Published in final edited form as:

Commun Stat Simul Comput. 2019 ; 48(6): 1637–1650. doi:10.1080/03610918.2017.1422748.

A study of the properties of Gaussian mixture model for stable isotope standard quantification in MALDI-TOF MS

John Christian G. Spainhour¹, Michael G. Janech³, Viswanathan Ramakrishnan²

John Christian G. Spainhour: john.spainhour@bme.gatech.edu; Michael G. Janech: janechmg@musc.edu; Viswanathan Ramakrishnan: ramakris@musc.edu

¹Georgia Institute of Technology, Department of Biomedical Engineering, 313 Ferst Drive, Room 2127 Atlanta, GA 30332

²Medical University of South Carolina, Department of Public Health Sciences, 135 Cannon Street, Rm. 305, Charleston, South Carolina 29425. 843-876-1153

³Division of Nephrology, Department of Medicine, Medical University of South Carolina, Clinical Science Bldg., 829. 96 Jonathan Lucas St, Charleston, SC 29425. 843-792-1083

Abstract

The quantification of peptides in Matrix assisted laser desorption/ionization time-of-flight mass spectrum analysis coupled with stable isotope standards has been used to quantify native peptides under many experimental conditions. This approach has difficulties quantifying samples containing peptides with ion currents in overlapping (convolved) spectra. In a previous article we proposed a reparametrized Gaussian mixture model based on the known characteristics of the peptides that could also accommodate overlapping spectra. We demonstrated the application of our model in a series of single and overlapping peptides quantification experiments. Here, we focus solely on studying the properties of our approach and examine the characteristics of the GMM approach in convolved peptides using simulated spectra and provide a method for simulating these spectra.

Keywords

MALDI-TOF MS; Gaussian mixture; spectra simulation

1. Introduction

The ability to quantify proteins or peptides using Matrix assisted laser desorption/ionization time-of-flight (MALDI-TOF) mass spectrum (MS) is hindered by inherent physical and chemical properties of these materials [Bantscheff, 2007, Beck, 2011, Bantscheff, 2012, Gerber, 2003]. Differences in the charge, matrix effects, hydrophobicity, or posttranslational modification are just some of the factors that affect the ion formation and time of flight of a sample peptide or their mixtures. The work flow of these experiments must always be carefully managed so that possible loss of some of the samples during preparation or contamination from unusual sources is minimized. The MALDI has the benefit of being able to handle high molecular weight organic compounds without fragmentation [Bantscheff, 2007, Bantscheff, 2012]. Due to these internal standards with properties as similar as

possible to the material in question are preferred during quantification experiments. Peptides containing heavy carbon and nitrogen are one source of internal standards and are referred to as stable isotope standard (SIS) peptides.

Stable isotope dilution theory is based on the concept that a stable isotope labeled protein or peptide would behave exactly the same as the native, unlabeled peptide during MS analysis. Since the mass difference between the labeled and unlabeled samples can be detected through MS, a comparison can be made based on their signal intensities and quantification achieved. The SIS method involves placing a known quantity of labeled peptide in a sample and comparing the peak intensities between the labeled and native peptide. The labeled peptides act as an internal standard which can be compared to the native peptides of the same amino acid composition in the sample. The SIS has an error range from 2% to 12% [Bantscheff, 2012] depending on the quantification method used. The MS intensities can be measured in several ways and have been compared [Spainhour, 2014] in previous work. The mixture model approach using Gaussian and other distributions has been proposed previously [Schwacke, 2013, Dijkstra, 2006, Dijkstra, 2007, Dijkstra, 2009]. However, these methods do not effectively make use of the mechanisms known about the peptides. In a previous paper, we proposed an approach using the Gaussian mixture models (GMM) under a parametrization that incorporates the known mechanisms of the peptides that are expressed in MALDI data. This method also allows for the deconvolution of peptides when they have overlapping peaks [Spainhour, 2014].

While the previous article compared methods that are useful for quantifying single peaks, including the proposed GMM approach the characteristics of the new approach needs further scrutiny, especially for quantifying multiple peaks. The previous work [Spainhour, 2014, Schwacke, 2013] has shown that in a limited set of peptides in real applications, GMM is capable of both single peptide and convolved peptide quantifications within a reasonable margin of error. The GMM can be used to estimate peak areas for both native and SIS peptides which can then be compared using a ratio of the known SIS peptide amount and its associated peak areas to estimate the amount of native peptide in the sample. For studying the properties of GMM under the new parametrization as it applies to SIS an algorithm for generating data that mimic SIS data is necessary. Here, we provide an algorithm for the simulation of mass spectra and using that approach, simulate data under various scenarios to examine the properties of the GMM approach. Further, an improvement in the search algorithm using a quadrant ascent method is described. Proposals for improvement and expansion of the GMM in future research are also discussed.

2. Reparametrized mixture models for peptide quantification

In MS, data are generated in the form of a histogram. That is, any given mass to charge ratio (M/Z) measured in Daltons divided by the ionization charge applied to the molecule, x_i , the frequency, h_i , as detected in the machine as ionic current is observed. These M/Z and frequency pairs are output from the machine in the form of an $n \times 2$ matrix, where the first column represents the x_i (in M/Z) and the second column is the h_i . The total number of observations is, $\sum_{i=1}^n h_i$, where n is the number of unique x_i 's observed.

2.1 Mixture models

The density of an observation from a K component mixture distribution might be written

$$\sum_{k=1}^K \lambda_k f_k(x; \mu_k, \sigma_k^2), \quad (1)$$

where $f_k(x; \mu_k, \sigma_k^2)$ is a two parameter probability density function with parameter (μ_k, σ_k^2) , for the k th component. Also, $0 \leq \lambda_k$ for all k and they sum to 1. This is a Gaussian mixture when $f_k(x; \mu_k, \sigma_k^2)$ is a Gaussian density with mean and variance that is represented by (μ_k, σ_k^2) . Dijkstra *et al* in a series of articles [Dijkstra, 2006, Dijkstra, 2007, Dijkstra, 2009] describe the use of Gaussian mixtures and log Gaussian mixture models for quantification of peptides in proteomics. Here, we adapt their approach to some known characteristics of peptides by reparametrizing the mixture distribution. Although our approach is easily extendable to all mixtures, for simplicity, we restrict to Gaussian mixtures.

The parametrization proposed here is as follows. In the use of GMM for MS quantification, the first Gaussian (μ_0) can be determined within an unknown mass error to account for error in the standard curve calibration of the mass spectra. Subsequent means are known to be separated by the mass of a single neutron. That is, $\mu_k = \mu_0 + (k-1)\nu^0$, where μ_0 is the mean of the first Gaussian (or where the first peak is expected to occur) and ν^0 is the mass of a neutron (1.00866912 Daltons). The variance of the k Gaussians can be assumed to be equal within the small range of M/Z being examined in SIS quantification [Gerber, 2003, Hoffmann, 2007]. Mixture coefficients (λ_k) can be determined by the isotopic distribution of the peptide and therefore should be treated known. Further, the number of components, K , could be large. In applications, this number is typically capped at 7 or 8, where it almost covers 100% of the mixture distribution. This new parameterization reduces the dimensionality of the parameter space from $3K - 1$ parameters to two parameters. Since the previous approaches [Gerber, 2003, Hoffmann, 2007] do not incorporate these known characteristics of the peptides, they over-parametrize the model, resulting in lack of efficiency.

Under the new parametrization, the two unknown parameters are, mass error, μ_0 , and variance σ^2 (or peak width, σ).

Since mass spectra have some form of baseline shift or noise a second trapezoidal distribution can be combined with the Gaussian mixture. This also can be modeled through a mixture of trapezoidal distribution and Gaussian mixtures as follows:

$$f_B(x) = \gamma g(x; \alpha, \beta) + (1 - \gamma) f_{GM}(x; \Delta, \sigma^2), \quad (2)$$

where $f_{GM}(x; \mu, \sigma^2)$ is the reparametrized Gaussian mixture of the density shown in equation (1), $g(x; \alpha, \beta)$ is the trapezoidal density that has a height of $\alpha + \beta_x$ at x , the mass to charge ratio along the x-axis of the spectra, and $0 < \gamma < 1$ is the proportion of the baseline

shift in the total area. For g to be a proper density $\alpha > 0$ and the support for x must be finite such that g integrates to unity. (Since x range is unrestricted, to properly define the baseline density, ideally x window should be truncated at the tails of the distribution. For example, one could truncate at 7 Daltons from the first and the last peaks.)

Now, to accommodate convolved peptides, (that is, in the case of multiple overlapping peptides), a mixture of Gaussian mixture distributions could be used. Thus the complete mixture model proposed, denoted by $f_{CGM}(x, \Delta, \sigma, \eta, \gamma, \alpha, \beta)$, is

$$f_{CGM}(x, \Delta, \sigma, \eta, \gamma, \alpha, \beta) = \gamma g(x; \alpha, \beta) + (1 - \gamma) \sum_{j=1}^M \eta_j f_{GM}(x; \Delta_j, \sigma_j^2), \quad (3)$$

where $f_{GM}(x; \Delta_j, \sigma_j^2)$ for $j=1, 2, \dots, M$ represents the mixture normal density for the j th peptide, and σ are the corresponding vector valued parameters. In general, while the mixture proportions for each peptide is known, the mixture proportions, γ , of the convolved peptides may not be known.

Although the reparametrized mixture model reduces the dimensionality of the parameter space, the maximum likelihood estimation (MLE) of the parameters becomes more complicated than the unrestricted parametrization. As in the general case, the Expectation-Maximization (EM) algorithm [Dempster, 1977] could be used to simplify the maximization of the likelihood. The general form of the likelihood is,

$$L(\Delta, \sigma, \eta, \gamma, \alpha, \beta | \underline{x}) = \prod_{i=1}^n f_{CGM}(x_i, \Delta, \sigma, \eta, \gamma, \alpha, \beta)^{h_i}$$

This likelihood does not lead to closed form solutions (even in the M-step of an EM-algorithm). Furthermore, convergence of the EM algorithm is painfully slow. We propose an alternate hybrid approach that combines the first few steps of the EM algorithm and then performs an exhaustive search of α, β, γ , using parameters Δ and σ to obtain estimates that maximizes the coefficient of determination.

2.1.1. Peak area estimation algorithm for single peptide—The GMM algorithm is designed to estimate the parameters such that the goodness-of-fit of the peak area or area under the curve of all peaks making up a single peptide is maximized. In cases of convolved peptides, where the isotopic clusters of individual peptides overlap, this algorithm simultaneously estimates the peak areas for each convolved peptide separately. First we describe in detail an algorithm for single peptide. To begin a search of spectra, a list of peptides needs to be provided. A workflow for the algorithm, written in R [R Core Team, 2013] that will be released as an R package for future public use, is presented in Figure 1.

Based on the known properties of the peptides, or convolved peptides, a search window is defined, preferably using the first few steps of an EM-algorithm. Once the search window is provided, a search space for the unknowns in the GMM, namely, Δ and σ is determined. For

each combination of μ and σ in the search space, models of the peptide are estimated. These models start with estimating the isotopic distribution of the peptide based on the isotopic distribution of its constituent elements. To avoid overflows in the computations, this estimation is capped at a number close to 100% of the peptide being accounted for, typically at least 99.99%.

Each peptide in the cluster is constructed separately and represented in a matrix that we call a model matrix. The baseline noise in the mass spectrum is also included in the model matrix. The baseline noise is defined by descriptors of the intercept (α) and slope (β) represented in the model matrix by a vector of 1's and the corresponding x_i 's in the window being examined, respectively. Each of the peptides modeled and baseline descriptors (α , β) are represented as columns in the matrix.

To estimate the peak areas and baseline noise the model matrix is equated to a vector of peak intensities and a QR decomposition [Stoer, 2002] is applied to produce the solution which constitutes a vector containing estimates of peak area for each peptide in the cluster and the baseline shift parameters α and β of the trapezoidal distribution. The use of QR decomposition acts as a linear regression where we estimate the peak areas and baseline.

The estimated peptide peak areas, slope and intercept are then used to construct a model to be compared against the observed histogram. The model is used to calculate the coefficient of determination (R^2) as a goodness-of-fit metric. In other words, a regression model of the observed versus the fitted histograms are evaluated using the R^2 in the search space of μ and σ .

2.1.2. Algorithm Enhancements—A difficulty in implementing the GMM to large data sets of spectra may arise due to the computational time required to examine the full search space of μ and σ , the dimensionality of which could be large. A quadrant ascent search could be used to speed up the process. This search operates from an initial set of values of the two parameter vectors within the specified ranges and searches through the range of one of the parameters (say, μ_k) for a fixed value of the other parameters. Once this is established, at that optimum value of μ_k the search continues on another member of the parameter space. This is repeated for all parameters in each step of the algorithm until convergence. The quadrant ascent method is useful in speeding up the convergence to the best fitting parameters when the distribution of the criterion (here R^2) is quadratic around the maximum. The distribution of R^2 can be shown to be quadratic around the maximum, especially for large samples. To further speed up the search, we recommend using the first few iterations of the EM algorithm. The first few EM steps typically leap closer to the maximum likelihood estimates and therefore result in a range that is much more efficient for our algorithm. While the implementation of the EM algorithm until convergence is tedious, the hybrid algorithm not only converges faster but also converges to a solution closer to the MLE and thus guarantees maximum R^2 in the parameter range searched.

Although only a linear baseline shift is discussed here, higher order polynomials could be included in the QR decomposition by simply adding columns that are powers of x_i . In our enhancement of the algorithm, we have also incorporated this additional nuance.

2.2. Algorithm for simulating data mimicking SIS peptides in quantification

In previous articles we compared the GMM approach with the traditional approaches used for SIS peptide quantification and established the superiority of GMM. However, since the other approaches are not applicable for the quantification of convolved peptides, to evaluate the usefulness of GMM a simulation study is warranted. The first challenge in such a study is generating data that represent convolved peptides.

Our approach for simulating relevant data consists of three main parts. First is the generation of the isotopic peaks in the spectrum. Second is the addition of noise using a first-order autoregressive (AR(1)) model [Wallace, 2002]. The third is a baseline shift to model that represents the gross changes in the spectrum due to noise. This is done for separate peptides, native and SIS, and for convolved peptides, where the isotopic peaks overlap one another. This algorithm was implemented in an R computational environment [R Core Team, 2013] and used to generate spectra under a number of scenarios to test the flexibility of the GMM approach. The simulation algorithm will be made user-friendly and subsequently published as an R function in the GMM package.

The algorithm breaks a spectrum down into several characteristics, which we will call, simulation conditions. These consist of: the peptide(s) being simulated, total peak area (A), the size of the window and its bin width (step size) around that peptide on the x (M/Z) axis, peak width (σ), mass error adjustment (ϵ), AR(1) parameter (ρ), uniform distribution for white noise, sample count (S), bin size, slope (β), and intercept (β). (In the simulation R package these conditions could be randomly generated or specified based on prior knowledge.) For the purposes of evaluating GMM, the combinations of values under these simulation conditions were selected based on the experience from the existing data, described earlier. While this is not an exhaustive list of possible influences of spectrum production, it does provide enough scenarios to produce a large range of spectra to adequately study the properties of the GMM approach. A summary of this algorithm workflow can be seen in Figure 2.

The algorithm starts by calculating the mass (μ_0) and isotopic distribution (λ_k) of the peptide being simulated. Isotopic distribution is estimated by convolving the elemental distributions for each atom in the peptide being simulated. Individual elemental isotopic distributions come from the CRC Handbook [Haynes, 2010] and are universal distributions that may have local variations. A multinomial distribution based on the vector, λ , is then sampled S times to determine an empirical probability vector (λ_k) describing the number of samples per isotopic peak (i.e., the k th mixture normal component) in the peptide. For each λ_k , the number of samples in λ_k is taken from the corresponding Gaussian density, namely $N(\mu_0 + (k-1)\nu^0 + \epsilon, \sigma)$. All of the samples from every Gaussian distribution sampled this way (one for each isotope of the peptide being modeled) for this peptide are then binned into a histogram for a specified bin size. This histogram is then normalized by dividing by S so that the area under the peaks approximates 1. The histogram bins are then centered to the bin midpoints, matching the normalized bin count to a single point on the x_i axis of a spectrum. The centered, normalized histogram is then multiplied by the selected peak area A for the peptide being simulated. The modified histogram is then truncated to fit into the selected window. Although, this essentially truncates the mixture density at both ends of the tail, the

density outside of the trimmed region is calculated to be less than 0.01% of the total peak mass. Once each of the individual peptides in a cluster are generated in this fashion they are combined by addition of a vector of the individual peptide modified histograms to create the cluster of overlapping isotopic peaks in the cluster.

Generating noise starts with the generation of an AR(1) noise sampled along the x -axis window, using the bin size to measure the steps within the window, with an extra 2 Dalton (200 element) added at either end of the window. This uses the autocorrelation ρ and the uniform distribution selected on a specified range. The AR(1) noise is centered at zero by subtracting the mean of the noise generated within the window. This centered noise is then combined with the simulated peaks with a simple substitution, where if the simulated peak is zero or less than the noise being generated it is replaced with the noise. This mimics our treatment of noise being an additive factor and not a masking one.

The baseline bulk noise is then added using a trapezoid constructed over the window using intercept α and slope β over the range of the window. This is then added to the AR(1) modified simulated spectrum to mimic a typical MALDI-TOF observed in our database. The parameters and simulated spectra are stored in separate files.

2.2.1 Setting the Simulation Parameters—Deciding on the simulation parameters should be guided by experience with the peptides of interest. The simulation scenarios were determined through an examination of 1789 MALDI-TOF MS spectra from previously published work containing singular and convolved Angiotensin (labeled here as Ang) peptides [Velez, 2007, Velez, 2009, Velez, 2012, Schwacke, 2013, Spainhour, 2014]. Based on the analysis of sections of various spectra in our database, we observed several noise and signal characteristics along with estimates of the corresponding noise and model parameters. The following are highlights of some of these characteristics; 1) The tails of the spectra around peptides demonstrated autocorrelations and baseline shifts that were often linearly changing, 2) this noise was constant between peaks and consistently exhibited autocorrelations, 3) The noise seemed additive, 4) the signal appeared to drown the noise out with relatively smooth increases in intensity, and 5) the peaks for a given compound were separated by constant intervals and the peaks appeared to have the same width regardless of intensity. The fifth characteristic observed confirms our reparametrization of the Gaussian mixture given in equation (2).

Based on the information above, combinations of peak areas (50, 100, 250 units), σ (0.05, 0.1, 0.2), and $(-0.1, 0, 0.1)$ were varied and 1000 spectra were simulated for each combination (Table 1). Baseline shift (α and β), noise (ρ), and peptide (representing Ang-1-8) were kept constant across the simulations.

For the convolved peptides, simulations were performed using reversed ratios of peaks ranging from 1:1 to 1:10 (Table 2). All parameters ($\sigma = 0.1$, $\rho = 0.1$, $\beta = -2$, $\alpha = 50$) were kept the same except for the peak areas for each peptide.

To test the efficacy of peak area ratio calculations, simulation of Ang-(2-10) and Ang-(1-9) with SIS- Ang-(2-10) and SIS-Ang-(1-9) were performed (Table 3) to see what effect

different ratios of the native peptides had on the ratio estimations using the GMM. All parameters ($\sigma = 0.1$, $\tau = 0.1$, $\beta = -2$, $\alpha = 50$) except for peak area were kept stable and 1000 samples were obtained in each case.

To test the range of convolved peptides capable of being estimated using GMM a series of two by two, four, eight and 16 sets of convolved peptides (Table 4) were simulated with each peptide having a peak area of 200 units with all other parameters held constant. These peptides were randomly generated and then ordered by mass. Peptides that had a minimum of two Daltons in difference of mass were used. The two by two peptide simulation refers to a series, where the second and third peptides are separated by a minimum mass difference of 5 Da but all four peptides are considered one cluster. A complete list of the peptides generated and used can be found in the supplemental information. All simulations used a random number seed of 031851 (R v 3.0.2, setseed() function).

To summarize the performance of the GMM in these simulations bias, mean square error (MSE), for the estimation of the model parameters and percent error (Tables 1-4) for the peak area were computed under the various simulation conditions. These are also used to discuss the peak area ratio estimations for convolved peptides and sets of convolved peptides as seen in SIS quantification. A graphical representation showing a comparison of the calculated MSE and the Bias squared (Figures 3-5) are presented, to illustrate the correlation and accuracy of simulation estimation.

3. Simulation results

Our previous work has shown, in applications, GMM has similar or superior characteristics [Spainhour, 2014] to other methods of peak quantification for single peptides. Our simulation study aimed to discover the behavior of GMM estimation over a variety of conditions, especially for the multiple convolved peptides.

In the single peptide scenarios (Table 1-2, Figure 3) simulation results show several trends. Simulation conditions that decrease the signal to noise ratio, which can be seen as decreasing peak (signal) intensity, has the effect of decreasing the accuracy of peak prediction. In the single peptide simulations (Table 1) the error in estimation seem to increase as peak width increases or as peak area decreases. Both of these conditions decrease the signal to noise ratio. A higher error is seen when the peak width is small. Small peak width results when the intensity of the peaks is too low, leading to fewer points to estimate the peak area and an appearance of higher resolution. Although there is increase in error, the quantification is still within acceptable range of error. This can also be seen in Figure 3, the increase in quantification error mirrors the increases in bias and MSE of the estimation of the GMM parameters. This also shows that the accuracy and precision suffer as the signal to noise ratio decreases. The simulation of a single set of convolved peptides (Ang-(2-10) and Ang-(1-9), Table 2) shows similar behavior in the ratio estimations.

When analyzing simulated convolved peptides, GMM shows a tendency to be more error prone in the estimation of the last peptide in the convolved cluster. However, this error is still within the allowed limits found in literature (<10% peak area, Table 4) [Bantscheff, 2012].

For different ratios of convolved peptides (Table 3, Figure 4) similar effects to single peptide estimation, where the signal to noise ratio plays an important role in the quality of peak estimation, are observed. This is compounded when the second peptide has less contributing area to the cluster (Table 3 500:50, 1000:100 ratios). In every scenario, a similar decrease in accuracy and precision is observed (Figure 4), where the second peak ratio is harder to estimate. There is an increased error in peak area estimation caused by baseline estimation in peptides with lower concentration. The peaks that are due to the peptides with the larger amount of material are proportionately less affected by errors in estimation of the baseline shift. In all cases, the errors are within acceptable range. (We have provided examples of simulated data and how the models fit these data as supplemental figures.)

In the series of convolved peptides that was performed to evaluate the ability of GMM to estimate long chains of convolved peptides for analysis, each of the simulations shows the same pattern of having the most estimation error in the quantification of the last peptide in the range. We show the mean estimation error (Table 4) and the individual peak estimations in Figure 5. While estimation error is once again within the established bounds, we see that the quantification of the last peptide cluster together in the upper right corner of Figure 1. This suggests an increase in error but that error is still within the acceptable bounds for SIS quantification.

Each set of simulations highlight the behavior of GMM. In almost all conditions, the estimation error is acceptable but follows a predictable pattern that shows the importance of signal to noise ratio and its effects on peak quantification.

4. Discussion

GMM is a robust method for peptide quantification in MALDI-TOF experiments using SIS peptides. Previously, we have adequately demonstrated the application of the GMM for SIS quantification of single and convolved peptides and showed it agrees with the traditional approaches where applicable in the single peptide case [Spainhour, 2014]. Simulations performed here confirm, using GMM approach, both single and convolved peptide areas and area ratios can be estimated within an acceptable range of error. While GMM shows some sensitivity to signal to noise ratio, as reflected with increases in estimation error with wider peaks or lower peak areas that decrease the signal strength, the approach is robust and can quantify peak areas and subsequent ratios for quantification precisely with adequate (95%) statistical confidence. The known characteristics of MALDI-TOF in which the occurrence of peaks that resemble Gaussian densities, the known number and spacing of peaks lend MALDI-TOF data to modeling using GMM and makes GMM an ideal tool for this class of problems. The algorithm described above is meant for high-throughput use, scanning multiple spectra for multiple peptides and their SIS counterparts. We showed application of GMM in a subset of circumstances to illustrate its utility. However, this method can be improved in several ways.

The isotopic distribution used in this algorithm is based on a universal elemental distribution from the literature. Since local isotopic distributions for a given element or compound can be different, regional elemental distributions need to be considered. We assumed that all

simulated peptides have equal flyability in these simulations but this can be relaxed in a strait forward manner by applying a flyability constant to the data to reflect differences in matrix effects or ionization. Essentially, this will render unequal variances across components of the Gaussian mixture.

Future work with GMM will include an analysis of the minimal separation between peaks that is needed for accurate peak quantification. Since this method has an obvious weakness with peptides or other compounds with identical (or nearly so) monoisotopic mass, a possible minimal peak separation based on spectra resolution needs to be explored. As noted previously, the estimation error is greater in the last peaks of convolved sets of peptides. This will need to be addressed by exploring alternate methods for peak analysis including solving for the first two peaks of a cluster to estimate unseen peaks using GMM, then subtracting estimated peaks from the convolved peaks. This allows for an estimation of the 2nd peptide as a single isotopic cluster.

In this article we restricted our attention to Gaussian mixtures because our main goal was to study the characteristics of the reparametrized mixture model approach. As pointed out in the literature [Spainhour, 2014, Schwacke, 2013, Dijkstra, 2006, Dijkstra, 2007, Dijkstra, 2009] often the distribution around the peaks exhibit a skewness and therefore other distributions such as log Gaussian [Dijkstra, 2009] might provide better fit for the area. We believe our approach under other distributions will perform equally well. However, a sensitivity study, where the underlying distribution is misspecified is warranted and we plan to include this in a subsequent article. Also, we have considered only the cases where the number of peptides is known because our primary goal was to test if our approach works when that information is available. The situation where the number of peptides is unknown is important.

Outside of the subset of proteomics that is MALDI-TOF based peptide quantification, the application of GMM to mass spectral quantification goes beyond peptides. GMM could be of use in other fields such as metabolite profiling or the rapidly growing field of MALDI-Imaging. GMM is a versatile and robust method for peak quantification that has utility in multiple instances of peptide quantification in MALDI-TOF MS. It is ideally designed for high-throughput analyses and can be implemented in a straight forward manor.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We wish to thank:

Dr. John Arthur, Jessalyn L. Ierardi and Alison Bland (Nephrology Proteomics Laboratory) in the Division of Nephrology, Department of Medicine at MUSC were instrumental in helping us acquire spectra for MALDI- TOF analysis of peptide samples.

Dr. Juan Carlos Velez of the Division of Nephrology for is expertise and use of his data.

Dr. Beth Wolf in the Department of Public Health Sciences at MUSC, for her assistance with computational methodology and theory.

Larry Harbin Director of Research Technology for the Department of Public Health Sciences at MUSC, for his assistance with image processing.

The MUSC Mass Spectrometry Facility for their assistance in gathering some of the data used in this study.

This project was supported by a grant from the National Institute for Arthritis and Musculoskeletal and Skin Diseases (1P60AR062755). The funding agency had no role in the study design; in the collection, analysis and interpretation of data; in the writing of the report; or in the decision to submit the article for publication.

References

- Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B. 2007; Quantitative mass spectrometry in proteomics: a critical review. *Analytical and Bioanalytical Chemistry*. 389:1017–1031. DOI: 10.1007/s00216-007-1486-6 [PubMed: 17668192]
- Beck M, Schmidt A, Malmstroem J, Claassem M, Ori A, Szymborska A, Herzog F, Rinner O, Ellenberg J, Aebersold R. 2011; The quantitative proteome of a human cell line. *Molecular Systems Biology*. 7:549.doi: 10.1038/msb.2011.82 [PubMed: 22068332]
- Bantscheff M, Lemeer S, Savitski MM, Kuster B. 2012; Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Analytical and Bioanalytical Chemistry*. 404:939–965. DOI: 10.1007/s00216-012-6203-4 [PubMed: 22772140]
- Gerber SA, Rush J, Stemman O, Kirschner MW, Gygi SP. 2003; Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *PNAS*. 100:6940–6945. DOI: 10.1073/pnas.0832254100 [PubMed: 12771378]
- Spainhour JC, Janech MG, Schwacke JH, Velez JC, Ramakrishnan R. 2014; The Application of Gaussian Mixture Models for Signal Quantification in MALDI-ToF Mass Spectrometry of Peptides. *PLOS One*. 9:e111016.doi: 10.1371/journal.pone.0111016 [PubMed: 25372836]
- Schwacke JH, Spainhour JCG, Ierardi JL, Chaves JM, Arthur JM, Janech MG, Velez JC. 2013; Network modeling reveals steps in angiotensin peptide processing. *Hypertension*. 61:690–700. DOI: 10.1161/HYPERTENSIONAHA.111.00318 [PubMed: 23283355]
- Dijkstra M, Roelofsen H, Vonk RJ, Jansen RC. 2006; Peak quantification in surface-enhanced laser desorption/ionization by using mixture models. *Proteomics*. 2006(6):5106–16.
- Dijkstra M, Vonk RJ, Jansen RC. 2007; SELDI-TOF mass spectra: a view on sources of variation. *J Chromatogr B Analyt Technol Biomed Life Sci*. 847:12–23.
- Dijkstra M, Jansen RC. 2009; Optimal analysis of complex protein mass spectra. *Proteomics*. 9:3869–76. DOI: 10.1002/pmic.200701064 [PubMed: 19639597]
- Siepen JA, Keevil EJ, Knight D, Hubbard SJ. 2007; Prediction of missed cleavage sites in tryptic peptides aids protein identification in proteomics. *J ProteomeRes*. 6:399–408.
- Pfeifer N, Leinenbach A, Huber CG, Kohlbacher O. 2007; Statistical learning of peptide retention behavior in chromatographic separations: a new kernel - based approach for computational proteomics. *BMC bioinformatics*. 8:468. [PubMed: 18053132]
- Lan K, Jorgenson JW. 2001; A hybrid of exponential and gaussian functions as a simple model of asymmetric chromatographic peaks. *J Chromatography A*. 915:1–13.
- Zhou C, Bowler LD, Feng J. 2008; A machine learning approach to explore the spectra intensity pattern of peptides using tandem mass spectrometry data. *BMC bioinformatics*. 9:325. [PubMed: 18664292]
- Chris, Bielow; Stephan, Aiche; Sandro, Andreotti; Knut, Reinert. 2011; MSSimulator: Simulation of Mass Spectrometry Data. *J ProteomeRes*. 10:2922–2929. DOI: 10.1021/pr200155f
- Schulz-Trieglaff O, Pfeifer N, Gröpl C, Kohlbacher O, Reinert K. 2008; LC-MSSim--a simulation software for liquid chromatography mass spectrometry data. *BMC bioinformatics*. 9:423. [PubMed: 18842122]
- Hoffmann, E, Stroobant, V. *Mass Spectrometry: Principles and Applications*. Wiley-Interscience; 2007.
- Dempster AP, Laird NM, Rubin DM. 1977Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B39*. :1–31.

- Spainhour, JCG. Doctoral dissertation. Medical University of South Carolina; 2014. A Gaussian Mixture Method for Modeling and Assessment of MALDI-TOF Mass Spectra for Stable Isotope Standard Quantification.
- R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing; Vienna, Austria: 2013.
- Stoer, J, Bulirsch, R. Introduction to Numerical Analysis. 3. New York: Springer; 2002.
- Velez JC. 2009; The importance of the intrarenal renin-angiotensin system. *Nat Clin Pract Nephrol.* 5:89–100. [PubMed: 19065132]
- Velez JC, Bland AM, Arthur JM, Raymond JR, Janech MG. 2007; Characterization of renin-angiotensin system enzyme activities in cultured mouse podocytes. *Am J Physiol Renal Physiol.* 293:F398–F407. [PubMed: 17429035]
- Velez JC, Ryan KJ, Harbeson CE, et al. 2009; Angiotensin I is largely converted to angiotensin 1-7) and angiotensin (2-10) by isolated rat glomeruli. *Hypertension.* 53:790–797. [PubMed: 19289651]
- Velez JC, Ryan KJ, Harbeson CE, Bland AM, Budisavljevic MN, Arthur JM, Fitzgibbon WR, Raymond JR, Janech MG. 2009; Angiotensin I is largely converted to angiotensin (1-7) and angiotensin (2-10) by isolated rat glomeruli. *Hypertension.* 53:790–797. [PubMed: 19289651]
- Velez JC, Ierardi JL, Bland AM, Morinelli TA, Arthur JM, Raymond JR, Janech MG. 2012; Enzymatic processing of angiotensin peptides by human glomerular endothelial cells. *Am J Physiol Renal Physiol.* 302:F1583–F1594. [PubMed: 22461301]
- Wallace WE, Guttman CM. 2002; Data analysis methods for synthetic polymer mass spectrometry: autocorrelation. *J Res Natl Inst Stand Technol.* 107:1–17. [PubMed: 27446716]
- Haynes, WM. CRC Handbook of Chemistry and Physics. 91. Boca Raton: CRC Press; 2010.

Figure 1.

Data processing flowchart for GMM processing of spectra This is a summary flowchart showing how the GMM algorithm processes spectra based on user input to estimate peak area.

Figure 2.

Data generation flowchart for Spectra simulation This is a summary flowchart showing how the spectra were simulated using input parameters.

Figure 3.

Bias and MSE of single peptide simulations by Peak Area and Peak σ This figure shows the correlation between MSE and Bias squared as shown above with the simulation conditions that most directly affect the signal to noise ratio. As this ratio is decreased (area decrease, σ increase) we see an increase in both MSE and Bias.

Figure 4.

Bias and MSE of convolved peptide simulations This figure shows the correlation between MSE and Bias squared as shown above with the simulation conditions that most directly affect the signal to noise ratio. As the ratio of peak areas increases in favor of the first peptide the second peptide becomes harder to estimate.

Figure 5. Bias and MSE of convolved multiple peptide simulations Here we see the individual peptide estimation error for each convolved peak regardless of which set of convolved peptides estimated.

Table 1

Single peptide simulations over multiple conditions

σ	mzError	-0.1			0			0.1		
		50	100	250	50	100	250	50	100	250
0.05	Area									
	Bias	1.987	1.744	2.093	1.963	1.717	1.936	1.946	1.692	1.848
	MSE	3.964	3.173	4.688	3.868	3.077	4.027	3.803	2.994	3.714
	Percent Error	3.974	1.744	0.838	3.925	1.717	0.776	3.893	1.692	0.742
0.1	Bias	3.387	3.273	3.028	3.397	3.287	3.067	3.416	3.314	3.056
	MSE	11.529	10.83	9.737	11.592	10.909	9.943	11.724	11.085	9.828
	Percent Error	6.774	3.273	1.211	6.795	3.287	1.227	6.833	3.314	1.222
	Bias	4.414	3.457	2.343	4.87	4.109	3.132	5.154	4.503	3.646
0.2	MSE	19.789	12.318	6.491	23.945	17.209	10.721	26.763	20.583	14.122
	Percent Error	8.828	3.457	0.938	9.739	4.109	1.253	10.307	4.503	1.458

GMM estimation is performed under several simulation scenarios. The effects of peak width and overall signal intensity should be noted. If the peaks are too thin ($\sigma=0.05$), with too few points to help define a curve, GMM has fewer points to use in estimation. If the peaks are too wide ($\sigma=2$), lowering the signal to noise ratio, GMM has more difficulty predicting peak area. This is due to the noise to signal ratio inherent in the given simulation. As the peaks become more distinct from the baseline noise they become easier to estimate. Each combination of parameters was simulated 1000 times.

Table 2

Convolved Peptide Peak Area Estimation

Bias	-0.0159	-0.000123	-0.17	0.00395	-1.13	-0.92
MSE	0.000272	7.02e-07	0.0301	1.58e-05	1.33	0.9
Percent Error	1.59	0.27	4.24	3.95	11.27	9.21
Ratio	1:1	1:4	4:1	1:10	10:1	10:1
Peak Areas	200:200	200:800	800:200	50:500	500:50	1000:100

To test how well GMM can separate convolved peaks, Ang-2-10 and Ang-1-9 peaks were simulated and GMM was used to estimate the ratio between the two peptides. The peak areas were estimated and the ratio of peaks was compared to the known real ratio. Note that the error increases greatly when the second peak is the smaller of the two. However, the error is always within acceptable limits making GMM useful in all scenarios. Each set of convolved peaks was simulated 1000 times.

Table 3

Separate estimations of convolved peak simulations of Ang-2-10 and Ang-1-9

Ratio	Ang-2-10 ratio			Ang-1-9 ratio		
	Bias	MSE	Percent Error	Bias	MSE	Percent Error
200:200	2.039	0.004	1.019	5.141	0.027	2.571
200:800	2.075	0.005	1.037	8.687	0.078	1.086
800:200	-2.491	0.008	0.322	7.517	0.059	3.758
50:500	2.546	0.007	5.093	5.968	0.036	1.194
500:50	-0.61	0.001	0.188	4.992	0.026	9.984
1000:100	-4.423	0.023	0.443	7.989	0.067	7.989

The independent peak estimations for Ang-2-10 and Ang-1-9 convolved peak simulations show that the error in estimation using GMM pools in the second peak. This is seen most clearly when the second set of peaks (Ang-1-9) in the spectrum is smaller than the first (Rows 3, 5, & 6). These are the same scenarios performing poorly in table 2 (Columns 3, 5, & 6).

Table 4

Summation of all peaks in different sets of convolved peptides

Peaks	Bias	MSE	Percent Error
2 by 2	0.118	0.314	0.222
4	0.080	0.335	0.229
8	0.206	0.342	0.233
16	0.185	0.363	0.242

This table shows the average error in peak estimation of different sets of convolved peaks. Peptides were randomly generated and grouped based on mass such that there was a maximum of 2 Da between monoisotopic masses of the peptides. These were simulated in the various groups and the average error of estimation for each peptide was computed. Each set of convolved peaks was simulated 1000 times.