Simulation of the Usage of Gaussian Mixture Models for the Purpose of Modelling Virtual Mass Spectrometry Data

Małgorzata PLECHAWSKA a,1, Joanna POLAŃSKA b

a Computer Science Institute, Lublin University of Technology, Poland
b Institute of Information Technology, The Silesian University of Technology, Gliwice, Poland

Abstract. This article presents the method of the processing of mass spectrometry data. Mass spectra are modelled with Gaussian Mixture Models. Every peak of the spectrum is represented by a single Gaussian. Its parameters describe the location, height and width of the corresponding peak of the spectrum. An authorial version of the Expectation Maximisation Algorithm was used to perform all calculations. Errors were estimated with a virtual mass spectrometer. The discussed tool was originally designed to generate a set of spectra within defined parameters.

Keywords. mass spectrometry, GMM, expectation maximization algorithm

1. Introduction

Mass spectrometry is a rapidly-developing technique of tissue-samples-processing. This process lies in the chemical fragmentation of the sample into ions. The aim of mass-spectrometry data processing is to find significant compounds of analysed data. In order to achieve this result, the data sets need complex processing. Nowadays computational methods of mass spectra data are expanding rapidly. Moreover, such fields as baseline removal, normalization [1], denoising, peak detection and alignment [1] constitute the area of many new improvements.

2. Virtual Spectrometer

Virtual mass spectrometer is a computer model of the spectrometer. It was developed by Coombes et al. [2]. The above-mentioned model is able to generate realistic mass spectra. It simulates physical principles occurring in the spectrometer [3]. This beneficial tool allows testing and evaluating of new methods of processing and analysing mass spectra. The conceptual model is based on principles of MALDI mass spectrometer work. Assuming that \( n \) spectra are observed, each of them is taken on the same equally-spaced grid of length \( T \) of TOF \( t_j, j = 1, \ldots, T \). The next move consists

1 Corresponding Author: Małgorzata Plechawska, Computer Science Institute, Lublin University of Technology, Nadbystrzycka 36b, 20-618 Lublin, Poland; E-mail: gosiap@pluton.pol.lublin.pl.
of modelling log-transformed intensities. Logarithmical transformation is used in this
case to make the data more symmetric and to reduce the dependence between values of
means and variances. The presented model for the observed log-value of intensity for
the $i^{th}$ spectrum at $t_j$ TOF) is given with Eq. (1).

$$y_j(t_j) = B_j(t_j) + N_j S_j(t_j) + \epsilon_j$$  \hspace{1cm} (1)

$S_j(t_j)$ is a true signal composed of a set of possibly overlapped peaks. These peaks
represent some biological molecules. $N_j$ is a constant normalisation factor [4] used to
diversify obtained spectra, whereas $B_j$ represents the baseline function. The last
component is error, which is implemented as standard Gaussian distribution with mean
value equal to zero $\mathcal{N}(0, \sigma_j^2)$.

Virtual spectrometry has some important noteworthy features. One of them is the
relationship between m/z values and characteristics of peaks. A low m/z value usually
determines a tall, sharp peak, whereas a higher one results in broader and shorter peak.
The shapes of received peaks are estimated in the process of physical simulation.
Parameters available in the model cannot be used to directly characterise these shapes.
However, they are affected by numerous factors like isotopic distributions of spectra
components, values of initial velocities or time resolution of the detector.

The virtual spectrometer implements simplified assumptions and employs only
basic characteristics of MALDI instruments. However, it successfully generates spectra
with features similar to real ones.

3. Mixture Models

The mixture model is a technique of presentation and modelling of huge data-sets.
Each distribution represents one sub-process of the modelled phenomenon. In the case
of mass-spectra processing, each component of the mixture describes one single peak.
Mixture model is a combination of a finite number of distributions (Eq. (2)).

$$f^{mix}(x, \alpha_1, \ldots, \alpha_K, p_1, \ldots, p_K) = \sum_{k=1}^{K} \alpha_k f_k(x, p_k)$$  \hspace{1cm} (2)

where $K$ is the number of components in the mixture and $\alpha_k, k = 1, 2, \ldots, K$ are
weights of the particular component. $\sum_{k=1}^{K} \alpha_k = 1$. The density distribution function is
denoted as $f_k(x, p_k)$.

One of the best-known methods of solving complicated structures is the Expectation
Maximisation (EM) Algorithm. This iterative method includes three main steps. The
preliminary step consists of searching for initial parameters. The most popular method of
performing this process is randomisation. Two further steps are repeated until the
algorithm achieves convergence. The first step, Expectation (E), consists of calculation of
conditional probability that sample $x_n$ belongs to the $k^{th}$ component (Eq. (3)):
The second step, Maximisation (M) is responsible for obtaining new parameter values. Steps E and M are repeated until the stop criterion with a defined accuracy is reached. There are a few effective stop criterions. One of the most common is based on the likelihood function (Eq. (4)) and the maximum-likelihood rule. The maximum-likelihood [5] rule states that the higher the value of likelihood function, the better the estimation of parameters. This rule is commonly recognised as the main principle of EM algorithm operation.

\[
L(p, x) = L(p) = f(x_1, x_2, ..., x_N, p) = \prod_{n=1}^{N} f(x_n, p)
\]  

(4)

4. Application of EM Algorithm to Mass Spectrometry Data

The aim is to model pre-processed mass spectra with the mixture model and to solve its parameters with the EM algorithm. The standard version of this method uses a one-dimensional vector of input data. The data received from mass spectrometer have two dimensions: x (m/z values) and y (intensities) scales. Therefore a weighed form of the algorithm has been developed.

Values of intensities given in spectrometry data determine the number of repetitions of corresponding m/z values. To obtain a one-dimensional vector, each single m/z value from X axis \( x_i \) should be repeated \( y_i \) times. The data formulated in that manner could be used by the regular version of EM. However, this solution does not seem to be optimal. The EM algorithm is a computational-demanding method and usage of data with repeated values would additionally prolong the duration of the operation and the number of required iterations.

The appropriate multiplications are applied in a weighted version of the algorithm. E-step calculations remain unchanged because the process of obtaining the value of the probability that sample \( x_n \) belongs to the \( k \)th component needs to be performed in a traditional way. Regardless of the number of repeated values of \( x_n \) in a vector, the calculated probability is always equal for all of them. There are some differences concerning M-step calculations. A new formula is given with Eq. (5).

\[
\mu^\text{new}_k = \frac{\sum_{n=1}^{N} x_n p(k | x_n, p^\text{old})}{\sum_{n=1}^{N} p(k | x_n, p^\text{old})}, \quad k = 1, 2, ..., K
\]

\[
(\sigma^\text{new}_k)^2 = \frac{\sum_{n=1}^{N} (x_n - \mu^\text{new}_k)^2 p(k | x_n, p^\text{old})}{\sum_{n=1}^{N} p(k | x_n, p^\text{old})}, \quad k = 1, 2, ..., K
\]

\[
\alpha^\text{new}_k = \frac{\sum_{n=1}^{N} p(k | x_n, p^\text{old}) y_n}{N}
\]

(5)
5. Details of the Simulation Study

The virtual mass spectrometer allows generation of spectra with different heights and shapes of peaks. Wide peaks are often merged, and that makes them hard to determine. In order to obtain the rate of peak-detection process error, the appropriate simulations have to be performed. First, simulation parameters have to be properly set. Gaussian parameters used for calculation of velocities are of especially-great importance in this case. Small values of means result in setting of peak locations close to positions defined in the simulation input data. Usage of higher values results in the movement of peaks by proportional value. Simultaneously, setting standard-deviation values influences the shapes of peaks. The higher the values of standard deviations, the shorter the wither and noised peaks.

For the purpose of error-rate calculation, peaks should be placed in well-known locations. That is why the values of means in velocity parameters should to be as small as possible. However, the values of standard deviations may be set freely. The simulations were performed with the use of a variety of parameter sets to examine the efficiency of the algorithm on different data sets.

The examples of the simulation study are presented in Figure 1. Each spectrum is presented on two figures: the original one and the one obtained from EM algorithm. Examples consist of rates of error for particular spectra. Original spectra are placed on the left side of the figure, while spectra decomposed with EM algorithm are presented on the right. The central part is reserved for spectra after preprocessing.

The EM algorithm considers each decomposed peak as a density function. Therefore, values on the Y axis need to be rescaled to achieve the original data scale. However, this process is not really necessary, because the most interesting values are presented on the X axis (means and standard deviations). Consequently, values of Y axis are ignored.

The important task is to calculate the error between original and obtained spectra. Only values of means were used for calculations. Determination of the distance between particular peaks is given in Eq. (6):

\[
d = \frac{\sum (\text{sort}(\mu^{\text{orig}}) - \text{sort}(\mu^{\text{result}}))}{\sum(\mu^{\text{orig}})}
\]  

(6)
Figure 2 depicts the comparison of error rates. Figure 2a illustrates the comparison between all received errors. Figures 2b–d characterise errors for the number of components presented in the ranges, respectively <1-15>;<16-30>;<31-45>. Error rates depend on the value of the defined accuracy of the EM algorithm. The proper accuracy value should give an acceptable value of error. Too high an accuracy may result in the extended duration of calculations.

6. Conclusion

The presented method appeared to be accurate, especially regarding spectra with a complex structure and overlapped peaks. Its disadvantage mainly concerns the duration of calculations, that is, longer than standard methods based on local maxima and minima. Nevertheless, the computational time is acceptable and the achieved results are satisfactory. In the simulation study a virtual mass spectrometer was used. It appeared to be a beneficial tool facilitating the generation of spectra with defined parameters and shapes. Thanks to this tool, it was possible to estimate the errors occurring in results of mass spectrometry data processing. Presented method may also be applied to time-course DNA microarrays data studies.

Acknowledgement. Special thanks due to Michał Marczyk for his help and instructions concerning the virtual mass spectrometer. This work was supported by the European Program FI6R-036452, GENEPI-lowRT.

References