Bioinformatics for mass spectrometry. Novel statistical algorithms
Dijkstra, Martijn

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2008

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 4

Advanced deconvolution analysis of complex mass spectra

ABSTRACT

Due to physical and chemical phenomena, a simple sample can give rise to a complex mass spectrum with many more peaks than the number of molecule species present in the sample. We link peaks within and between different spectra, and come up with an advanced analysis approach to produce reliable estimates of the molecule masses and abundances. By linking peaks, (i) we can locate multiple charge peaks at the correct position in the spectrum, (ii) we can deconvolute complex regions with many overlapping peaks by including information from related regions with lower complexity and higher resolution, and (iii) we reduce the total number of observed peaks in a spectrum to a much smaller number of underlying molecular species. This (iv) reduces the statistical test multiplicity for biomarker discovery and therefore increases its power significantly.

4.1 Introduction

A simple sample containing a few molecule species can generate a complex mass spectrum with many peaks. Various chemical and physical phenomena can explain this (Dijkstra et al. 2007). For example, a molecule species can have different forms (isotopes) with different numbers of neutrons. These isotopes give rise to peaks at multiple locations $\mu + n$ in the spectrum (mono-isotopic molecule mass $\mu$ with $n = 0, 1, 2, \ldots$ neutrons). High resolution Time-Of-Flight (TOF), Quadrupole, Orbitrap and Fourier transform ion cyclotron resonance mass analyzers can detect isotopes as separate peaks in the spectrum. In surface enhanced laser/desorption and ioni-
4. Advanced deconvolution analysis of complex mass spectra

...
4.2 Materials and methods

4.2.1 SELDI-TOF MS data

Figure 4.1 presents real data from serum samples which were measured with a low-resolution SELDI-TOF mass spectrometer from Ciphergen. The serum samples were taken from patients treated for colon cancer; (Roelofsen et al. 2007) give a detailed description of the samples. The ‘SELDI method’ involves three steps: a specific fraction of molecules is enriched from the sample; the selected molecules are then embedded in a lattice of energy absorbing molecules (EAM) (also known as matrix molecules); and the EAM uses the energy from a laser to sublimate and ionize the selected molecules. The ‘TOF method’ makes use of an electric field to separate and detect the charged molecules based on their mass-to-charge ratio ($m/z$).

The upper panel in Figure 4.1 apparently shows two peaks which correspond to two singly charged molecule species. These two peaks are skewed and show shoulders. This is explained by the formation of intermolecular complexes of sample molecules with 0, 1, 2, and 3 matrix adducts, which here leads to $2 \times 3 = 6$ extra peaks in the spectrum. However the extra peaks can hardly be seen; a simple deconvolution method would probably just fit two skewed distributions to the spectrum.

The complexes can also get 1, 2, and 3 charges. The lower panel in Figure 4.1 shows that molecules with $>1$ charges generate peaks with higher resolution so that more peaks can be detected. Double and triple charge peaks can therefore provide helpful insight in the complexity of a mixture of single charge peaks suffering from overlapping of peaks.

4.2.2 Calibration of MS data

One is generally interested in the mass of the molecules and not in their time-of-flight (TOF). Therefore, the processing of samples typically starts with a calibration run. The measured TOFs of molecules with known masses in a synthesized sample can be used to set the calibration parameters. These parameters are used for the conversion from TOF data to mass-to-charge
4. Advanced deconvolution analysis of complex mass spectra

Figure 4.1: Row 1 and 2 show the single charge peaks (bold black curves) in Spectrum 1 and 2, respectively. Row 3 shows the corresponding triple charge peaks in Spectrum 1. Every column shows a fit of a mixture model (bold brown curves) to the data, ranging from a simple model in the left column to a complex/improved model in the right column, as indicated in the headings. The thin blue and dashed green curves correspond to the individual mixture distributions; blue means 0 adducts, green means $>0$ adducts. The thin vertical black lines indicate the estimated peak locations. Incorporating adducts in the model improves the fit, and linking peaks reduces the total number of parameters. Self-calibration reduces the mis-alignment between corresponding peaks within the spectrum.

$(m/z)$ data in the next runs. The derived $m/z$ data may be displayed and analyzed visually and computationally.

Unfortunately, calibration parameters derived from one spectrum do not always apply well to other spectra, i.e. (i) locations of corresponding peaks can be shifted across different spectra, and (ii) within a single spectrum, double charge peaks are not located exactly at half the mass position of the single charge peaks. Peak shifts between spectra (issue $i$) are small if the spectra are measured with a single instrument and within a short period of time (Jeffries 2005). Here we propose to self-calibrate a spectrum and address issue $ii$ so that multiple charge peaks locate at the correct positions as com-
pared to the single charge peak.

Figure 4.2: A detailed illustration of self-calibration of the spectrum shown in (a). Given two locations, $y_{\text{left}}$ and $y_{\text{right}}$, on the m/z-axis, the regions 1 and 2 are defined as shown in (b). Region 1 contains $z_1$ peaks (single charge) which correspond to $z_2$ peaks (double charge) in region 2. However, the relative locations of the $z_1$ and the $z_2$ peaks do generally not exactly match. Figure (c) illustrates this in close-up by doubling the m/z locations of $z_2$ peaks and plotting them on top of the corresponding $z_1$ peaks. Self-calibration optimizes the correlation between the intensities in the two regions as function of the calibration parameters. As a result, the locations of the $z_2$ peaks match the locations $z_1$ peaks, as is shown in (d).

We consider an experiment which consists of $K$ spectra, numbered $k = 1, 2, \ldots, K$. A spectrum can be interpreted as a histogram with TOFs, say $t_1, t_2, \ldots, t_I \in \mathbb{R}$, ordered from small (left) to large (right) on the x-axis, where the histogram of the spectrum has $I \in \mathbb{N}$ bins. Generally, the spectra $(1, 2, \ldots, K)$ have $x$-axes with corresponding TOF-labels, because the detector frequency is generally not altered between different measurements. Let $n_{k,1}, n_{k,2}, \ldots, n_{k,I} \in \mathbb{R}_{\geq 0}$ denote the intensities of the detection signal.
in spectrum $k$ after $t_1, t_2, \ldots, t_I$, respectively. We transform the TOF on the $x$-axis spectra to $m/z$-values, by means of a quadratic calibration equation,

$$\frac{y(t, \gamma)}{U} = \alpha (t - t_0)^2 + \beta$$

with calibration parameters $\gamma = (\alpha, \beta, t_0)$, and the known, applied electric field voltage $U$ (Ciphergen manual 2002).

A random spectrum can be self-calibrated by determining optimal values for calibration parameters $t_0$ and $\beta$. The values are optimal if the locations of the double charge peaks ($z_2$ peaks) in the spectrum best match the locations of the corresponding single charge peaks ($z_1$ peaks) in the spectrum, as is illustrated in Figure 4.2. We use the correlation between the measured intensities on the normal $m/z$-axis and the measured intensities in the same spectrum at $2 \times m/z$, to indicate the goodness of a match. Technically, we search calibration parameters such that the locations of the $z_1$ peaks in region 1, i.e.

$$[2 \times y_{\text{left}}, y_{\text{right}}]$$

match the locations of the $z_2$ peaks in the region two, i.e.

$$[y_{\text{left}}, \frac{y_{\text{right}}}{2}]$$

where $y_{\text{left}}$ and $y_{\text{right}}$ are two locations on the $m/z$-axis in the spectrum, e.g. the boundaries of the spectrum.

First, we calculate interpolated intensities in region 1, at twice the $m/z$-values in the region 2. Next we calculate the correlation between the intensities in the region 2 and the interpolated intensities in region 1. We use a non-linear Newton-type optimization method to optimize the correlation as function of the parameters $t_0$ and $\beta$. Prior baseline subtraction is recommended for spectra which suffer severely from chemical noise.

Figure 4.3 illustrates self-calibration in a spectrum with complex regions and many overlapping peaks. Figure 4.3(a) shows that self-calibration located the double and triple charge peaks at the correct relative position in the spectrum.
4.2 Materials and methods

Figure 4.3: Detailed illustration of the self-calibration of spectrum (c). The solid black curves in (a) and (b) plot the $z_1$ peaks between 11–12.5 kDa (left) and between 27–30 kDa (right), in close-up. The $z_2$ (dashed curves) and $z_3$ (dotted curves) peaks are superimposed after multiplying their locations by 2 and 3, respectively. Figure (b) shows that the relative locations of the $z_1$, $z_2$ and $z_3$ peaks in the original data do not exactly match. Figure (a) shows that self-calibration matches the relative locations of the peaks.

After finding optimal values for the parameters, $t_0$ and $\beta$, in the calibration equation 4.1, we determine an optimal value for the other parameter, $\alpha$. Changing the value of $\alpha$ results in a proportional scaling of the x-axis. Therefore, if the mass of one of the peaks in the spectrum is known, then $\alpha$ can be used to scale this peak, and thereby all other peaks, to the correct m/z-value. Alternatively, if none of the peak masses is known a priori, one can proceed
to fit our mixture model to the data, and use the estimated mass of the matrix adducts, $\mu_a$, to determine an optimal value for $\alpha$. The adduct mass for the SPA matrix is 206 Da, according to (Dijkstra et al. 2007). Therefore, multiplying the original value of $\alpha$ by a factor of $206 / \mu_a$ should result in optimal scaling of the $m/z$-axis, too.

If two (or more) self-calibrated spectra (still) mis-align relative to each other but just a proportional scaling of the $x$-axis of one of the spectra will solve this.

If the spectra in an experiment are acquired shortly after each other, and different spectra do not severely mis-align, we apply self-calibration to all spectra in an experiment by optimizing the sum of the correlations over all spectra, as function of the two calibration parameters $\beta$ and $t_0$. Next we determine an optimal value for $\alpha$, as discussed above.

Let $y_1, y_2, \ldots, y_I$ denote the $m/z$-values in the self-calibrated spectra, which correspond to the times-of-flight, $t_1, t_2, \ldots, t_I$, respectively, where $y_i = y(t_i, \gamma)$.

### 4.2.3 Models interconnecting peaks

Suppose the sample contains $M$ major molecule species, numbered $j = 1, 2, \ldots, M$, with molecular masses, $\mu_j$. A molecule can form intermolecular complexes with other molecules. The matrix molecules, which are abundant in the SELDI analysis, frequently react with the molecules of interest by forming intermolecular complexes (particularly the sinapinic acid (SPA) matrix). However, complexes between different sample molecules are less abundant and often do not exceed the noise level in the spectrum. We assume that a detected molecule can form non-covalent adducts with $a$ matrix molecules, for $a \in \{0, 1, \ldots, a_{\text{max}}\}$, where $a_{\text{max}}$ is the maximum number of molecules reasonably involved in a single complex. In Figure 4.1 we use SPA as matrix and we take $a_{\text{max}} = 3$, because peaks containing 4 adducts do not exceed the noise level in the spectrum.

We assume that a detected molecule (or intermolecular complex) can carry $z$ charges, for $z \in \{1, 2, \ldots, z_{\text{max}}\}$, where $z_{\text{max}}$ is the maximum number
of charges which a molecule reasonably carries. For the analysis in Figure 4.1, we take $z_{\text{max}} = 3$, because peaks with 4 charges do not exceed the noise level in the spectrum.

In addition to different numbers of adducts and charges, isotopes also contribute to the multitude of peaks which can originate from a single molecule species. In SELDI data the resolution is generally too low to observe the individual isotopic peaks. We refer to Section 4.4 for a detailed discussion about how our models can be used and extended for the analysis of high resolution data with isotopes.

We consider $y_i$ as the observed m/z-values and $n_{k,i}$ as the corresponding frequencies of the observations ($i \in \{1, \ldots, I\}$), in spectrum $k$. We assume that the m/z-values which are observed in a spectrum, $k$, derive from a mixture of a baseline distribution and $M \times (a_{\text{max}} + 1) \times z_{\text{max}}$ normal distributions. The normal distributions correspond to the observed peaks, and are defined by

$$f_{j,a,z}(y) = \frac{1}{\sigma_{j,a,z}\sqrt{2\pi}} \exp\left(-\frac{(y - \mu_{j,a,z})^2}{2\sigma_{j,a,z}^2}\right)$$

(4.4)

where $y$ is the observed m/z-value, the expected peak locations

$$\mu_{j,a,z} = \frac{\mu_j + a \cdot \mu_a}{z}$$

(4.5)

are the means of the distributions, and

$$\sigma_{j,a,z} = r \cdot \mu_{j,a,z}^2$$

(4.6)

are the standard deviations of the distributions, for a parameter $r \in \mathbb{R}_+$ which is related to the resolution of the peaks in the spectra.

We here model the baseline ($f_{k,\text{bl}}(y)$) with a lowess curve, which uses locally-weighted polynomial regression to enable a flexible fit to the baseline in the data (Cleveland 1979).

For spectrum $k$, the mixture distribution of the observed m/z-value $y$, is

$$f_k(y) = \sum_{j,a,z} p_{k,j,z} \cdot f_{j,a,z}(y) + p_{k,\text{bl}} \cdot f_{k,\text{bl}}(y)$$

(4.7)
where the parameters, \( p_\ast \in \mathbb{R} \) (\( *: \) any indexes) are the proportion parameters of the corresponding distributions, such that \( 0 \leq p_\ast \), and such that the area under each mixture distribution equals 1, \( i.e. \) for each \( k \), \( \sum_{j,a,z} p_{k,\ast} = 1 \).

### 4.2.4 Parameter estimation

We apply the iterative EM-algorithm (Dempster et al. 1977) to calculate maximum likelihood values for the parameters in the model. In the SELDI pre-processing pipeline, any peak detection method which can identify single charge peaks, can be used to initialize the parameters \( \mu_j \). We identified a peak at \( \mu_j \) as single charge, if peaks were detected at \( \mu_{j,a,z} \), for \( a = 0 \) and \( z = 1, \ldots, z_{\text{max}} \). Multiple charge peaks (\( z = 2, \ldots, z_{\text{max}} \)) have generally smaller proportions than the corresponding single charge peak. The proportion parameters can be initialized randomly, as long as \( p_\ast \in (0,1] \) and \( \sum p_{k,\ast} = 1 \), per mixture \( k \). However, a good guess is preferable to speed up the convergence of the algorithm. We initialize \( r = 2 \cdot 10^{-7} \). Each iteration consists of an E-step and an M-step. The E-step calculates the component membership probabilities for the normal distributions, by

\[
p_{k,j,a,z|i} = \frac{p_{k,j,a,z} \cdot f_{j,a,z}(y_i)}{f_k(y_i)} \quad (4.8)
\]

and

\[
p_{k,b|i} = \frac{p_{k,b} \cdot f_{k,b}(y_i)}{f_k(y_i)} \quad (4.9)
\]

given the current parameter estimates.

The M-step calculates the updated estimations of the parameters in the model. Let

\[
\varphi_{k,j,a,z,i} = \frac{n_{k,i} \cdot p_{k,j,a,z|i}}{z^2 \cdot \sigma_{j,a,z}^2} \quad (4.10)
\]

The updated estimates for the molecular masses are

\[
\hat{\mu}_j = \frac{\sum_{k,a,z,i} \varphi_{k,j,a,z,i} \cdot (z \cdot y_i - a \cdot \mu_a)}{\sum_{k,a,z,i} \varphi_{k,j,a,z,i}} \quad (4.11)
\]
4.2 Materials and methods

and the updated estimate for the mass of the adduct is

$$\hat{\mu}_a = \frac{\sum_{k,j,a,z,i} \varphi_{k,j,a,z,i} \cdot (z \cdot a \cdot y_i - a \cdot \mu_j)}{\sum_{k,a,z,i} \varphi_{k,j,a,z,i} \cdot a^2}$$ (4.12)

for $k = 1, \ldots, K; \ j = 1, \ldots, M; \ a = 0, \ldots, a_{\text{max}}; \ z = 1, \ldots, z_{\text{max}}; \ i = 1, \ldots, I$.

The newly obtained $\hat{\mu}_j$'s and $\hat{\mu}_a$ are used to calculate the resolution parameter

$$\hat{r}_2 = \frac{\sum_{k,j,a,z,i} \frac{n_{k,i} \cdot p_{k,j,a,z|i} \cdot (y_i - \hat{\mu}_{j,a,z})^2}{\hat{\mu}_{j,a,z}^2}}{\sum_{k,j,a,z,i} n_{k,i} \cdot p_{k,j,a,z|i}}$$ (4.13)

The baseline is updated for each spectrum individually. In spectrum $k$, the fractions $p_{k,\text{bl}|i}$ of the data $n_{k,i}$ are attributed to the baseline. The lowess curve is fit to

$$p_{k,\text{bl}|i} \cdot n_{k,i}, \quad \text{for} \quad i = 1, 2, \ldots, I$$ (4.14)

Finally, the proportion parameters are updated by

$$\hat{p}_{k,j,a,z} = \frac{\sum_{i} n_{k,i} \cdot p_{k,j,a,z|i}}{\sum_{i} n_{k,i}}$$ (4.15)

and

$$\hat{p}_{k,\text{bl}} = 1 - \sum_{j,a,z} \hat{p}_{k,j,a,z}$$ (4.16)

for $k = 1, \ldots, K; \ j = 1, \ldots, M; \ a = 0, \ldots, a_{\text{max}}; \ z = 1, \ldots, z_{\text{max}}; \ i = 1, \ldots, I$.

The $E$-step and the $M$-step are alternated until convergence. Some minor peaks, which are not included in the model, may turn out to bias parameter estimates of nearby peaks. We tackle this issue by implementing robustness weights in the parameter estimates. We explained the details on robust estimation in (Dijkstra et al. 2006). Alternatively, these biases can be corrected by including such minor peaks in the model.
4.2.5 Visualization

The spectrum intensities can be plotted on the $y$-axis versus the observed TOF values or $m/z$ values on the $x$-axis. Note that converting TOF into $m/z$ will change the area under the spectrum. Also note that equally sized TOF intervals correspond to differently sized $m/z$ intervals. The fit of a mixture distribution, $f_k$, to spectrum $k$ can be visually inspected on the $m/z$ scale by plotting $f_k$ on top of the spectrum, after taking the following two steps.

First, we multiply the mixture distribution ($f_k(y_i)$), which has area 1, with the area under the spectrum on the time-scale,

$$A_k = \Delta t \cdot \sum_{k,i} n_{k,i}$$

where

$$\Delta t = t_{i+1} - t_i$$

is the regular distance between the bins on the axis, which corresponds to the detector frequency. The area under the mixture distribution is now equal to the area under the spectrum on the time-scale.

Second, we scale the fitted intensities. This is necessary because the regular distances between the bins on the TOF scale become variable on the $m/z$-scale, which affects the area under the peaks. We multiply the mixture distribution with the Jacobian of the transformation ($\frac{\delta}{\delta t} y(t_i, \gamma)$), described by the calibration equation (equation 4.1). The Jacobian of the transformation is the derivative of the calibration function with respect to time,

$$\frac{\delta}{\delta t} y(t, \gamma) = \frac{\delta}{\delta t} \left( U\alpha (t - t_0)^2 + U\beta \right)$$

$$= 2 \cdot U\alpha(t - t_0)$$

Plotting

$$f_k(y_i) \cdot A_k \cdot \frac{\delta}{\delta t} y(t_i, \gamma)$$

on top of spectrum $k$, shows the fit of the model to the spectrum on the $m/z$-scale.
4.3 Results

Figure 4.1 step by step (column by column) extends a simple mixture model to a more advanced mixture model. The peaks in this figure correspond to two detected molecule species in two spectra. The two species generate multiple (overlapping) peaks within one spectrum because of matrix adducts and multiple charges. The upper two rows display the single charge peaks in spectrum 1 and 2. The third row displays the triple charge peaks in spectrum 1. The first column shows the simple approach, one normal distribution per local mode in the data. The vertical lines (see left green rectangle) in column 1 and 2 illustrate the discrepancies between the locations of the single charge peaks and the expected locations of the triple charge peaks. The second column incorporates the formation of matrix adducts in the model by adding an extra normal distribution for each matrix adduct. This improves the fit of the model to the data, and diminishes the discrepancies between the vertical lines. The third column links the parameters of peak components in our mixture models by making use of the known relationships between the locations of the peaks. Location estimations of corresponding peaks are linked across different spectra (row 1 and 2), and within each spectrum (row 1 and 3). Moreover, the parameters for the standard deviations are linked between all peaks in all spectra; i.e., we only use 1 parameter ($r$) to model the standard deviations of all peaks. However, the goodness of fit is diminished in the third column. This is mainly because the spectra are not self-calibrated, or in other words, the triple charge peak is not detected at $1/3$ of the molecular mass of the single charge peak. And, the double charge peak is not detected at $1/2$ of the mass of the single charge peak (not shown here). Therefore, we self-calibrate the spectra, as illustrated in the fourth column. The fourth column displays a parsimonious model (i.e., with a few parameters) which closely fits the data. We hereby reduce the total number of observed peaks to a much smaller number of underlying molecular species.

Figure 4.4 shows the fit of the parsimonious mixture model to another spectrum from the same data set. The right column (Cluster B) shows peaks in the same mass region as the peaks analyzed in Figure 4.1. We have ana-
4. Advanced deconvolution analysis of complex mass spectra

Figure 4.4: Deconvolution of two complex clusters with many overlapping peaks. Before deconvolution, the spectrum was self-calibrated, as shown in Figure 4.3. Row 1, 2 and 3 show the single, double and triple charge peaks, respectively. The data is plotted in black. The fitted mixture distribution is plotted in red. The thin blue and dashed green curves indicate the individual mixture components; blue means 0 adducts, green means > 0 adducts. We expect that the peaks below the curly bracket originate from matrix adducts.

We analyzed the single charge peaks in Cluster A (shown in upper left plot) before in (Dijkstra et al. 2006). However, in that previous analysis we did not link the \( z_1 \) peaks to the corresponding \( z_2 \) peaks, as we do here. The \( z_2 \) peaks have higher resolution and help to better deconvolute the \( z_1 \) peaks. We believe that the 36 peaks in this plot originate from only 8 molecule species. Six of these species giving rise to the peaks in Cluster A, and two giving rise to
the peaks in Cluster B. Other methods might not detect the peaks below the curly bracket in Cluster A, or, might explain these peaks as different molecules, i.e. independent from the other six molecules in Cluster A. As illustrated with the green peaks, our model can explain this complex region below the curly bracket by matrix adducts.

We can even go a step further and make use of the adduct mass $\mu_a$ to come up with optimal $m/z$-values on the $x$-axis. The parameter $\mu_a$ is estimated after fitting our model to the data, and it should have a value of 206 Da, according to (Dijkstra et al. 2007). The $x$-axis can be proportionally scaled by a factor of $206/\mu_a$, as is explained in detail in Section 4.2.2. This means that it is possible to come up with $m/z$ values on the $x$-axis, purely on the basis of adduct formation and the combination of single and double charge peaks in the spectrum.

4.4 Discussion

In this article we developed novel methods and models for the optimal deconvolution analysis of mass spectrometry data. We illustrated our models on the most complex and low resolution SELDI-TOF MS data with commonly observed phenomena such as adduct formation and varying numbers of charges. However, we anticipate that our method and models have a general applicability to, and are very useful for, many commonly used MS separation, ionization and detection techniques.

4.4.1 One general model for many MS technologies

Commonly used separation techniques which can be applied prior to MS analysis include liquid and gas chromatography (LC and GC), capillary electrophoresis, iso-electric focusing and 1-D and 2-D gel electrophoresis. Our models can be used to link peaks across the different fractions that are separated by these techniques.

Besides MALDI, a commonly used ionization technique is Electrospray Ionization (ESI). With ESI, a molecule can get many more charges than with
SELDI. We can take this into account by setting a higher value for $z_{\text{max}}$, e.g. 30.

Figure 4.2 illustrated that our current method for self-calibration optimizes the correlation between the intensities of $z_1$ peaks in region 1 and the intensities of the corresponding $z_2$ peaks in region 2. We notice that in addition to the considered peaks per region, region 1 may contain $z_2$ peaks, and region 2 may contain $z_1$ peaks, too. These peaks may slightly contribute to the self-calibration of the spectrum: $z_2$ peaks in region 1 may correspond to $z_4$ peaks in region 2, and $z_1$ peaks in region 2 may correspond to dimers in region 1. A dimer consists of two molecules of the same species which are linked together (Dijkstra et al. 2007). $z_4$ Peaks and dimers generally have low relative abundances compared to $z_1$ and $z_2$ peaks, and may therefore only slightly contribute to the self-calibration. Absence of $z_4$ peaks and dimers, is not expected to have a negative effect on the outcome of the self-calibration.

For the analysis of spectra, produced by other MS technologies such as ESI, in which molecules generally hold more charges, the Pearson correlation coefficient,

$$\rho(n_1, n_2) = \frac{\sum_i (n_{1,i} - \bar{n}_1)(n_{2,i} - \bar{n}_2)}{\sqrt{\sum_i (n_{1,i} - \bar{n}_1)^2 \sum_i (n_{2,i} - \bar{n}_2)^2}}$$ (4.22)

can be generalized to

$$\rho(n_1, n_2, n_3) = \frac{\sum_i (n_{1,i} - \bar{n}_1)(n_{2,i} - \bar{n}_2)(n_{3,i} - \bar{n}_3)}{\sqrt{\sum_i (n_{1,i} - \bar{n}_1)^2 \sum_i (n_{2,i} - \bar{n}_2)^2 \sum_i (n_{3,i} - \bar{n}_3)^2}}$$ (4.23)

and so on, where $n_{1,i}, n_{2,i}$ and $n_{3,i}$ are intensities in region 1, 2 and 3, and $\bar{n}_1$, $\bar{n}_2$ and $\bar{n}_3$ are the means of the intensities per region, respectively.

Commonly used detection techniques are Time-of-flight (TOF), multipole, Fourier transform (FT), and orbitrap. These techniques can produce high resolution spectra with peaks that show little or no overlap. Less overlap between peaks is favorable for the spectrum analysis because it simpli-
4.4 Discussion

fies the deconvolution analysis considerably. We have evidence that the par-
simonious interrelationship between the standard deviation of the normal
distributions in our models, which we defined in Equation (4.6), also applies
to data acquired with ESI FT-MS. The authors of (Marshall and Hendrickson
2001) analyzed Bovine Ubiquitin with ESI FT-MS and showed that the res-
olution of the resulting peaks was proportional to the charge on the mole-
cule. Our definition of the standard deviations implies that the resolution of
our model peaks is proportional to the “z/m-ratio” of the analyzed molecu-
les. Therefore, we anticipate that our models are directly applicable to data
acquired with ESI FT-MS; and probably to data that is acquired with other
detection techniques as well.

An isotope is one of the several forms which molecules of a given spe-
cies can have. In high resolution spectra individual isotopes can be detected
as separate peaks in the spectrum, whereas in SELDI the isotopic peaks are
nearly never fully separated and show almost always full overlap with each
other.

Different isotopes of the same molecule species have different numbers
of neutrons. A neutron has a mass of about 1.008664915 Da = δ, and does
not have a net charge. Therefore, the different isotopes will give rise to peaks
at multiple locations μ + n × δ in the spectrum (mono-isotopic mass μ with
n = 0, 1, 2, . . . neutrons). The interrelationships between peak locations can
easily be incorporated in our model, by adding extra peak components for
the isotopes of each of the analyzed molecule species. We can do this in a
parsimonious way, because we don’t need extra parameters to model the lo-
cations of these ‘extra’ isotopic peaks. Remark: the value of δ can be slightly
different for the isotopes of different molecules, due to differences in binding
energies between the atoms in the molecules.

We anticipate that, in protein spectra, we may moreover define parsimo-
nous interrelationships between the proportions of isotopic peaks. Proteins
consist of amino acids. An interesting property of amino acids is that their
composition is mainly limited to the chemical elements carbon (C), hydrogen
(H), nitrogen (N), oxygen (O), and sulfur (S) atoms. Based on the average
amino acid composition, Senko et al. derived a model amino acid ‘averagine’
4. Advanced deconvolution analysis of complex mass spectra

(Senko et al. 1995). The molecular formula of averagine:

\[ \text{C}_{4.9384}\text{H}_{7.7583}\text{N}_{1.3577}\text{O}_{1.4773}\text{S}_{0.0417} \]

Senko et al. can accurately predict the isotopic distribution of proteins of each given mass, based on this average protein composition. We can use the ‘average protein composition’ to incorporate the predicted isotopic distribution in our model. This will result in parsimonious models with interconnected proportions of interrelated isotopes, which are very suitable for the analysis of high resolution spectra with peaks on the isotopic mass level.

Used in the same way as we analyze the interconnected locations between adduct peaks and isotopes, our models can also be used for the analysis of samples in which common transformations take place; see (Breitling et al. 2006) for a list of common transformations which apply to proteins and metabolites. This should improve the estimates of molecular masses, and moreover reduce complex spectra with many peaks to a much smaller number of molecule species. Extending our model to analyze multiple common transformations is straightforward.

Optimization of the correlation between peaks with a known mass difference, as function of the calibration parameters, can improve the self-calibration of a given spectrum. In a similar way, the regular distances between isotopes in high resolution spectra can be used to further improve the self-calibration.

4.4.2 Biomarker discovery improved

Our novel models help to improve biomarker discovery for the following reasons. We (i) can detect peaks in complex regions of the spectrum since we make use of information from related regions with lower complexity and higher resolution, by linking peaks. This is important because each peak is a potential biomarker. Moreover, we (ii) produce appropriate estimates of the peak positions and the molecule masses. These estimates can help in subsequent (biomarker) molecule identification steps. We also (iii) improve the estimates of the molecule abundances, which increases the chance on finding
‘real’ biomarkers in the discovery phase. An additional improvement for biomarker discovery is that by linking peaks we (iv) reduce the total number of observed peaks in a spectrum to a much smaller number of underlying molecule species. This reduces the statistical test multiplicity in the biomarker discovery phase and therefore increases the power, and ultimately the chance on finding real biomarkers, even further.

4.4.3 Conclusion

In this paper, we presented a novel method, called ‘self-calibration’, to locate peaks at the correct locations in the spectrum. Moreover, we pointed out that our novel statistical models interconnecting peaks, have a wide applicability to commonly used MS techniques, improve biomarker discovery and have better power to get more out of your mass spectrometry data.