	*-7.1*	*0.21*	*-71*	*-4.7*	*0.21*	*-85*	*-6.1*	*0.21*
	V59	141	8.7	0.26	118	7.6	0.21	109	7.7	0.21
	V72	6152	380.6	0.34	5463	353.5	0.29	5215	370.9	0.29
	V84	-360	-22.3	0.21	-279	-18.1	0.21	-245	-17.4	0.21
	V103	*47*	*2.9*	*0.21*						
	V104									
		Step 6			**Step 7**			**Step 8**		
CF	*V4*									
vs	*V9*									
AO	V25	-519	-37.5	0.34	-477	-37.5	0.29	-375	-30.1	0.29
	V34	*445*	*32.3*	*0.16*						
	V40									

V41	81	6.0	0.21	69	5.3	0.21	18	1.5	0.24
V42									
V43	*-5*	*-0.5*	*0.18*	*-17*	*-1.3*	*0.16*			
V59	85	6.2	0.21	72	5.6	0.21	42	3.4	0.21
V72	4455	323.1	0.26	3863	305.3	0.24	3201	259.3	0.24
V84	-187	-13.7	0.21	-150	-11.7	0.21	-81	-6.6	0.24
V103									
V104									

Table A.7: Discriminant values for the considered sample set of lung cancer patients for different comparisons, using the entire sample set (overall) and the leave-one-out method (loo), respectively: the pairwise comparisons are indicated by the headings using the nominations CF for circular focuses, EB for endobronchial carcinomas, and NN for other kinds of carcinoma, as well as AO comprising all tumors exclusive circular focuses. Lines of patients that do not contain discriminant values were not included in the considered comparison.

		CF vs EB		**CF vs NN**		**EB vs NN**		**CF vs AO**	
		full	loo	full	loo	full	loo	full	loo
Circular focus (CF)	1	-1.40	-1.19	-1.80	-1.54			-0.87	-0.77
	2	-0.51	-0.45	-3.97	2.01			-0.31	0.70
	3	-0.64	-0.54	-3.13	-2.51			-0.10	0.09
	4	-1.65	-1.60	-4.82	-4.91			-2.61	-2.77
	5	-0.06	0.18	-5.58	-9.36			-0.38	-0.33
	6	-1.43	-1.41	-5.50	-7.94			-1.17	-1.14
	7	-1.49	-1.46	-4.30	-4.15			-1.22	-1.13
	8	-0.34	-0.09	-4.40	-4.57			-0.67	-0.63
	9	-0.21	-0.05	-2.85	-2.21			-0.81	-0.77
	10	-1.38	-1.35	-5.51	-8.93			-1.94	-1.99
Endobronchial	1	0.41	0.27			-2.25	-2.39	0.11	0.02
tumor (EB)	2	-0.05	-0.27			-2.03	-2.08	1.08	1.05
	3	0.61	0.20			-0.70	-0.48	0.47	0.14
	4	1.14	1.11			-2.42	-2.55	0.31	0.25
	5	1.13	1.05			-0.70	-0.52	1.54	1.50
	6	0.37	0.15			-0.19	-0.02	0.28	0.11
	7	1.08	1.05			0.06	0.33	1.42	1.40
	8	-0.36	-0.41			-0.28	-0.07	-0.26	-0.31
	9	0.84	0.69			-0.34	-0.13	0.11	0.06
	10	2.01	2.07			0.25	0.30	2.23	2.31
	11	-0.88	-1.50			-0.18	-0.03	-0.73	-1.02
	12	2.57	3.44			-0.61	-0.49	1.97	2.01
	13	-0.89	-0.99			-1.87	-1.91	-0.54	-0.66
	14	1.37	1.36			-1.53	-1.52	1.40	1.38
	15	-0.47	-1.09			-1.56	-1.53	0.14	0.03
	16	0.59	0.51			0.94	1.31	2.55	2.80
	17	0.50	0.40			-1.53	-1.51	-0.32	-0.40

		CF vs EB		CF vs NN		EB vs NN		CF vs AO	
		full	loo	full	loo	full	loo	full	loo
	18	2.42	2.75			-0.61	-0.05	0.69	0.33
	19	2.29	2.47			-1.38	-1.33	NA	NA
	20	3.19	5.13			-2.03	-2.07	2.95	4.11
	21	1.30	1.25			-1.95	-1.99	0.35	0.26
Neither nor (NN)	1			4.04	2.11	0.19	0.12	1.34	1.33
	2			4.28	3.18	0.92	0.76	2.86	3.89
	3			3.57	2.89	-0.11	-1.12	2.32	2.41
	4			3.80	3.38	3.48	4.69	0.33	0.26
	5			4.47	29.41	1.86	0.55	1.75	6.77
	6			3.39	1.56	0.73	0.43	1.04	0.97
	7			5.21	9.31	1.16	0.89	0.47	0.18
	8			4.55	5.17	0.99	-1.35	NA	NA
	9			4.38	4.20	-0.24	-0.61	2.37	2.56

Appendix B

Figures

This chapter contains illustrations of relevant variables for comparisons of different tumor kinds (Section 7.2), presented in density plots for three different comparisons (Fig. B.1, B.2, and B.3) and marked in the heatmap of an instance measurement for one of the comparisons (Fig. B.4). To avoid the interruption of the reading flow in the main part of the thesis, these figures were put into the appendix to still allow a complete overview about the peak variables giving the base for the determined discriminant functions.

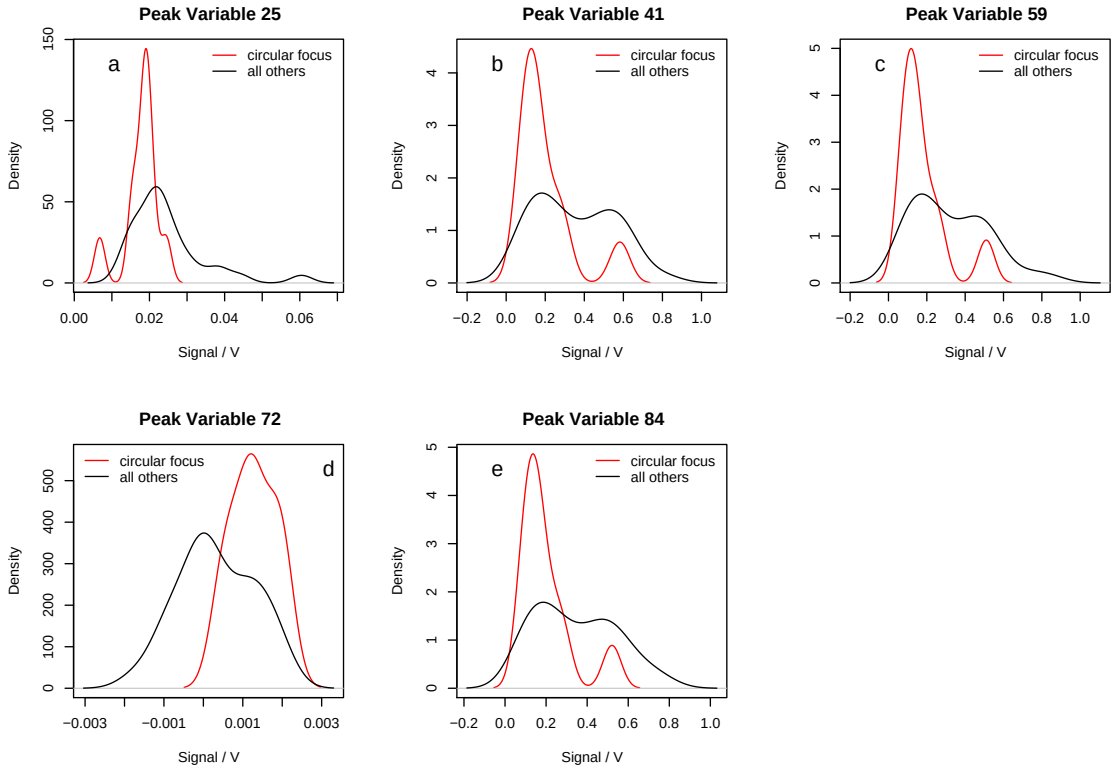

Figure B.1: Plots overlaying the density curves for the two groups of patients with circular focuses and all other carcinoma inclusive endobronchial tumors for the relevant peak variables of this comparison (a) $V25$, (b) $V41$, (c) $V59$, (d) $V72$, and (e) $V84$.

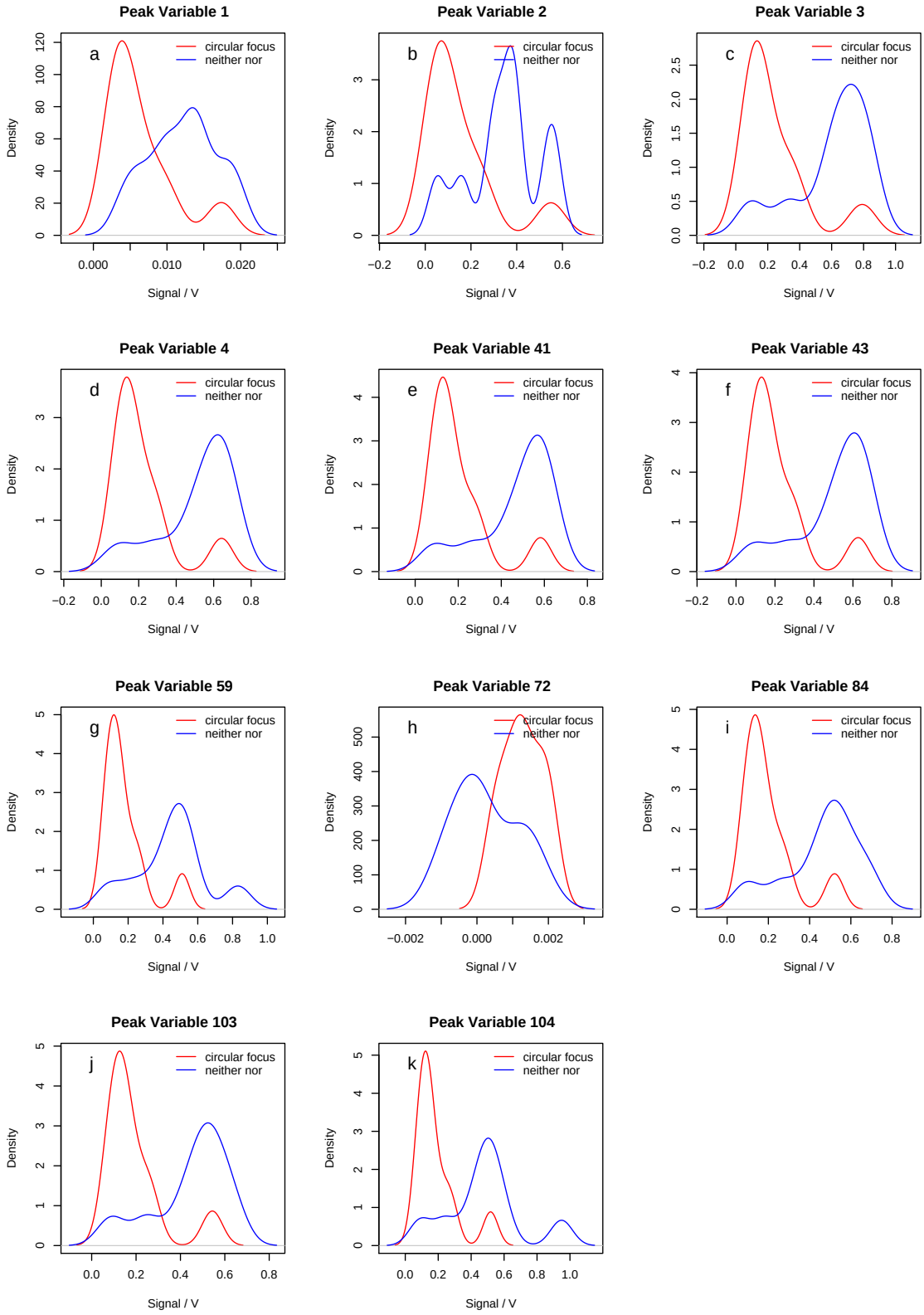

Figure B.2: Plots overlaying the density curves for the two groups of patients with circular focuses and other tumor kinds exclusive endobronchial tumors for the relevant peak variables of this comparison (a) $V1$, (b) $V2$, (c) $V3$, (d) $V4$, (e) $V41$, (f) $V43$, (g) $V59$, (h) $V72$, (i) $V84$, (j) $V103$, and (k) $V104$.

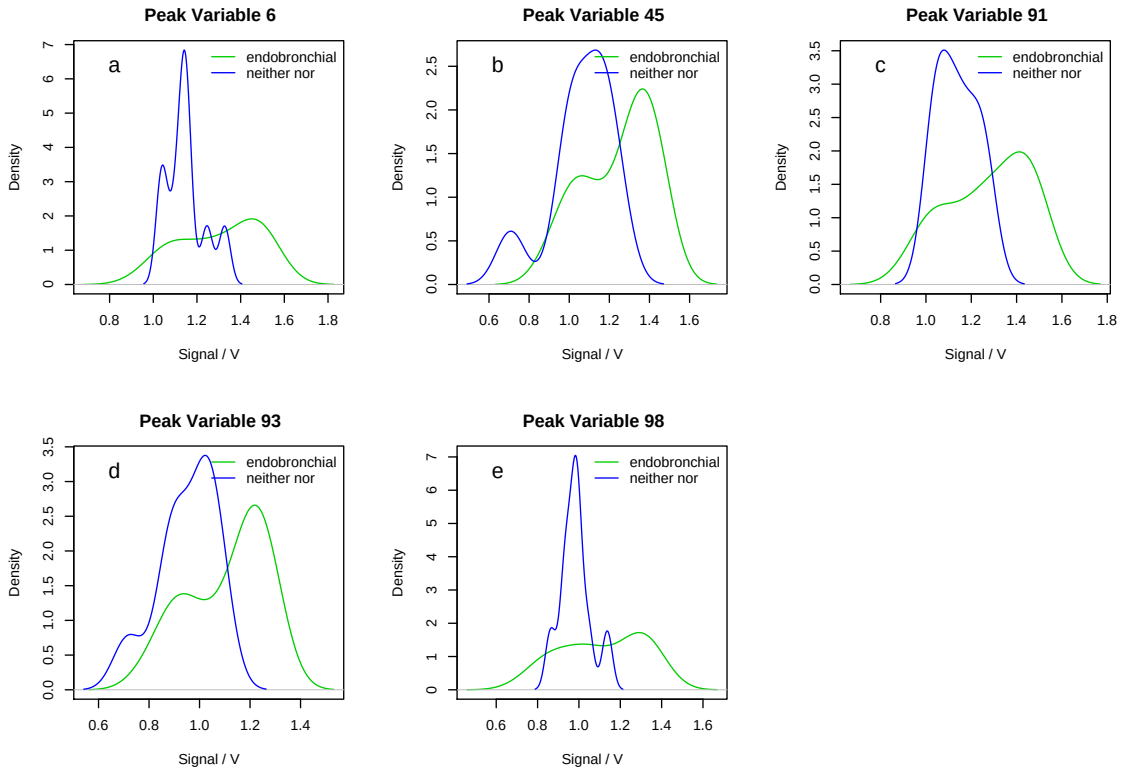

Figure B.3: Plots overlaying the density curves for the two groups of patients with endobronchial tumors and other carcinoma exclusive circular focuses for the relevant peak variables of this comparison (a) $V6$, (b) $V45$, (c) $V91$, (d) $V93$, and (e) $V98$.

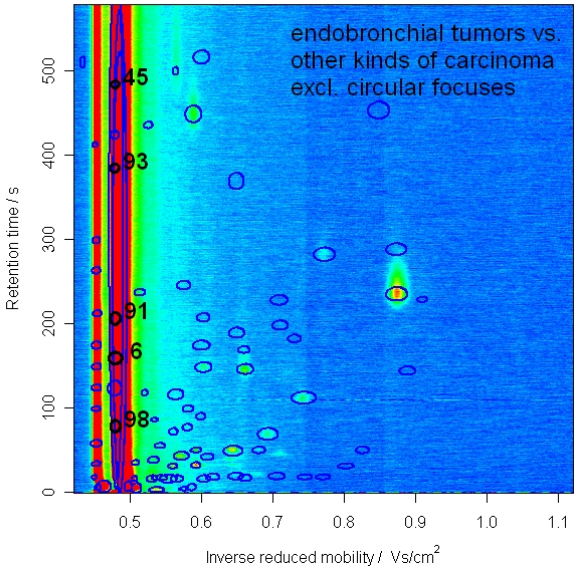

Figure B.4: Heatmap of an instance breath measurement showing the general peak areas corresponding to the relevant variables for the comparison of endobronchial tumors with other carcinoma excluding circular focuses.

Bibliography

Bader, S., Baumbach, J., and Urfer, W. (accepted in 2008): Preprocessing of Ion Mobility Spectra by Lognormal Detailing and Wavelet Transform. *International Journal of Ion Mobility Spectrometry*.

Bader, S., Urfer, W., and Baumbach, J. (2005): Processing Ion Mobility Spectrometry Data to Characterize Group Differences in a Multiple Class Comparison. *International Journal of Ion Mobility Spectrometry*, **8**, 1–4.

Bader, S., Urfer, W., and Baumbach, J. (2006): Reduction of Ion Mobility Spectrometry Data by Clustering Characteristic Peak Structures. *Journal of Chemometrics*, **20**, 128–135.

Baumbach, J., Bader, S., Urfer, W., Ruzsanyi, V., Westhoff, M., Litterst, P., and Freitag, L. (submitted in 2007): Rapid Classification of Lung Diseases by GC Coupled to Ion Mobility Spectrometry. *Journal of Chromatography A*.

Baumbach, J. I. (2006): Process Analysis Using Ion Mobility Spectrometry. *Analytical and Bioanalytical Chemistry*, **384**, 1059–1070.

Baumbach, J. I., Vautz, W., Ruzsányi, V., and Freitag, L. (2005): Metabolites in Human Breath: Ion Mobility Spectrometers as Diagnostic Tools for Lung Diseases. In: A. Amann and D. Smith (Hrsg.) *Breath Gas Analysis for Medical Diagnostics*. World Scientific.

Bruce, L., Balch, T., and Veloso, M. (2000): Fast and Inexpensive Color Image Segmentation for Interactive Robots. In: *Proceedings of IROS-2000*, 2061–2066.

Byrd, R., Lu, P., Nocedal, J., and Zhu, C. (1995): A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, **16** (5), 1190–1208. URL `http://link.aip.org/link/?SCE/16/1190/1`.

Cai, C. and Harrington, P. (1998): Different Discrete Wavelet Transforms Applied to Denoising Analytical Data. *Journal of Chemical Information and Computer Sciences*, **38** (6), 1161–1170.

Calinski, R. B. and Harabasz, J. (1974): A Dendrite Method for Cluster Analysis. *Communications in Statistics*, **3**, 1–27.

Cleveland, W. S. (1979): Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, **74**, 829–836.

Daubechies, I. (1992): *Ten Lectures on Wavelets*. SIAM.

Donoho, D. and Johnstone, I. (1995): Adapting to Unknown Smoothness Via Wavelet Shrinkage. *Journal of the American Statistical Association*, **90**, 362–366.

Eiceman, G. and Karpas, Z. (2005): *Ion Mobility Spectrometry*. CRC Press, Taylor & Francis.

Fisher, R. A. (1936): The Use of Multiple Measurement in Taxonomic Problems. *Annals of Eugenics*, **7**, 179–188.

Hartigan, J. A. and Wong, M. A. (1979): A K-Means Clustering Algorithm. *Applied Statistics*, **28**, 100–108.

Holm, S. (1979): A Simple Sequentially Rejective Bonferroni Test Procedure. *Scandinavian Journal of Statistics*, **6**, 65–70.

Kaufman, L. and Rousseeuw, P. J. (1990): *Finding Groups in Data - An Introduction to Cluster Analysis*. Wiley.

Lachenbruch, P. A. and Mickey, M. R. (1968): Estimation of Error Rates in Discriminant Analysis. *Technometrics*, **10**, 1–11.

Liu, Z., Patterson, D., and Lee, M. (1995): Geometric Approach to Factor Analysis for the Estimation of Orthogonality and Practical Peak Capacity in Comprehensive Two-Dimensional Separations. *Analytical Chemistry*, **67**, 3840–3845.

Milligan, W. G. and Cooper, M. C. (1985): An Examination of Procedures for Determining the Number of Clusters in a Data Set. *Psychometrika*, **50**, 159–179.

Percival, D. B. and Walden, A. T. (2000): *Wavelet Methods for Time Series Analysis*. Cambridge University Press.

Prasad, S., Pierce, K., Schmidt, H., Rao, J., Güth, R., Bader, S., Synovec, R., Smith, G., and Eiceman, G. (2007): Analysis of Bacteria by Pyrolysis Gas ChromatographyDifferential Mobility Spectrometry and Isolation of Chemical Components with a Dependence on Growth Temperature. *The Analyst*, **132**, 1031–1039.

Punj, G. and Stewart, D. W. (1983): Cluster Analysis in Marketing Research: Review and Suggestions for Application. *Journal of Marketing Research*, **20**, 134–148.

R Development Core Team (2007): *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

Randolph, T. and Yasui, Y. (2006): Multiscale Processing of Mass Spectrometry Data. *Biometrics*, **62**, 589–597.

Ripley, B. D. (1996): *Pattern Recognition and Neural Networks*. Cambridge University Press.

Rousseeuw, P. J. (1987): Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, **20**, 53–65.

Ruzsanyi, V., Baumbach, J., Sielemann, S., Litterst, P., Westhoff, M., and Freitag, L. (2005): Detection of Human Metabolites Using Multi-Capillary Columns Coupled to Ion Mobility Spectrometers. *Journal of Chromatography A*, **1084**, 145–151.

Schmidt, H., Tadjimukhamedov, F., Smith, G., Mohrentz, I., and Eiceman, G. (2004): Micro-Fabricated Differential Mobility Spectrometry with Pyrolysis Gas Chromatography for Chemical Characterisation of Bacteria. *Analytical and Bioanalytical Chemistry*, **76**, 5208–5217.

Urbas, A. and Harrington, P. (2001): Two-Dimensional Wavelet Compression of Ion Mobility Spectra. *Analytica Chimica Acta*, **446**, 391–410.

Westhoff, M., Litterst, P., Freitag, L., Bader, S., and Baumbach, J. (submitted in 2008): Ion Mobility Spectrometry for Detection of Volatile Organic Compounds in Exhaled Air of Patients with Bronchial Carcinoma Results of a Pilot Study. *Chest*.