Solving the correspondence problem in analytical chemistry

Automated methods for alignment and quantification of multiple signals

Erik Alm

Doctoral Thesis
Department of Analytical Chemistry
Stockholm University
2012
Academic dissertation for the Doctor of Philosophy degree in Analytical Chemistry to be publicly defended on Friday, May 25th in Magnélisalen, Kemiska Övningslaboratoriet, Svante Arrhenius väg 16, Stockholm, Sweden.

© Erik Alm, Stockholm 2012
Printed by Universitetsservice US-AB, Stockholm
**Abstract**

When applying statistical data analysis techniques to analytical chemical data, all variables must have correspondence over the samples dimension in order for the analysis to generate meaningful results. Peak shifts in NMR and chromatography destroys that correspondence and creates data matrices that have to be aligned before analysis. In this thesis, new methods are introduced that allow for automated transformation from unaligned raw data to aligned data matrices where each column corresponds to a unique signal. These methods are based on linear multivariate models for the peak shifts and Hough transform for establishing the parameters of these linear models. Methods for quantification under difficult conditions, such as crowded spectral regions, noisy data and unknown peak identities are also introduced. These methods include automated peak selection and a robust method for background subtraction. This thesis focuses on the processing of the data; the experimental work is secondary and is not discussed in great detail.

The developed methods are put together in a full procedure that takes us from raw data to a table of concentrations in a matter of minutes. The procedure is applied to $^1$H–NMR data from biological samples, which is one of the toughest alignment tasks available in the field of analytical chemistry. It is shown that the procedure performs consistently on the same level as much more labor intensive manual techniques such as Chenomx NMRSuite spectral profiling.

Several kinds of datasets are evaluated using the procedure. Most of the data is from the field of *Metabolomics*, where the goal is to establish
concentrations of as many small molecules as possible in biological samples.

**Figure 1:** Schematic representation of correctly assigned correspondence for a peak in a $^1$H–NMR dataset measured on urine samples.
Populärvetenskaplig sammanfattning

Ett stort problem när man med avancerad kemisk analys försöker bestämma absoluta eller relativa halter av många ämnen i många biologiska prover samtidigt är att fastställa vilka instrumentsignaler som hör till vilken kemisk förening. Om man inte säkerställer att alla signaler man mätt har korrekt korrespondens så är risken stor att man i den efterföljande statistiska analysen jämför signaler som inte alls hör till samma kemiska förening, man jämför äpplen med päron. Detta kan leda till totalt felaktiga slutsatser om det system man undersöker och detta kan i värsta fall få långtgående konsekvenser där man arbetar utifrån felaktiga antaganden under lång tid.


I avhandlingen diskuteras även några olika applikationer av de
tekniker som presenteras. Främst appliceras teknikerna inom fältet Metabolomics, där man analyserar alla små molekyler i biologiska prover såsom urin, blodplasma etc. Metoderna kan även användas för att skapa data med flerdimensionell korrespondens som kan modelleras med multidimensionella metoder.
List of papers

I. **Vibrational overtone combination spectroscopy (VOCSY)—a new way of using IR and NIR data**
   Erik Alm, Rasmus Bro, Søren B. Engelsen, Bo Karlberg and Ralf J. O. Torgrip
   Analytical and Bioanalytical Chemistry, 2007, Volume 388, Number 1, Pages 179–188
   The author was responsible for part of the idea, all the experimental work and data evaluation, and partly for writing the paper.

II. **Proof of principle of a generalized fuzzy Hough transform approach to peak alignment of one–dimensional 1H NMR data**
    Analytical and Bioanalytical Chemistry, 2007, Volume 389, Number 3, Pages 875–885
    The author was responsible for part of the idea, part of the data evaluation and partly for writing the paper.

III. **A solution to the 1D NMR alignment problem using an extended generalized fuzzy Hough transform and mode support**
    Erik Alm, Ralf J. O. Torgrip, K. Magnus Åberg, Ina Schuppe–Koistinen and Johan Lindberg
    Analytical and Bioanalytical Chemistry, 2009, Volume 395, Number 1, Pages 213–223
    The author was responsible for the idea, all the data evaluation and for writing most of the paper.

IV. **Time–resolved biomarker discovery in 1H–NMR data using generalized fuzzy Hough transform alignment and parallel factor analysis**
    Erik Alm, Ralf J. O. Torgrip, K. Magnus Åberg, Ina Schuppe–Koistinen and Johan Lindberg
    Analytical and Bioanalytical Chemistry, 2010, Volume 396, Number 5, Pages 1681–1689
The author was responsible for the idea, all the data evaluation and for writing most of the paper.

V. **Automated annotation and quantification of metabolites in $^1$H–NMR data of biological origin**  
Erik Alm, Tove Slagbrand, K. Magnus Åberg, Erik Wahlström, Ingela Gustafsson and Johan Lindberg  
Analytical and Bioanalytical Chemistry, 2012, Volume 403, Number 2, Pages 443–455  
The author was responsible for the idea, the supervision of the project, most of the data evaluation and for writing the paper.

**Other papers by the author**

VI. **A note on normalization of biofluid 1D $^1$H–NMR data**  
Ralf J. O. Torgrip, K. Magnus Åberg, Erik Alm, Ina Schuppe–Koistinen and Johan Lindberg  
Metabolomics, 2008, Volume 4, Number 1, Pages 114–121

VII. **The correspondence problem for metabonomics datasets**  
K. Magnus Åberg, Erik Alm and Ralf J. O. Torgrip  
Analytical and Bioanalytical Chemistry, 2009, Volume 394, Number 1, Pages 151–162

VIII. **Warping and alignment technologies for inter–sample feature correspondence in 1D H–NMR, chromatography–, and capillary electrophoresis–mass spectrometry data**  
Ralf J. O. Torgrip, Erik Alm, K. Magnus Åberg  
Bioanalytical Reviews, 2010, Volume 1, Number 2, Pages 105–116
Abbreviations

COSY Correlation spectroscopy
COW Correlation optimized warping
DTW Dynamic time warping
FID Free induction decay or flame ionization detector
GC Gas chromatography
GFHT Generalized fuzzy Hough transform
HIT Hough indicator tensor
HSQC Heteronuclear single quantum coherence spectroscopy
IR Infrared
LC Liquid chromatography
MALDI Matrix-assisted laser desorption/ionization
MS Mass spectrometry
NMR Nuclear magnetic resonance
NIR Near infrared
PARAFAC Parallel factor analysis
PARS Peak alignment using reduced set mapping
PCA Principal component analysis
PCR Principal component regression
PLS Partial least squares
RMSEP Root mean square error of prediction
SIM Selected ion monitoring
S/N Signal to noise ratio
SRM Selected reaction monitoring
STOCSY Statistical total correlation spectroscopy
TIC Total ion current
UV Ultraviolet
VIS Visual
VOCSY Vibrational overtone combination spectroscopy
XC Chromatography (unspecified)
The correspondence problem

Introduction

The topic of this thesis is the correspondence problem that arises when analyzing many kinds of first order data. (Although the problem also arises with higher order data, which is beyond the scope of this thesis.) The correspondence problem can be stated as a lack of correspondence between variables:

When each column (variable) of the data matrix to be statistically analyzed does not originate from the same source, the results of the analysis will be meaningless.

Figure 2: Sample $^1$H-NMR spectra from a plant metabonomic dataset. The top pane shows superimposed spectra from four samples and the bottom pane shows a heat map of all the samples in the dataset. White
indicates low intensity and black indicates high intensity. It can clearly be seen from this figure that the positions of the peaks vary from sample to sample in a way that would be detrimental to the quality of results from multivariate analysis techniques.

This problem has hampered techniques that yield complex first-order data such as $^1$H-NMR spectroscopy of complex samples$^1$, figure 2, and chromatography-mass spectrometry (XC/MS) of complex samples for a long time; and renders the high resolution of modern-day instruments useless when comparing a large number of signals. The problem mainly occurs when many compounds are analyzed in many analysis runs over a long period of time and the results are compared between samples. This is because it becomes more difficult to manually assign correspondence as the number of sample constituents to be analyzed increases$^2$, $^3$ and instrument drift tends to be more pronounced over longer time-spans.

To take full advantage of the high resolving power of modern-day equipment, automated algorithms for assigning correspondence are required. In this thesis, new methods for this purpose will be considered and compared to older methods, as well as to the direct use of raw data.

**Correspondence algorithms**

I will now briefly discuss some of the techniques that have previously been developed for correspondence assignment in first-order datasets.

**Bucketing (Binning)**

The most straightforward way to deal with the correspondence
problem is to segment the independent variable and sum the signals within each segment, forming a new aggregate variable. This technique is called *bucketing* or *binning*\(^4\), **figure 3**. It has one very strong point, its simplicity. A simple technique usually works well straightaway, without the need for tedious development and refinement. Bucketing however also has one very severe drawback; it effectively reduces the resolution of the analytical technique to the point where one may as well use a much cruder instrument. It also has some less obvious drawbacks. For example, it cannot establish correspondence when peaks swap places between samples and it runs the risk of splitting peaks across neighboring segments. For these reasons, to fully exploit the potential of a high-resolution analytical instrument, one should look further than bucketing. Recently, more refined bucketing algorithms have been devised which may be able to overcome some of the drawbacks\(^5\).
Figure 3: The bucketing method in its most basic form, with the same data as in A typical bucket size for $^1$H-NMR, 0.04 ppm, has been used (boundaries are indicated by vertical lines). The bottom pane shows a heat map of the data after bucketing. It is easy to see that the resolution of the data has been drastically reduced. Even with this great loss in resolution, we have still not been able to solve the correspondence problem in this case. Peaks shift between buckets, giving rise to false signal increases and/or decreases in all the buckets except 1 and 8. This would lead to incorrect conclusions about the system if multivariate data analysis were performed on the bucketed data.

Warping

The next class of techniques that have been developed to establish correspondence between samples are known as warping techniques$^{6-10}$. These techniques require the specification of a target spectrum/chromatogram. A piecewise stretching function is constructed to maximize the similarity between the spectrum/chromatogram under
consideration and the target. Warping, unlike bucketing, does not reduce the resolution of the data matrix; so that the subsequent data analysis can benefit from the data having been generated by a high resolution instrument. The drawbacks of warping techniques are however not negligible. They cannot handle peaks that change place from sample to sample (this causes other peaks to lose correspondence as well); they have parameters that have to be optimized; and they tend to yield unsatisfactory results in crowded regions of the spectra/chromatograms, figure 4. They are a step up from bucketing in that they retain the resolution of the instrument, but a step down in simplicity and robustness. The differences between the various warping techniques are essentially in how they measure similarity between the target and the warped spectrum/chromatogram. Examples of warping techniques are correlation optimized warping (COW) and dynamic time warping (DTW)\textsuperscript{11,12}. 
The COW technique applied to the spectra shown in figures 2–3. The bottom pane shows a heat map of the results from the warping. It is obvious that many of the peaks have been incorrectly or insufficiently aligned.

**Discrete alignment techniques**

The final class of techniques are known as *peak alignment* techniques. I will refer to them as *discrete alignment techniques* in order to not confuse them with alignment techniques in general. The goal of discrete alignment is to reduce the raw data matrix to a smaller, *fully synchronized data matrix* in which every column has correspondence (i.e., describes the same chemical phenomenon for each sample) and the number of columns is equal to the number of unique peaks in the dataset. This representation is the most informative for statistical analysis since its mathematical rank is much closer to the chemical rank of the system than is the rank of other representations.
The goal of peak alignment is to establish the position of a peak in all the samples for which the peak is present and to note that the peak is missing in all the samples for which it is not present. Ultimately this process has to be repeated for all peaks in the dataset. One method for doing this is peak alignment using reduced-set mapping (PARS\textsuperscript{13-15}). The problem with this class of techniques has to date been lack of reliability. One cannot afford erroneous assignments when dealing with a fully automated process that generates output that is input directly into the statistical analysis. Another potential weakness is that it relies on the results of a peak-detection algorithm. If the peak-detection results are unsatisfactory, a discrete alignment algorithm cannot produce satisfactory results.

In this thesis I investigate peak alignment at a deeper level, using as much information about the inherent structure of the peak shifts as possible.

Data

Sources of data

Before one even considers the task of peak alignment, it is necessary to know the properties of the data being examined. Different experimental techniques yield data that are different in some respects and similar in others. The most common techniques that generate first-order data are chromatography and spectroscopy. Light spectroscopy (for example, UV/VIS, mid-IR\textsuperscript{16}, and NIR\textsuperscript{17, 18}) data generally suffer from few correspondence problems. This is one of the major reasons that the field of
chemometrics has historically been vastly more successful in dealing with this kind of data than with chromatographic data or other kinds of spectroscopic data. NMR spectroscopy data suffer from many peak shifts, owing to variable sample properties, and would benefit greatly from a reliable technique for correspondence assignment; one that does not reduce the data resolution the way bucketing does. Chromatography data\textsuperscript{19} suffer from correspondence problems as well, mainly because of the dynamic nature of the chromatographic system. Columns degrade and are replaced, injections are not completely reproducible, sample constituents may interact with each other and affect the retention time, etc.

Most of these factors are more-or-less out of our control; we cannot foresee the exact peak shifts that will be present in the next sample run, even knowing the shifts in all previous samples. The factors involved are just too difficult to predict and often behave in a seemingly random manner.

**Infrared and Near-infrared spectroscopy**

In paper I, the alignment of combined mid-infrared and near-infrared data was presented; the aim being to exploit the additional modes of spectral overtones. Mid-infrared (mid-IR) and near-infrared (NIR) spectroscopy measure the differences in energy levels of molecular vibrational modes. The energy of a mode can be expressed as the frequency of electromagnetic radiation that excites the mode.

A molecule has $6N-3$ fundamental vibrational modes, where $N$ is the number of atoms in the molecule. Each fundamental mode has corresponding $2^{\text{nd}}$, $3^{\text{rd}}$, etc., overtones at approximately twice, three times,
etc., the frequency of the fundamental mode. Because of anharmonicity in the vibrations, the overtone frequencies are not exact multiples of the fundamental frequency\textsuperscript{20}. The mid-IR and NIR absorption spectra measured on a mixture of ketones (\textbf{paper I}) are shown in \textbf{figure 5}. It can be seen that the fundamental and overtone signals of carbonyl are spread out over the spectrum

\begin{center}
\includegraphics[width=\textwidth]{figure5.png}
\end{center}

\textbf{Figure 5:} Combined mid-IR and NIR absorbance spectra measured on a mixture of ketones.

The statistical correlation between signals originating from the same analyte is very high. This can be used to assign signals to a specific analyte and to find the overtones of a specific fundamental. This property was exploited in the work presented in \textbf{paper I}. Further details can be found in the applications section of this thesis.

\textbf{Nuclear Magnetic Resonance spectroscopy}

Nuclear magnetic resonance (NMR) spectroscopy is a very versatile analytical tool that is applied mostly to structure elucidations and in conformational studies of organic small molecules\textsuperscript{21} and macromolecules\textsuperscript{22}.  

20
In this thesis (papers II–V), 1D $^1$H-NMR has been applied to biological samples in order to study the abundances of small organic molecules (metabolomics).

**One-dimensional $^1$H-NMR**

1D $^1$H-NMR is performed by placing the sample in a strong homogeneous magnetic field. This creates an energy difference between the two spin-states (up and down) of the $^1$H nuclei in the sample. A spectrum is obtained by exciting the nuclei to the higher energy state using RF radiation, measuring the free induction decay (FID) as the energy is released through relaxation, and then moving from the time domain to the frequency domain using the Fourier transform (FT). To assure that the magnetic field is homogeneous over the entire sample volume, *shim coils* are used to compensate for small inhomogeneities. The frequency of the spectrum is linearly transformed by subtracting the frequency of a reference compound, commonly *tetramethylsilane (TMS)*, *4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS)*, or *trimethylsilylpropionate (TSP)*. The shifted frequency is then divided by the operating frequency of the instrument, commonly 400–1000 MHz, to yield a relative frequency measured in ppm. On this scale, a chemical compound will have resonances essentially independent of the instrument used for the experiment.

**$^1$H-NMR in metabolomics and STOCSY**

One-dimensional $^1$H-NMR is one of the dominant analytical techniques for metabolomics$^4, 23-30$. The main advantages it has over LC-MS are related to experimental reproducibility, although it cannot compete
with LC-MS in terms of the latter’s low limit of detection.

Usually, very little sample pretreatment is applied to biofluid samples; being limited to removal of proteins by precipitation and pH buffering, described in papers II-V. Typically, around 1500 signals can be observed in a high-resolution (600 MHz and above) $^1$H-NMR spectrum measured on a urine sample. Some regions of the spectrum are very crowded and some peaks may be impossible to integrate or even detect because of overlap. Figure 6 shows a typical crowded region in a urine spectrum at around 3.5–4 ppm.

![Figure 6: A portion of a $^1$H-NMR spectrum measured on a urine sample. The sample has been pretreated by protein precipitation and pH buffering.](image)

The assignment of peaks to analytes in crowded spectral regions is a very challenging task. We do however have some useful tools at our disposal. Peaks that originate from the same analyte generally correlate very well with each other. This is used in statistical total correlation spectroscopy (STOCSY$^{31}$), a purely statistical 2D correlation technique (not to be confused with true 2D NMR techniques such as COSY, $J$-
resolved NMR\textsuperscript{32} and HSQC\textsuperscript{33}). STOCSY creates a statistical correlation map between variables, and therefore requires data from several samples. A STOCSY correlation map computed from raw \textsuperscript{1}H-NMR data is shown in figure 7. STOCSY is affected by peak misalignments in the data, since correlations become blurred for misaligned peaks. Another problem is that the baseline regions tend to correlate a lot, making analysis of the correlation map difficult.

![Figure 7](image)

**Figure 7:** STOCSY on corresponding subsections (3.4–4.2 ppm) of raw \textsuperscript{1}H-NMR spectra measured on 21 urine samples. Black indicates high correlation.

When data are aligned and baseline-corrected, the STOCSY map is much easier to interpret, as shown in figure 8. The data used to create the map in figure 8 were created using the generalized fuzzy Hough transform.
(GFHT) alignment and quantification procedure, described in paper V. The correlation map from aligned data is much easier to interpret, the spurious correlations from the baseline are now gone and each pixel in the correlation map corresponds to one pair of peaks, with no blur from peak shifts. The only remaining issue is that in crowded regions of the spectra, peak quantification is rather uncertain, and this can reduce the observed correlation coefficient between peaks from the same analyte.

![Image](image.png)

**Figure 8:** STOCSY on GFHT-aligned ¹H-NMR spectra measured on 21 urine samples. The same spectral region as in figure 7 is shown. Variable numbers are shown instead of ppm values. Black indicates high correlation.

With this correlation map, peaks can be grouped together by analyte. Peaks that originate from the same analyte generally have a Pearson’s correlation coefficient larger than 0.95 when properly aligned and quantified (the correlation coefficient is lower for peaks with very low
S/N). With a list of ppm values for each analyte it is possible to identify the analyte by comparison with a pure-analyte spectrum. Usually, some peaks present in the analyte spectrum are missing from the sample spectra; owing to overlap, low S/N, or misalignments. This STOCSY-based procedure for assigning peaks was used in the work described in paper V to create a database of metabolites.

**Data dimensionality and rank**

Scientific instruments typically output data of zero, one, or two dimensions; depending on the instrument and its mode of operation. Most common are *first-order* data, for which each sample can be stored as a one-dimensional vector. For a multi-sample dataset, the data are stored in a two-dimensional matrix. Typical instrumental techniques that yield first-order data are LC-UV, GC-flame ionization detector (FID), 1D-NMR, IR spectroscopy, chromatography- selected ion monitoring (SIM) and selected reaction monitoring (SRM), and matrix-assisted laser desorption/ionization (MALDI).

*Solving the correspondence problem for first-order data comes down to deciding which peaks correspond to which over the samples dimension (alignment) and resolving the peak intensities in crowded regions (curve resolution).*

Some analytical techniques yield *second-order* data. These techniques yield one two-dimensional matrix per sample and thus one three-dimensional data tensor per multi-sample dataset, also known as three-way data. Techniques that yield this kind of data include chromatography-MS in
fullscan mode, chromatography–diode array arrangements, and fluorescence spectroscopy. Sometimes it can also be useful to rearrange first order data into three-dimensional form; for example, in vibrational overtone combination spectroscopy (VOCSY), described in the applications section of this thesis.

The focus of this thesis is on the solution of the correspondence problem for first-order data, but many of the principles used can be extended to second-order data.

**Peak detection**

Peak detection (also called *peak picking*\textsuperscript{34, 35}) is a difficult but necessary step in the automated analysis of analytical chemistry data. There is no universal definition of a peak nor what mathematical features it should have. Sometimes, the words *peak* and *signal* are used interchangeably in analytical chemistry.

The objective of peak detection is to go from continuous data such as spectra or chromatograms to a discrete representation where each peak is characterized by a position and intensity (or just one of the two in some cases). Not much information regarding peak detection for NMR metabolomics data could be found in the literature; therefore new methods were developed for peak detection, as described in papers II–V. In the work presented in paper I, no peak detection was used, instead the continuous spectra were analyzed directly.
Zero-area filters

In papers III-V, zero-area filters\textsuperscript{36} based on peak-shapes derived from the spectral data itself are used for peak-detection. Zero-area filters are very selective, robust in the presence of noise, and can even detect shoulder peaks. Filter-based methods have one drawback: they fail in cases where the spectral peak width deviates too much from the peak width used to create the filter. This can be solved, for example, by using \textit{filter bank} \textsuperscript{37} comprising several filters of varying peak widths.

Figure 9: Zero-area filter (top right) based on the TMS peak (top left). The spectrum (middle) is filtered, generating the convolved spectrum (bottom). Detected peaks are marked (o) and peaks that were detected but then removed in the curve-fitting validation step are marked (x).
Figure 9 describes the procedure used in papers III–V, a single zero-area filter based on the TMS peak was used for the detection. Each local maximum in the filtered spectrum is a candidate peak.

**Curve fitting**

After candidate peaks have been identified, each candidate is subject to a curve-fitting test as follows. A Lorentzian curve is fit to the data around the detected peak. The curve is specified by peak width and peak height, which are optimized using the simplex algorithm. If the optimization fails to converge, the peak is discarded. The entire peak-detection algorithm is described below.

**Algorithm 1: Peak detection using a zero-area filter**

1. Calculate the standard deviation of the noise in an empty part of the spectrum.

2. Take a 0.06-ppm–wide portion of the spectrum centered on the internal standard (TSP/DSS) peak at 0 ppm. Process this with a Savitzky-Golay second derivative operation with reversed sign and adjust the resulting filter to sum to zero by multiplying the positive part of it by an appropriate factor, resulting in $g_{\text{norm}}$.

3. Convolve the entire spectrum ($z$) with $g_{\text{norm}}$.

$$z_c(n) = \sum_{k=-\infty}^{\infty} z(n-k) \cdot g_{\text{norm}}(k)$$
4. Identify all local maxima in the convolved spectrum \((z_c)\) and record their position and sample number in a list of candidate peaks.

5. For each candidate peak, fit a Lorentzian-plus-linear-baseline model to the raw data \((z)\), in a window around its position, using least squares. The objective function to be minimized in this peak-fitting step is the residual \(e\):

\[
e(k, m, b) = \sum_{x=x_0-i}^{x_0+i} \left( y(x) - \left( k(x - x_0) + m + \frac{ab}{(x - x_0)^2 + a^2} \right) \right)^2
\]

where \(a\) is the peak width, derived from the TSP/DSS peak and held constant throughout the spectrum; \(x_0\) is the peak mode (the position of the maximum); \(y(x)\) is the intensity; \(2i+1\) is the width of the local segment in data points; and \(m\) is a constant baseline component. Discard peaks with S/N < 3.

6. The position of each undiscarded candidate peak is stored in a list, together with the maximum value of the fitted Lorentzian intensity, and a reference to the spectrum in which it was found. This yields a list of dimensions \(3 \times \) number of detected peaks.

This peak-detection algorithm is an integral part of the GFHT alignment procedure described later in this thesis. The most common source of error encountered so far with this procedure is the peak detection. Erroneous peak detection is a fatal error since once the data have been
reduced to a list of peaks there is no way for the software to recover the lost information about the peak. The peak-detection algorithm described above is not universal; there are $^1$H-NMR datasets where it does not yield satisfactory results, usually as a result of shim problems or incorrect phasing of the spectra.

**Sample, system, and analyte properties**

When analyzing data from chemical systems, it is important to understand the origin of the data. In this chapter, important features of analytical chemistry data will be discussed. In particular, features of the analytical systems and of the samples that can cause peak shifts will be described.

**Instrument drift**

In gas- and liquid chromatography, a major source of peak shift and peak-shape distortions is *instrument drift*\textsuperscript{2, 3}. That is, the signals arising from specific analytes shift continuously over time. This phenomenon can largely be attributed to degradation and pollution of the column, the separation efficiency in the column decreases as the column degrades and the retention mechanism can be altered due to irreversible bonding to matrix components. This effect is more pronounced for analytes that spend more time in the stationary phase of the column, since it is that phase that degrades over time. When the column is replaced or cut, the retention times of all analytes change drastically. Further, differences between individual columns may cause the retention times between resulting chromatograms to differ even before any degradation has taken place.
In figure 10, a portion of 155 GC-MS TIC chromatograms from an indoor air study (not published) are shown as a heat map. The chromatograms are sorted by run order, with the first GC run at the top of the figure. It is clear that there are peak shifts in the data and that the retention time suffers from continuous drift, discontinuous jumps, and sample-by-sample variation. This is most easily realized by looking at the behavior of the strong peak eluting at ~1000 s.

Figure 10: A portion of a heat map of 155 GC/MS TIC chromatograms from an indoor air study. Darker shades of gray indicate higher signal intensity.
These variations in retention time are not that difficult to compensate for when only a few specific peaks in the chromatograms need to be identified. Such peaks can usually be identified from experience and the use of supporting information from the detector (for example, fragmentation patterns from electron-ionization mass spectrometry). The situation is much more complicated for comprehensive studies where we want to establish correspondence for all the signals in the chromatograms; the manual approach is usually not feasible owing to time constraints and the sheer difficulty of some of the assignments.

**The Matrix effect**

In addition to instrument drift, the properties of the sample itself can cause peaks to shift, both in chromatography and in NMR spectroscopy. In the case of $^1$H-NMR, the pH of the sample has a large impact on where on the ppm-axis peaks appear\textsuperscript{38}. The impact of pH varies greatly from peak to peak because of the various degrees of interaction of the analytes with solvent protons/deuterons, depending on the chemical environment of the analyte proton. The magnitude of the effect is correlated with the location of the peak on the ppm axis, since protons with low chemical shift tend to be of aliphatic character and thus do not interact that much with the solvent. However, at higher chemical shift, protic and aprotic protons result in nearby signals in the NMR spectrum, which can cause peaks to switch positions when the pH of the sample is changed. This situation is illustrated in figure 11, which shows a small section of 376 $^1$H-NMR spectra (600 MHz) of human blood plasma. After sorting the spectra on the position of the largest peak within the ppm window it becomes clear that the pattern of
peak shifts is similar across samples but with varying magnitude, from no shift at all, to a shift of ~0.3 ppm between samples. Shifts of that magnitude severely alter the spectrum and make multivariate statistical comparison of different spectra very difficult. Since some peaks switch places between samples, methods that rely on piecewise stretching or compression of the ppm axis (warping) will never work in these situations.

![Figure 11](image)

**Figure 11:** A small section of the ppm axis of 376 ¹H-NMR spectra (600 MHz) from QC samples of human blood plasma is shown. The left pane shows the spectra in run order and the right pane shows them sorted on the position of the largest peak.

Shifts due to the matrix effect are not unique to NMR spectroscopy; chromatography suffers from similar problems. Ideally, chromatographic systems have negligible molecular interaction between different analytes, but with increasingly concentrated samples and fewer purification steps
before analysis, non-ideal behavior will be more pronounced. In untargeted analysis, where little or no sample workup is performed, it is more likely that analytes will interact significantly not only with the stationary phase in the chromatographic system, but also with other analytes. This can lead to peak shifts that are not due to instrument drift, but rather can be attributed to characteristics of the sample itself. It is not a trivial task to determine whether peak shifts are due to instrument drift or due to the matrix effect, the user will only observe that peaks shift and how much. An educated guess can be made in some cases, e.g. if diluted samples have higher retention times in a chromatographic system, the shift in retention time is probably due to the matrix effect. If there are continuous systematic shifts in retention-time over time, it is probably mainly due to instrument drift.

**Random events**

In addition to peak shifts caused by the instrument and the samples, there are also effects that are seemingly random. These include phenomena such as slight variations in the injection and sample-pretreatment procedures, small temperature variations, instrument noise, etc. None of these effects are truly random; they just do not follow any simple pattern. Peak shifts due to these effects are usually smaller than those from instrument drift and matrix effects; except for cases where a major random event causes the analytical system to fail completely. Such events are best handled by redoing the analysis.

*Peak shifts are probably due to a mix of instrument drift, the matrix effect, and random events.*
Defining the correspondence problem

I have now established a couple of partial explanations for peak shifts in chromatography and $^1$H-NMR spectroscopy. I will expand these concepts to include any number of unknown sources of peak shifts, in data from any analytical technique that yields first-order data, such as 1D-spectra and chromatograms. This process is described in paper III using slightly different terminology. I will here simplify the equations and generalize them to describe peak shifts in any kind of first-order data.

**Hypothesis:** *All the peak shifts in a dataset can be explained by a linear model with only a few terms.*

Assume that the instrument/sample-system has a finite number of one-dimensional properties (e.g., pH, column length) that affect the positions of peaks in the resulting output data. Also assume that the position of each peak in the resulting output data is linearly dependent on each of these properties, with varying sensitivities. If these assumptions can be shown to hold, the hypothesis above is reinforced.

The hypothesis can also be formulated with matrix algebra. The matrix containing the positions of all peaks ($S$) in a dataset can be decomposed to instrument/sample properties ($M$) and peak properties ($A$) using PCA with only a small number of principal components:

$$ S = MA \quad (1) $$

At this point we need a dataset with correctly aligned peaks to assess whether the hypothesis holds or not. In papers II–V, the development of the GFHT alignment method, which is based on the hypothesis above, was
described. The method worked for the alignment of $^1$H-NMR data; and that the model was able to describe reality gives weight to the above hypothesis. In **Figure 12**, the relationship described in **equation 1** is shown in more detail.

![Figure 12: The relationship between M, A and S.](image)

$\begin{align*}
M \text{ Analyte properties NxQ} & = A \\
P \times N & = S \\
\text{Peak shifts PxQ}
\end{align*}$

**Figure 12:** The relationship between $M$, $A$ and $S$. $M$ is a matrix that contains the properties of the instrument/samples that give rise to peak shifts. $A$ is a matrix that contains the sensitivity of the peaks (analytes) to the properties in $M$. $S$ is the resulting matrix and contains the peak shifts observed in the dataset. Of these matrices, only $S$ can be observed directly from the data. $N$ is a scalar that describes how many relevant shift phenomena there are in the dataset. $P$ is the number of samples and $Q$ is the number of peaks per sample.

If we can find $M$ and $A$ for all pairs of samples (present and future) and all peaks in our dataset, we have solved the correspondence problem for that particular dataset under the hypothesis. The positions of the peaks can then be calculated using **equation 1**. A feature of this decomposition is that $A$ is constant for all systems with the same analytes and the same set of relevant sample/instrument properties. Once $A$ has been calculated for one dataset, we only need to find $M$ for future similar datasets in order to
align all peaks described by $A$.

The following sections will outline a method for decomposing $S$ into $M$ and $A$; an initial automated assignment of a few easily identifiable peaks is followed by expansion of $A$ to include as many of the dataset’s peaks as possible.

**Establishing peak-shift models**

**The Hough transform**

The concept of using empirical linear models to describe peak shifts in $^1$H-NMR spectroscopy was first described in paper II. The concept was discovered when a large number of spectra of human plasma samples were visualized as a heat map. The peak shifts seemed more-or-less random until the spectra were sorted on the position of one peak as it then became obvious that the positions of the other peaks were shifted in a similar pattern, *figure 11*. The initial impression was that it had to be possible to use image analysis to find the patterns.

An important tool for finding parameterized shapes in images is the generalized Hough transform$^{39}$. It can take any shape that can be described by a finite number of parameters, and transforms the image into a space that is spanned by these parameters. As an extension, the generalized *fuzzy* Hough transform (GFHT)$^{40-42}$ also allows for the shape to deviate somewhat from the parametric form by using a fuzzy parameter, $\sigma$. These methods are based upon the original Hough transform$^{43}$ which is used for finding lines in images, *figure 13*. 
In paper III we introduced the term *Hough indicator tensor* (HIT) to describe the multidimensional data structure that results from the Hough transform. It is also called the *accumulator array*, or simply the *Hough space*. The HIT will have local maxima at coordinates that correspond to the parameter values that describe the shapes found in the image.

![Diagram](image)

**Figure 13:** The original version of the Hough transform, a line is described using offset and angle as parameters. The image to the left contains two lines that yield two local maxima in the two-dimensional Hough indicator tensor (HIT) that is shown to the right. The locations of the maxima give the angle and offset for both lines.
Figure 14: The GFHT applied to a $^1$H-NMR metabolomic dataset. The raw data (shown as a heat map in the bottom pane) are transformed via peak-detected data (shown in the middle pane) to the HIT (shown as a heat map in the top pane). This particular example uses only two parameters for the peak-shift model, which leads to a two-dimensional Hough space. The black lines indicate alignment suggestions produced by the GFHT alignment algorithm.

The GFHT method for establishing shift models

The GFHT using equation 1 as the shape parameterization model (plus a constant term $k$) is given in equations 2–4. For this transform to be useful, we must define the sample properties $M$. The approach described in papers II–V was to assign the positions of around 10 easily identifiable
peaks for every sample in the dataset. They may be easy to identify because they appear in empty regions of the spectra or because they are much larger than the surrounding peaks. The positions of these peaks were then analyzed using PCA. The first few score vectors from the PCA were used as $M$. The first implementation of the GFHT alignment algorithm, described in papers II–III, was very computationally demanding. This restricted the number columns of $M$, to a maximum of 3 or 4, corresponding to 3 to 4 unique sample properties. This problem was later solved by a more efficient algorithm, see algorithm 2 below.

$$h_{k,f,m,...} = f(a, k)$$
$$f(a, k) = \sum_{i} \sum_{j} x_{ij} \exp \left[ -\frac{1}{2} \left( \frac{j - k - M_{ij}a}{\sigma} \right)^2 \right] \quad (2-4)$$

$$a = [(a_1)_1, (a_2)_m, \ldots]$$

When applying this transform to peak-detected data, the resulting HIT will contain local maxima at points corresponding to each peak, with consistency over the entire dataset, figure 14 top pane. The consistency does not have to be 100%, lower consistency yields a lower score of the local maximum. The shapes corresponding to these local maxima are shown in figure 14, middle and lower panes.

In later work (papers IV and V) it became clear that calculating the entire HIT is neither necessary nor desirable; the latter because of the very long computational times (days, sometimes weeks) involved. An alternative method using one of the spectra as an anchor spectrum was developed. The sample properties in $M$ are transformed by subtracting the
property values of the anchor spectrum (similar to mean-centering). This forces the parameterized shapes to have constant values in the anchor spectrum, regardless of the values of the peak properties $A$. Then for each detected peak in the anchor spectrum, a numerical search is performed to align that particular peak. The complete algorithm was presented in paper V and is described below. The transform was also simplified slightly to save computational time, equations 5–6. A side-effect of this simplification was that the resolving power of the algorithm improved in crowded regions of the spectra. This is due to the fact that only the closest peak is used when calculating the score in equation 6, any other peaks within the fuzzy region are disregarded.

$$\mathbf{s} = \mathbf{M}\mathbf{a} + k \quad (5)$$

$$h = \frac{\sum_{i=1}^{N} \exp\left(-\left(\frac{s_i - z_i}{\sigma}\right)^2\right)}{N} \quad (6)$$

Equation 6 is the core of the current GFHT alignment algorithm as described in paper V, the terms in the equation are as follows: $z_i$ is the location of the detected peak closest to $s_i$ in spectrum $i$ (both measured in data points), $N$ is the number of spectra (including the anchor spectrum), $h$ is the target function to be maximized by the algorithm, $\mathbf{M}$ is the shift model matrix ($N \times r$, where $r$ is the number of selected model peaks), $k$ is the position of the peak in the anchor spectrum, and $\mathbf{a}$ ($r \times 1$) is the set of independent peak properties. $z$ is found by a nearest-neighbor search in the peak list from the peak-detection algorithm. $N$ is the number of spectra in
the dataset. In practice, the value of $h$ will be close to $1/N$ if the peak cannot be aligned using this method (usually because it is a false peak, has low consistency, or overlaps too much with other peaks) and close to 1 if a perfect match is found. For the trivial case of a peak with 100% consistency that does not shift, $h$ will be close to 1 with all values of $\alpha$ being zero.

**GFHT alignment algorithm**

The objective of the GFHT alignment algorithm is to maximize $h$ in equation 6 for each peak in the anchor spectrum. The value of $h$ is bounded between 0 and 1, where $h = 1$ corresponds to a perfect match in all spectra.

**Algorithm 2. GFHT alignment**

<table>
<thead>
<tr>
<th>Definitions:</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>X</strong> – The raw data matrix ($N \times m$), where $N$ is the number of spectra and $m$ is the number of data points per spectrum.</td>
</tr>
<tr>
<td><strong>Y</strong> – The aligned data matrix ($N \times p$), where $p$ is the number of successfully aligned peaks.</td>
</tr>
<tr>
<td><strong>S</strong> – The locations (in data points) on the ppm axis for each peak in each sample ($N \times p$).</td>
</tr>
<tr>
<td><strong>MAX_MATCH</strong> – The maximum number of data points of “slack” the algorithm has when assigning a peak.</td>
</tr>
<tr>
<td><strong>H_CUTOFF</strong> – The minimum value of $h$ that is required to regard an alignment as successful.</td>
</tr>
</tbody>
</table>
σ – The GFHT fuzzy parameter.

Algorithm:

1. Define MAX.MATCH, H_CUTOFF and σ. The values we used are found in the results section. We recommend values of 0–3 data points for MAX.MATCH, a value of 0.5–0.6 for H_CUTOFF and a value of 1–5 data points for σ.

2. Append the anchor spectrum to the dataset as a new row in X.

3. Peak-detect all spectra in X.

4. Select a few (5–10) easily assignable peaks in X as model peaks. This is most easily done by finding local maxima in small segments of the spectra where one peak is known to dominate. Arrange the positions (in data points) of each selected peak as a column in the shift model matrix M. Subtract the position of the peak in the reference spectrum for each column from all values in that column (every column will have a zero in the row of the reference spectrum).

5. For every peak in the reference database, perform the following steps:
   a) In the peak-detected reference spectrum, find the peak corresponding to the entry in the database. Set k equal to its position (in data points).
   b) Generate an initial guess for each element of α. The default is to use all zeros, if the method fails to converge from this initial guess; a random initial guess is drawn from a standard normal distribution instead.
   c) Maximize the target function \( h \) in equation 6 as a function of \( \alpha \) using
simplex optimization (or gradient descent). Repeat steps 5b and 5c several times with higher values for the random initial guess if the optimization does not converge.

d) Adjust $s$ slightly to match the closest peak in each spectrum if it is within MAX_MATCH data points. Do not recalculate $h$ in this step.

e) If $h > H_{\text{CUTOFF}}$ add $s$ as a new column in the peak location matrix $S$, otherwise disregard it.

6. For every entry $(i, j)$ in $S$, perform the following steps:

a) Calculate a baseline value, $b$, equal to the minimum value in row $i$ of $X$ between the peaks immediately bracketing $S(i, j)$.

b) Assign an intensity for the peak in the aligned data matrix:

$$Y(i, j) = X(i, S(i, j)) - b.$$ 

Using shift models to establish correspondence

The GFHT alignment algorithm described above yields an aligned matrix $Y$ where each column represents one unique peak with correspondence over the samples dimension and each row represents one sample in the dataset. Each value in the matrix represents the intensity of a peak. The current algorithm yields somewhere in between 800 to 1100 aligned peaks in a 1D $^1$H-NMR urine dataset. Another output from the algorithm, $S$, contains the corresponding ppm values for all of the peaks in $Y$.

Principal component analysis of $S$ reveals information about the structure of the peak shifts in the dataset. The hypothesis that the shifts
follow a simple linear model is reinforced if the following two statements hold:

A. The alignment results are correct.

B. The peak positions S can be decomposed, modeling a large portion of the variance, using PCA with only a few principal components.

Statement B is needed because the GFHT alignment algorithm allows some slack in the model, a portion of each peak shift is allowed to be random (as defined by the MAX_MATCH parameter setting). Because of the slack, the peak locations do not automatically follow a linear model.

Statement A was verified in the work described in papers III and V for several different approaches. Paper III describes the modeling of metabolite concentrations using PLS on aligned data. Successive removal of variables showed that even the smallest aligned peaks in the dataset contained relevant concentration information. The method was also compared to manual alignment using Chenomx NMRSuite44, as described in paper V.

Statement B is verified in figure 15 below. The figure shows the variance explained by PCA models of the shifts of several peaks from the same dataset, dataset B in paper V. The shifts are explained well by a model with few PCs, even when all 965 peaks are included.
Figure 15: Variance explained by PCA of the shift matrix $S$ from dataset B in paper V (rat urine $^1$H-NMR data). The 10 peaks are the model peaks used to model the shifts, the 93 peaks are the peaks used for targeted quantification, and the 965 peaks are all the peaks that were aligned in the dataset.

The results from papers II-V strengthens the hypothesis that a simple model of peak shifts derived from only a few peaks works on a global scale over the entire dataset.

It is apparent that the choice of model peaks included in $M$ is important, they have to span the sample properties that give rise to peak shifts. One way to check this is to perform PCA on $M$. There are two criteria that should both be fulfilled when adding a new peak to $M$:

1. The total variance in $M$ should increase as much as possible; this corresponds to choosing peaks that shift a lot. Larger shifts are probably more likely to be relevant and less likely to be
random.

2. The newly added peak should have low correlation with the already selected peaks; this ensures that the domain is spanned, by choosing peaks with shift patterns that are close to being orthogonal.

These criteria ensure that there are as many different relevant phenomena represented in $M$ as possible. Of course it is also important to ensure that the selected peak actually has correct correspondence over the dataset.

Since the properties of the peaks $A$ are independent of the samples, it can be reused for future datasets. The requirement is that the set of peaks used to form $M$ are not changed and that the same instrument/sample properties are relevant for both datasets. To predict the peak shifts of the set of peaks in $A$ for future samples, the outer product of $M$ and $A$ is calculated:

$$S_{new} = M_{rev}A$$  \hspace{1cm} (7)

In practice, there are also some minor random variations in the shifts that are not spanned by the model. The best way to handle this seems to be to use the columns in $A$ from a previous dataset as initial guesses for $\alpha$ when aligning a new dataset using the GFHT alignment algorithm described above. This ensures much shorter computational times and higher probability of convergence than using random initial guesses; because the guesses will be nearly optimal.
Quantification of signals

The implementation of the GFHT alignment algorithm described in papers IV and V allows the user to select a subset of peaks to align, if so desired, by using only a few peaks in the anchor spectrum. This is useful for targeted quantification of metabolites because peaks originating from a specific analyte can be aligned and grouped separately. Each hydrogen-containing compound yields several signals in $^1$H-NMR, except for special cases where all hydrogen atoms in the molecule are chemically equivalent. The ratios between the signal intensities reflect the relative abundance of the hydrogen atoms that give rise to the signals. This peak multiplicity is an advantage when performing quantification of metabolites in complex samples because we can use the fact that the signal intensity ratios are almost constant across samples. An example of this is shown in figure 16. In paper V, the development of methods for peak selection and quantification that exploited the multiplicity of signals was described. The methods are based on PCA of all peaks assigned to a single analyte and are described below.
Peak selection

The purpose of peak selection in targeted analysis is to exclude peaks that do not carry high-quality information. Peaks carrying high-quality information are here defined as peaks that are non-overlapped and correctly annotated and aligned. The method for peak selection presented in paper V is based on the assumption that ratios of peak intensities from a single analyte remain constant between runs. This is equivalent to a PCA model with one PC explaining 100% of the variance of an aligned data
matrix containing only the peaks associated with a particular analyte, $Y_{\text{analyte}}$. The procedure used for quantification is presented in Algorithm 3 and Figure 17 below.

**Algorithm 3. Peak selection using PCA**

1. The spectral intensities at the positions of all the aligned peaks for the metabolite are put into a (samples × peaks) matrix $Y_{\text{analyte}}$.

2. $Y_{\text{analyte}}$ is centered and scaled to unit standard deviation.

3. PCA is performed; the peak with the highest absolute PC2 loading value is removed from $Y_{\text{analyte}}$.

4. The procedure is repeated until the cross-validated explained variance (leave-one-out, projecting the left-out sample onto the PCA loadings calculated from the remaining samples) reaches a pre-determined cut-off (set by the user to 95–99%, depending on the noisiness and complexity of the data).

5. If only two peaks remain and the explained variance has still not reached the cut-off, the peak with the highest relative standard deviation is removed and the remaining peak is used for quantification.
Peaks used for quantification

Peaks not used for quantification

Figure 17: The loadings from a PCA of the aligned data matrix for a single analyte (asparagine). The peaks that cluster along PC1 are deemed to be of high quality and are kept, the rest of the peaks are discarded before quantification.

Results from the peak-selection algorithm for asparagine from paper V are shown in figure 17. Some peaks cluster nicely along PC1 and will be used for quantification. Peaks that do not carry high-quality information do not cluster with the rest and are removed. Each peak is scaled to unit standard deviation before peak selection. There are two main reasons for this:

- The PCA should not be controlled by a single large peak; that
would lead to the risk of letting an erroneous large overlapping peak bias the peak selection algorithm.

- The loadings cluster in a better way when the data are scaled because each peak will have approximately the same distance to the origin in the loadings plot for scaled data. For unscaled data, the distance to the origin for each peak varies with peak intensity.

**Quantification using PCA**

When a subset of the peaks have been selected using algorithm 3 described above, the signals of the selected peaks are used to quantify the analyte. The method presented in paper V is based on PCA and gives relative concentrations of all the samples in the dataset. If absolute concentrations are the goal, the absolute concentration in one sample in the dataset has to be established using a reference method.

The quantification described in paper V is performed using a single-point calibration with the PC1 score as the independent variable. This is a simplistic, special case of *Principal component regression* (PCR).
Algorithm 4. Quantification using selected peaks

1. PCA is performed on the unscaled and uncentered intensity matrix of the selected peaks for the analyte, $Y_{\text{analyte}}$.

2. The PC1 scores, $t_1$, from the PCA are used as relative concentrations for the analyte.

3. If absolute concentrations for a reference sample are available, absolute concentrations of all samples are determined from PC1 scores by single-point calibration.

\[
c = \frac{t_1 c_{\text{ref}}}{t_{1,\text{ref}}} \quad (8)
\]

The reasoning behind using this PCR-like calibration model is that it weights all the collinear peak intensities together, disregarding variance directions other than the major one and giving higher weight to larger peaks (which are assumed to have higher S/N), and that it only requires one reference concentration.

Software

All of the algorithms described above have been implemented in a MATLAB toolbox with a graphical user interface (GUI), developed at Stockholm University. The interface has gone through several transformations over time as new algorithms were developed and old ones were replaced. The current (March 2012) GUI is shown in figure 18.
Having all the algorithms collected in a toolbox with a GUI is very helpful, as the process of testing combinations of methods and evaluating the results becomes very fast.

When the program is started, automated GFHT alignment, peak selection, and quantification are performed. All the analytes with assigned peaks are listed in the GUI and all the samples are available for individual inspection. There is the option to test different user-defined peak selection and quantification methods. There is also the option to realign peaks using...
different parameter settings, which is useful for peaks that the GFHT algorithms failed to align automatically. When the results appear to be satisfactory, the analyte concentrations and the intensities of all unassigned peaks can be exported in either Excel or MATLAB format.

**Modeling of synchronized data**

**Principal component analysis**

Principal component analysis (PCA\textsuperscript{45}) is the most widely used multivariate technique for exploratory data analysis and dates back to the beginning of the 20\textsuperscript{th} century (or possibly even further back). It was introduced by Karl Pearson in 1901 and has since then been applied in probably all fields of quantitative research. PCA is a successive-projections algorithm, where the directions of the loading vectors (the normalized vectors to project the data onto) are constrained to:

- Give the lowest sum of squared residuals after projection (best least squares fit).
- Be orthogonal to all previously calculated loading vectors.

The projections for each row in $X$ onto the loadings are called the scores of the matrix and are used to determine similarities and differences between the rows.

PCA is most easily implemented in scientific computational software such as MATLAB through singular value decomposition of the data matrix to be analyzed, $X$ (of size $N \times M$):

$$X = U \Sigma V^T \quad \text{(9)}$$
U is a $N \times N$ orthonormal matrix, the columns of $U$ are normalized scores for successive PCs. If $U$ is multiplied by $\Sigma$, the scores are obtained. $\Sigma$ is a $N \times M$ rectangular diagonal matrix of singular values where the diagonal entries are sorted by descending magnitude. The singular values correspond to the magnitude of the scores for each successive PC. $V$ is an $M \times M$ orthonormal matrix where the columns correspond to loading vectors for successive PCs.

The application of PCA was described in papers II–V, to both data preprocessing (peak selection, quantification) and to exploratory data analysis (analysis of peak intensities and metabolite quantities).

**Partial least squares**

Partial least squares (PLS, also called *projection to latent structures* \(^{46-49}\)) is a linear regression method introduced by Herman Wold in 1975. It is similar to PCA in that it decomposes collinear data into less-correlated *latent variables* which are then used for regression. The difference between PCA and PLS in terms of projections lies in that the dependent variables $Y$ in the regression are used in the decomposition for PLS. The directions of the components of PLS (called *latent variables*) are constrained to:

- Maximize the covariance between the scores $t$ and $Y$.

After each latent variable is calculated, both $X$ and $Y$ are deflated, that is the part of $X$ and $Y$ that was described by the projection is subtracted. After the deflation, the next latent variable is calculated. There are several algorithms available for PLS. In this thesis, the PLS toolbox for MATLAB was used for all PLS models. The use of PLS as a reference method for
mid-IR and NIR data was described in paper I, and as a tool for validating the alignment of NMR data in paper III.

**Parallel factor analysis**

Parallel factor analysis (PARAFAC\textsuperscript{50}), also known as *canonical polyadic decomposition* (CPD) is a generalization of PCA for application to higher dimensional data structures (tensors). The method was described as early as 1927 by F.L. Hitchcock\textsuperscript{51} but was made popular in chemometrics by Rasmus Bro\textsuperscript{52} in 1997. PARAFAC is not the most general extension of PCA to tensors. It is a constrained *Tucker\textsuperscript{3}51* model (described by F.L. Hitchcock in 1927, named after Ledyard R. Tucker\textsuperscript{53}, 1966) with the restriction that the core tensor of the decomposition has to be diagonal. Tucker3 is not described here, for further information refer to the publications by Hitchcock and Tucker.

PARAFAC decomposes the multi-way data tensor \( \mathbf{X} \) into one \( N_i \times f \) loading matrix for each mode of the tensor. \( N_i \) is the number of data points along mode \( i \), and \( f \) is the number of factors (similar to PCs in PCA) in the PARAFAC decomposition. For example, a \( 71 \times 81 \times 3 \) tensor of mid-IR and NIR absorbances was decomposed using a three-factor PARAFAC model to three matrices of sizes \( 71 \times 3 \), \( 81 \times 3 \), and \( 3 \times 3 \); see paper I.

In PARAFAC, all the factors are determined simultaneously by minimizing the sum of squared residuals for the entire decomposition, shown here in element-wise form:

\[
x_{ijk...} = \sum_{i=1}^{f} a_{ij} b_{jf} c_{kf} ... + \epsilon_{ijk...} \quad (10)
\]
Where \( A, B, C \ldots \) are the loadings for the different modes of the tensor and \( E \) are the residuals whose sum of squares is to be minimized. There are several algorithms available for calculating the PARAFAC decomposition. In this thesis, the N-way toolbox\(^{54}\) for MATLAB by Rasmus Bro et al. was used for all applications. The use of PARAFAC for modeling multi-way mid-IR and NIR data is described in paper I and for modeling \(^1\)H-NMR metabolomic time-series data in paper IV.

Applications

**Infrared and near-infrared spectroscopy**

Vibrational overtone combination spectroscopy (VOCSY), a method for rearranging mid-IR and NIR data into a three-way tensor that can be used for modeling with PARAFAC, was described in paper I. The main obstacle in developing VOCSY was an alignment problem, how do we match the overtones with the correct fundamentals? This problem was solved using statistical total correlation spectroscopy (STOCSY). The procedure used was to create a correlation map between the fundamental region of the data and each overtone region, figure 19. The overtone correlation signals fall approximately on a diagonal line in the map. The equation describing this line is the warping transform from the fundamental region to the overtone region. Inverting it yields the transform from overtone to fundamental, which is used to resample the overtone data to create overtone slabs with the same number of data points. The overtone slabs are then combined with the fundamental to form the third mode of the VOCSY tensor. The tensor is then modeled using a variety of multi-way techniques such as PARAFAC. The method should in theory have
advantages over two-way multivariate techniques such as PLS because there are constraints on the analyte signals in that they have to match in all slabs. Additional details can be found in paper I.

Figure 19: A STOCSY map of the fundamental carbonyl region (horizontal axis) and the first-overtone carbonyl region (vertical axis). The straight line represents the signal matching.

The VOCSY method was shown to work in a very simple case that used standard solutions created in the laboratory. Some attempts were made to model NIR data from more complex food samples, and the results looked somewhat promising at the time. My work on that project was however discontinued because my focus shifted towards NMR data and metabolomics (papers II–V). For further details regarding VOCSY,
including results, see paper I.

**Nuclear magnetic resonance**

**Datasets**

Papers II–V describe the development of new tools for peak alignment and for modeling of $^1$H-NMR metabolomics data. The datasets used for demonstrating the methods are:

- One large dataset (188 samples in duplicate) of human blood plasma samples. These are QC samples created by AstraZeneca.
- One large dataset (336 samples) of rat urine samples. These are from a toxicology study of L-ethionine at AstraZeneca.
- One small dataset (24 samples) of synthetic mixtures designed to mimic plant extracts. This dataset is used with permission from Dr. Ian Lewis$^{55}$.
- One small dataset (20 samples) of rat urine samples. These are from a tetracycline toxicology study conducted at AstraZeneca.

Out of these datasets, the synthetic mixtures have the lowest complexity with ~350 peaks per sample. The plasma dataset is of intermediate complexity with ~700 peaks per sample, and the two urine datasets are of high complexity, with ~1100 peaks per sample. The peak shifts in the urine data are also more complex than for the other data sets, making them more challenging for the GFHT alignment procedure.
Results

In this section a brief overview of the main results in papers II–V will be given. Some additional results and discussion combining the results from the older papers with the methods from paper V will also be presented.

The main result of paper II is that one peak with a chemical shift that depends on the pH of a sample can be used to linearly model the shifts of other pH-sensitive peaks in plasma samples. This model works because pH is the dominating contributor to the peak shifts in plasma samples. The paper also introduces the GFHT method for determining the model parameters. If we look carefully at figures 2 and 3 in paper II we can see that other minor phenomena also contribute to the overall shift. By using the updated version of the GFHT alignment algorithm presented in paper V on this dataset and performing PCA on the peak-location matrix, it becomes apparent that: (a) The vast majority of the shifts are explained by a one-PC model, and (b) that the second component seems to be a compensation for a small non-linearity with respect to the first component, figure 20. The variance explained by one PC is 99.96%, which is even higher than the variance explained by a one-PC model of the six model peaks used for the alignment (99.49%).
From this we can conclude that a single property (pH) is responsible for the vast majority of the shifts in human plasma $^1$H-NMR datasets and that the shifts deviate only slightly from a linear model.

In paper III it was established that the GFHT method works for urine data. PLS calibration models using only low intensity peaks for determining metabolite concentrations were shown to work on the synthetic mixtures dataset. It was also shown that PCA models of the large rat urine dataset could be interpreted in terms of the study design even after all peaks larger than 0.005% of the maximum spectral intensity were removed from the dataset. It was also suggested that the study groups
could be scored in the GFHT separately (using the maximum score among the groups) to avoid the low consistency for peaks that only appear after dosing (or peaks that disappear after dosing). Note that even though the parts of the dataset are scored separately, the peak should still be searched for within MAX_MATCH distance in all spectra, to avoid bias.

In paper IV, no new alignment or quantification technologies were introduced. The focus was instead on modeling aligned time-series data. The large rat urine dataset consists of urine samples from 35 animals sampled at up to 16 different time points. It was shown that by arranging all peaks with significant regulation (determined by Student’s \( t \)-test) into an \((\text{animals} \times \text{peaks} \times \text{days})\) tensor, it was possible to build a PARAFAC model that described different time-dependent phenomena. This is useful for studying time-series data without assuming any specific behavior, as detailed information about the regulation over time can still be obtained. With the annotations made in paper V, some of the analytes that regulate significantly according to the different time-profiles can now be identified: alanine, glucose, methylsuccinate, \( N,N \)-dimethylglycine, and 3-indoxylsulphate seem to behave according to the slowest of the three positive fold-change profiles shown in figure 21.
Some metabolites also follow the modeled negative fold-change time-profile (figure 10 in paper IV): succinate, 2-oxoglutarate, and citrate. None of the identified metabolites follow the two fast, positive fold-change time-profiles shown in figure 21. This might be an indication that those profiles describe drug metabolite excretions but it is also possible that they describe endogenous metabolites that were not identified in paper V.

The main result from paper V is a fully automated quantification procedure for $^1$H-NMR data. Concentrations of metabolites in the synthetic mixtures dataset were determined with a relative RMSEP value of 6.4%. The method was also shown to be in agreement with Chenomx NMRSuite manual spectral profiling for most metabolites in the small rat urine dataset. The validation for the large rat urine dataset was only indirect; the method was shown to give results that were relevant to the clinical question at hand. No reference method was applied, because of the time-consuming nature of the manual spectral profiling. Using Chenomx, only
around 20 spectra per day could be processed and many of the assignments were ambiguous to say the least.

**Conclusions and outlook**

Assignment of correct correspondence of analytical chemical data prior to statistical analysis is crucial for the quality of the obtained results. For typical targeted analysis with only a few analytes, the problem is usually solved manually. For *omics* studies with many variables it is not feasible to synchronize the entire data by hand.

The methods for modeling the shifts of peaks presented and discussed in this thesis provide a general solution that can be applied to any type of multivariate dataset where the following prerequisites are met:

1. The signal shifts have an underlying structure that has low rank.

2. There exists a working automated peak-detection algorithm for that type of data.

3. There are a few easily identifiable peaks that are present in all samples. (This can be solved by adding internal standards if no endogenic peaks are available).

4. The number of samples available has to exceed the rank of the shifts enough to avoid overfitting the model.

Prerequisite 1 holds for all chromatographic and spectroscopic data I have examined thus far, but it is of course possible to imagine cases where it does not hold. Prerequisite 2 is trickier to fulfill as there is no universal peak-detection algorithm available that works satisfactorily for all kinds of
data. Prerequisite 3 can be met by adding a few internal standards to the samples, preferably with varying chemical properties in order to span the domain of the shifts. Prerequisite 4 holds for most studies, around 20 samples is usually enough for successful alignment of a typical $^1$H-NMR dataset; in paper V, two datasets with 20 and 24 samples respectively were aligned.

Future work in this field could be concentrated on a more general framework for peak-detection. This has been the bottleneck when working with the GFHT alignment procedure; the peak detection method usually has to be tweaked in order to accommodate different S/N ratios, different peak shapes, different spectral/chromatographic resolutions, and different numbers of data points. I also have an idea that working with normalized raw spectra instead of peak-detected spectra in the GFHT might improve the alignment method and remove the need for peak detection. It is also possible that using the peak intensities and correlations with other peak intensities from the same analyte when scoring the GFHT alignment could also lead to improvements.

Another area with room for improvement is the optimization algorithm for the GFHT. The algorithm used in paper V ($fminsearch$ in MATLAB) sometimes converges to a local maximum—instead of the global maximum—for peaks with large shifts. One work-around that has been tried is to generate many starting guesses for separate optimizations and use the highest score value. The problem with that approach is that the computational workload increases significantly. A more efficient optimization algorithm would be desirable. Simulated annealing seems
like a good candidate but has not yet been evaluated.
References


(11) Tomasi, G.; van den Berg, F.; Dept. of Food Science, The Royal Veterinary and Agricultural University, Denmark, 2006.


(45) Pearson, K. *Philos. Mag.* **1901**, *559-572.*


(52) Bro, R. *Chemometrics and Intelligent Laboratory Systems* **1997**, *149-171.*


(54) [http://models.kvl.dk Andersson, C. A.; Bro, R., 2003.](http://models.kvl.dk)

Acknowledgements

I would like to acknowledge the Department of Analytical Chemistry at Stockholm University for spending the time and resources required for me to create this Thesis. I would also like to acknowledge AstraZeneca for co-financing my work and for allowing me to publish papers based on their data.

I also have a lot of people that I would like to thank individually, sorry if I forgot anyone.

I would like to thank Ralf Torgrip for being my supervisor and friend, for giving me a lot of support, for being a constant source of new (and sometimes revolutionary) ideas and for being a great guy.

I would like to thank Magnus Åberg for also being my supervisor, for being very thorough when writing and reviewing papers and for long discussions over coffee about different ideas and concepts. Extra thanks for the great times we had in North America in 2011.

I would like to thank Erik Tengstrand for taking conceptual discussions many steps further than most people would, and for having very, very strong shoulders.

I would like to thank Malin Kölhed for being part of the chemometrics course teaching crew one year and for lending me your IR instrument for paper I.

I would like to thank Bo Karlberg, Jenny Forshed and Ludvig Moberg for over and over delivering quality guest lectures. I would also like to thank Bo and Jenny for rewarding scientific collaborations.
I would like to think the great Danes; Rasmus Bro and Søren Engelsen for helping out with paper I and of course for the PARAFAC algorithms.

I would like to thank the great people at AstraZeneca R&D Södertälje, which is now unfortunately being closed down. Special thanks to Johan Lindberg, Ingela Gustafsson and Erik Wahlström.

I would like to thank Ian Lewis for allowing me to use his synthetic NMR mixture dataset and for writing great papers.

I would like to thank the administrative personnel at the Department of Analytical Chemistry; Anita, Ann-Marie, Lena and Jonas for making the place actually work.

I would like to thank my fellow PhD students and former fellow PhD students for making the department a fun place to be in. Special thanks to the party crew: Caroline Bergh, Helena Hansson, Axel Mie, Leila Zamani, Petter Olsson and Giovanna Luongo.

I would like to thank my master thesis students; Tove Slagbrand, Dina Lakayan and Elin Paulsson for being great students with a lot of energy. Special thanks to Tove who helped me with the grueling Chenomx validation for paper V.

I would like to thank the heads of the Department of Analytical Chemistry, Anders Colmsjö and Ulrika Nilsson for being nice people and for running a small but thriving department. I wish you well in the future!

I would like to thank the morning coffee crew at the department, Roger Westerholm, Gunnar Thorsén and Conny Östman for great inspiration and insights in the mornings, plus more than a few (usually bad)
jokes.

I would like to thank the great people at Springer for being very fast in handling papers, reviews and revisions. You can just take a look at the header of paper I to see what I am talking about.

I would like to thank my high-school teachers Anders Sjöström, Anders Johansson and Peter Sand for demonstrating that chemistry and math can be a lot of fun.

I would like to thank my family, Margareta Ross Bergsten, Steve Alm, Jonas Bergsten, Karin Alm and Oskar Alm for being a great family and role models.

I would also like to thank some of my role models among my close relatives; Gunnar Ross and Svante Ross, for setting the standards high.

I would also like to thank all of my great students at the chemometrics course for being a constant source of inspiration and energy for me. Teaching is one of my callings, a deeply satisfying experience.

I would also like to thank my friends, too many to mention them all, but you know who you are. Thanks for making my life a better place to be in!

Finally, thanks to all the people who have given me suggestions and comments on this thesis; Conny, Roger, Ulf, Bo, Magnus, Ioannis, Caroline, Leopold.