

Mass Spectrometry Data Processing Using Zero-Crossing Lines in Multi-Scale of Gaussian Derivative Wavelet

Nha Nguyen^{1,2}, Heng Huang^{1*}, Soontorn Orintara², An Vo³

¹Department of Computer Science and Engineering, ²Department of Electrical Engineering, University of Texas at Arlington, TX, USA.

³The Feinstein Institute for Medical Research, North Shore LIJ Health System, New York, USA.
Email: nha.nguyen@mavs.uta.edu, heng@uta.edu, orintar@uta.edu, anphuocnhu.vo@mavs.uta.edu.

ABSTRACT

Motivation: Peaks are the key information in Mass Spectrometry (MS) which has been increasingly used to discover diseases related proteomic patterns. Peak detection is an essential step for MS based proteomic data analysis. Recently, several peak detection algorithms have been proposed with good performance. However, in these algorithms, there are three major deficiencies: 1) because the noise is often removed as much as possible, the true signal could also be removed; 2) baseline removal step may get rid of true peaks and create new false peaks; 3) in peak quantification step, a threshold of signal-to-noise ratio (SNR) is usually used to remove false peaks, however, noise estimations in SNR calculation are often inaccurate in either time or wavelet domain. In this paper, we propose new algorithms to solve these problems. First, we use bivariate shrinkage estimator in stationary wavelet domain to avoid removing true peaks in denoising step. Second, without baseline removal, zero-crossing lines in multi-scale of derivative Gaussian wavelets are investigated with mixture of Gaussian to estimate discriminative parameters of peaks. Third, in quantification step, the frequency, standard deviation, height, and rank of peaks are used to detect both high and small energy peaks with robustness to noise.

Results: We propose a novel Gaussian Derivative Wavelet (GDWavelet) method to more accurately detect true peaks with a lower false discovery rate than existing methods. The proposed GDWavelet method has been performed on the real Surface-Enhanced Laser Desorption/Ionization Time-Of-Flight (SELDI-TOF) spectrum with known polypeptide positions and on two synthetic data with Gaussian and real noise. All experimental results demonstrate our method outperforms other commonly used methods. The standard receiver operating characteristic (ROC) curves are used to evaluate the experimental results.

Availability: <http://ranger.uta.edu/~heng/MS/GDWavelet.html>
or <http://www.naaan.org/nhanguyen/archive.htm>

Contact: heng@uta.edu

1 INTRODUCTION

Mass Spectrometry (MS) is a crucial analytical tool in proteomics research to provide tremendous information for disease proteomics study and drug targets identification at the protein/peptide level. Due to measurement error, chemical and other background noise, MS usually contains high frequency noise and consequently a multitude

of misleading peaks. Peak detection is one of the most important steps in MS data analysis because its performance directly effects the final proteomics study results.

Because the noise in MS comes from different resources and cannot be estimated, false positive peak detection results are unavoidable. This makes peak detection as a challenging problem. In recent years, several peak detection methods have been proposed (Morris *et al.*, 2005; Coombes *et al.*, 2005; Du *et al.*, 2006; Nguyen *et al.*, 2009). Most previous algorithms have four common pre-processing steps: denoising, baseline correction, alignment of spectrograms, and normalization. After pre-processing, local maxima usually is used to detect peak positions and design rules to quantify peaks. In this paper, we will explore the limitations of existing peak detection methods and propose several new algorithms to solve them.

Most peak detection methods employed denoising step by removing noise in each scale of wavelet, such as commonly used Cromwell (Morris *et al.*, 2005; Coombes *et al.*, 2005) and continuous wavelet transform (CWT) (Du *et al.*, 2006) methods. However, true peaks in MS could have large frequency response and be removed by denoising step. As a result, these true peaks cannot be detected. We propose using bivariate shrinkage model, which considers relationship of two neighbor scales, to remove noise in stationary wavelet domain. Because utilizing relationship between two neighbor coefficients or two scales of wavelets can keep high frequency true signal (Selesnick *et al.*, 2001). Stationary wavelet transform (SWT) utilizes spatial information of signals and suppress artifacts by redundant representations.

Baseline removal step was widely used in peak detection methods, but it often got rid of true peaks and created new false peaks. To avoid removing baseline, the CWT-based pattern-matching algorithm was introduced in (Du *et al.*, 2006). Using Mexican Hat wavelet in multi-scale, this method gave good results in peak detection with high sensitivity and low false discovery rate (FDR). However, the more important property of multi-scale in wavelet domain wasn't used in this method (WaveletTour Book, 2009). Instead of considering peaks as the sum of delta functions, more generally, we consider MS peaks as a mixture of Gaussian in which each peak corresponds to one Gaussian. We propose to use Gaussian derivative wavelet, instead of Mexican Hat wavelet which is only the second derivative of Gaussian wavelet. Zero-crossing lines which are robust to noise are also introduced to replace Ridge-lines in (Du *et al.*, 2006). We study the zero-crossing lines in multi-scale wavelet and provide new theoretical analysis.

*to whom correspondence should be addressed

In most peak detection methods, signal to noise ratio (SNR) was used to remove the small energy peaks with SNR values less than a threshold. But MS noise can not be correctly estimated in either time domain or wavelet domain. Thus, in this paper, instead of SNR, frequency response, height, and standard deviation of Gaussian peaks calculated by zero-crossing in Gaussian derivative wavelet domain are used to remove false peaks. In order to improve sensitivity, the Envelope analysis (Nguyen *et al.*, 2009) is also used to save some important peaks with small energy.

In this paper, we propose a new Gaussian derivative wavelet based peak detection method (GDWavelet) for Surface-Enhanced Laser Desorption/Ionization Time-Of-Flight (SELDI-TOF) spectrum. Both simulated and real spectrum with known polypeptide compositions and positions are used to evaluate our method. With simulated data, we compare different peak detection algorithms by both Gaussian noise and real noise. All experimental results show that our new approach can detect more peaks (in both high and low amplitude) with a lower false discovery rate than state-of-the-art methods.

2 METHODS

In this section, our new GDWavelet method will be introduced. In GDWavelet, we utilize bivariate smoothing model, Gaussian derivative wavelet, and envelope analysis. First, bivariate shrinkage estimator in SWT domain will be used to reduce noise and to keep whole true signal. Second, we will introduce how to detect peaks using Gaussian derivative wavelet through peak properties such as frequency response, standard deviation and height. Finally, envelope analysis is performed to save true small energy peaks which will be missed if only peak properties are used.

2.1 Smoothing by bivariate shrinkage function

Noise smoothing in MS is an important step which should remove noise and keep true peaks. In (Myers *et al.*, 2004), they tried to remove noise as much as possible, hence some true peaks were also removed. We propose to utilize bivariate shrinkage estimator in SWT domain to reduce noise and keep whole true signal. More precisely, we decrease the noise level without removing most of them. SWT is chosen due to its fast speed and redundant representations. The later step will further handle the remain noise.

To estimate wavelet coefficients, the most well-known rules are universal thresholding and soft thresholding (Donoho *et al.*, 1995) which was applied to Cromwell method (Morris *et al.*, 2005; Coombes *et al.*, 2005). These algorithms assume that wavelet coefficients are independent. Unfortunately, frequency response of peak is rather wide. Hard or soft thresholding only considers coefficients in a sub-band with narrow frequency. Recent research shows that algorithms utilizing the dependency between coefficients can give better results than those using the independency assumption (Sendur *et al.*, 2002). Sendur *et al.* exploited this dependency between coefficients and proposed a non-Gaussian bivariate pdf for the child coefficient w_1 and its parent w_2 as follows

$$p_{\mathbf{w}}(\mathbf{w}) = \frac{3}{2\pi\sigma^2} \exp\left(-\frac{\sqrt{3}}{\sigma} \sqrt{|w_1|^2 + |w_2|^2}\right). \quad (1)$$

The marginal variance σ^2 is dependent on the coefficients index k . By this bivariate pdf and the Bayesian estimation theory, the MAP

estimator of w_1 (Sendur *et al.*, 2002) is derived as

$$\hat{w}_1 = \begin{cases} 0 & \text{if } \sqrt{|y_1|^2 + |y_2|^2} < \frac{\sqrt{3}\sigma_n^2}{\sigma}, \\ \frac{\sqrt{|y_1|^2 + |y_2|^2} - \frac{\sqrt{3}\sigma_n^2}{\sigma}}{\sqrt{|y_1|^2 + |y_2|^2}} \cdot y_1 & \text{otherwise.} \end{cases} \quad (2)$$

where y_1 is child noisy coefficient, y_2 is parent noisy coefficient. This estimator is a bivariate shrinkage function. It has been used to smooth many kinds of signals such as image (Sendur *et al.*, 2002), DNA copy number (Huang *et al.*, 2008; Nguyen *et al.*, 2010), *etc.* In this paper, bivariate shrinkage estimator is used to smooth MS signals. An example of denoising result is shown in Figure 3(a). This example will be discussed in § 2.4.

2.2 Peak detection by Gaussian derivative wavelet

In previous work (Morris *et al.*, 2005; Coombes *et al.*, 2005), MS peaks were considered as the sum of delta functions. That means only heights of peaks have been used for peak detection throughout SNR. Du *et al.* (2006) utilized width of peaks to improve peak detection results a lot. We consider MS peaks as a mixture of Gaussian in which each peak corresponds to one Gaussian:

$$f(t) = \sum_{i=1}^N f_i(t) = \sum_{i=1}^N A_i \exp\left(-\frac{(t - \mu_i)^2}{2\sigma_i^2}\right). \quad (3)$$

With this assumption, four parameters providing intrinsic differences between true peaks and noise are peak position, standard deviation, height, and frequency response of peak. To find these parameters of a peak, we use zero-crossing lines in multi-scale of Gaussian derivative wavelet instead of ridge-lines in multi-scale of Mexican hat wavelet that was used by (Du *et al.*, 2006).

2.2.1 Theory of zero-crossing lines in multi-Scale Scaling theory for zero-crossings has been studied and applied to many applications. Yuille *et al.* (1986) assumed that signal is the sum of delta functions. Another similar assumption of signal, bandlimited signal, has been studied in (Vo *et al.*, 1996). However, studying zero-crossing of signals with Gaussian mixture assumption still is a new and challenging problem. We will build new theory of zero-crossing lines in multi-scale in following sections. Through our theory, parameters (position, standard deviation, height, and frequency response) of a Gaussian peak can be accurately estimated.

We use the first derivative of $f_i(t)$ to locate local maxima corresponding peak position: $f'_i(t_0) = 0$ with $t_0 = \mu_i$. We continue using the second derivative and third derivative of $f_i(t)$ to estimate height and standard deviation of Gaussian peak: $f''_i(t_0) = 0$ with $t_0 = \mu_i \pm \sigma_i$, $f'''_i(t_0) = 0$ with $t_0 = \mu_i$ and $t_0 = \mu_i \pm \sqrt{3}\sigma_i$.

Since smoothing performed in denoising step only reduces noise and keeps small true peaks, multi-scales of Gaussian derivative wavelet are used to make local maxima and minima more robust to noise instead of only one Gaussian filter in (Nguyen *et al.*, 2009). The wavelet transform can be written as convolution product in (4):

$$Wf(u, s) = \int_{-\infty}^{+\infty} f_i(t) \frac{1}{\sqrt{s}} \Psi^*\left(\frac{t-u}{s}\right) dt. \quad (4)$$

According to chapter 6 in (WaveletTour Book, 2009), the wavelet transform in (4) can be rewritten as a multi-scale differential

operator in (5)

$$W_n f(u, s) = s^n \frac{d^n}{du^n} (f_i \star \bar{\theta}_s(t))(u). \quad (5)$$

In this paper, the Gaussian wavelet is used. So, $\bar{\theta}_s(t)$ can be followed as (6):

$$\bar{\theta}_s(t) = \frac{1}{\sqrt{s}} \exp\left(-\frac{t^2}{s^2}\right). \quad (6)$$

If convoluting $f_i(t)$ and $\bar{\theta}_s(t)$, we get result in (7)

$$(f_i \star \bar{\theta}_s)(u) = K_1 \exp(-K_2(u - \mu_i)^2), \quad (7)$$

where $K_1 = A\sqrt{\frac{2\pi\sigma_i^2 s}{2\sigma_i^2 + s^2}}$ and $K_2 = \frac{1}{s^2 + 2\sigma_i^2}$.

Remark: The zero-crossing of $W_1 f(u, s)$ and $W_2 f(u, s)$ belong to connected curves that are never interrupted when the scale decreases.

Proof: With the first derivative, (5) can be rewritten as (8)

$$W_1 f(u, s) = -2sK_1K_2(u - \mu_i) \exp(-K_2(u - \mu_i)^2). \quad (8)$$

If $W_1 f(u, s) = 0$, we got $u_0 = \mu_i$ and $u_0(s+1) - u_0(s) = 0$ with any scale s . With the second derivative, (5) can be rewritten as (9):

$$W_2 f(u, s) = -2s^2K_1K_2[1 - 2K_2(u - \mu_i)^2] \exp(-K_2(u - \mu_i)^2). \quad (9)$$

If $W_2 f(u, s) = 0$, we get $u_0 = \mu_i \pm \sqrt{\sigma_i^2 + \frac{s^2}{2}}$, then $0 < u_0(s+1) - u_0(s) \leq 1$ with any scale s .

With the third derivative, (5) can be rewritten as (10)

$$W_3 f(u, s) = -4s^3K_1K_2^2 \cdot (u - \mu_i)[2K_2(u - \mu_i)^2 - 3] \exp(-K_2(u - \mu_i)^2). \quad (10)$$

If $W_3 f(u, s) = 0$, we get $u_0 = \mu_i$ or $u_0 = \mu_i \pm \sqrt{3}\sqrt{\sigma_i^2 + \frac{s^2}{2}}$. If we select $s = 100$ and $\sigma_i = 0.1$, then $u_0(100+1) - u_0(100) = 1.2247$. In conclusion, $0 \leq u_0(s+1) - u_0(s) \leq 1$ with the first and second derivative and zero-crossing lines belong to connected curves. Another conclusion is that zero-crossing lines is discontinuous lines if the third derivative Gaussian wavelet is used. Thus, only the first and second derivative Gaussian wavelets should be used in peak detection.

If f_i is a discrete signal, (4) can be rewritten as follows:

$$Wf(u, s) = \sum_k f_i(k) \int_k^{K+1} \frac{1}{\sqrt{s}} \Psi^*\left(\frac{t-u}{s}\right) dt. \quad (11)$$

We get $f(k)$ by sampling $f_i(t)$ with T_s :

$$f_i(k) = f_i(kT_s) = A_i \exp\left(-\frac{(k - \frac{\mu_i}{T_s})^2}{2(\frac{\sigma_i}{T_s})^2}\right). \quad (12)$$

If $W_2 f(u, s) = 0$, we get $u_0 = \mu_i \pm \sqrt{\sigma_i^2 + \frac{(s \times T_s)^2}{2}}$. if

$W_3 f(u, s) = 0$, we get $u_0 = \mu_i$ or $u_0 = \mu_i \pm \sqrt{3}\sqrt{\sigma_i^2 + \frac{(s \times T_s)^2}{2}}$.

Note: Zero-crossing line is more robust to noise than ridge-line. This conclusion is illustrated by an example in Figure 1. Figure 1(c)(e) show that zero-crossing lines are easier to detect than ridge lines linking local maxima or local minima points.

2.2.2 *Applying zero-crossing to peak detection* From section 2.2.1, parameters of a Gaussian peak could be estimated by follows:

Estimation of Peak Position: There are three ways to estimate peak positions throughout zero-crossing of three kind Gaussian derivative wavelets.

1. The First Gaussian Derivative Wavelet (Gaus1): zero-crossing line corresponds peak position. In multi-scale, this zero-crossing line is a continuous line with length N . Peak position should be estimated by

$$\mu_i = \frac{1}{N} \sum_{s=1}^N u_0(s). \quad (13)$$

2. The Second Gaussian Derivative Wavelet (Gaus2): there are two zero-crossing lines correspond two edges of Gaussian peak. They are u_{0left} and u_{0right} . Because two zero-crossing lines are symmetric at peak position, peak position should be estimated by

$$\mu_i = \frac{1}{N} \sum_{s=1}^N \frac{u_{0left}(s) + u_{0right}(s)}{2}. \quad (14)$$

3. The Third Gaussian Derivative Wavelet (Gaus3): there are three zero-crossing lines if using the third Gaussian derivative wavelet. They are u_{0left} , $u_{0middle}$ and u_{0right} . Because u_{0left} and u_{0right} are non-continuous lines, they should not be used to estimate peak position. From $u_{0middle}$, we can find peak position by

$$\mu_i = \frac{1}{N} \sum_{s=1}^N u_{0middle}(s). \quad (15)$$

Estimation of Peak's Standard Deviation: Another parameter of Gaussian peak is standard deviation σ_i . There are two ways to estimate σ_i by follows

1. The Second Gaussian Derivative Wavelet (Gaus2): from Remark, σ_i at scale s could be calculated by

$$\sigma_{i-left}(s) = \sqrt{(u_{0left}(s) - \mu_i)^2 - \frac{s^2}{2}}, \quad (16)$$

$$\sigma_{i-right}(s) = \sqrt{(u_{0right}(s) - \mu_i)^2 - \frac{s^2}{2}}. \quad (17)$$

After calculating $\sigma_{i-left}(s)$ and $\sigma_{i-right}(s)$ at all scales, σ_i should be estimated by

$$\sigma_i = \frac{\frac{1}{N_l} \sum_{s=1}^{N_l} \sigma_{i-left}(s) + \frac{1}{N_r} \sum_{s=1}^{N_r} \sigma_{i-right}(s)}{2}, \quad (18)$$

where N_l and N_r are length of left and right zero-crossing lines.

2. The Third Gaussian Derivative Wavelet (Gaus3): from Remark, σ_i at scale s could be calculated by

$$\sigma_{i-left}(s) = \sqrt{\frac{1}{3}(u_{0left}(s) - \mu_i)^2 - \frac{s^2}{2}}, \quad (19)$$

$$\sigma_{i-right}(s) = \sqrt{\frac{1}{3}(u_{0right}(s) - \mu_i)^2 - \frac{s^2}{2}}. \quad (20)$$

After calculating $\sigma_{i-left}(s)$ and $\sigma_{i-right}(s)$ at all scales, σ_i should be estimated by

$$\sigma_i = \frac{\frac{1}{N_l} \sum_{s=1}^{N_l} \sigma_{i-left}(s) + \frac{1}{N_r} \sum_{s=1}^{N_r} \sigma_{i-right}(s)}{2}, \quad (21)$$

where N_l and N_r are length of left and right zero-crossing lines. However, zero-crossing lines at left and right sides of the third

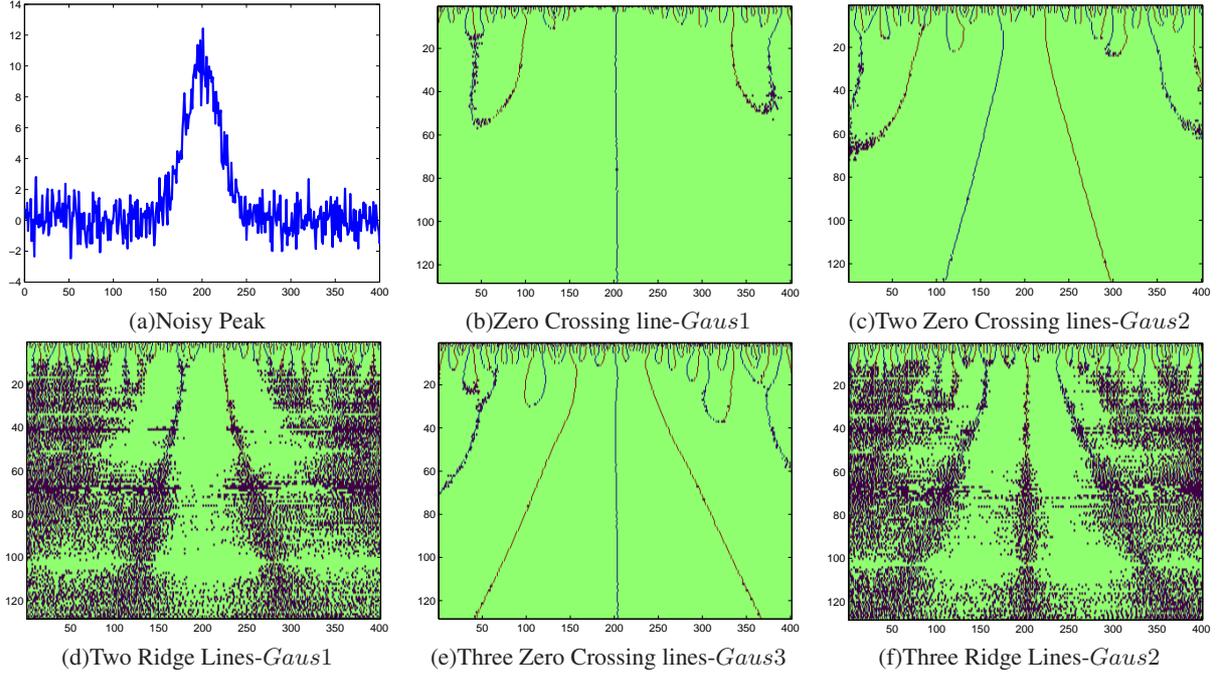


Fig. 1. An illustration of zero-crossing lines and ridge lines comparison. (a) A peak sample with shape followed $(10 \exp(-\frac{(t-5)^2}{2 \times 0.5^2}))$ and Gaussian noise ($\text{std}=1$); (b) Using *Gaus1*, the zero-crossing line corresponds to peak position, $t = 5$; (c) Using *Gaus2*, two zero-crossing lines correspond to two peak edges whose distances to peak position are $\sigma_i = 0.5$; (d) Using *Gaus1*, two ridge lines are corresponding to two peak edges whose distances to peak position are $\sigma_i = 0.5$; (e) Using *Gaus3*, three zero-crossing lines are corresponding to one peak position and two peak edges whose distances to peak position are $\sqrt{3}\sigma_i = 0.866$; (f) Using *Gaus2*, three ridge lines are corresponding to one peak position and two peak edges whose distances to peak position are $\sqrt{3}\sigma_i$.

Gaussian derivative wavelet are disconnected lines, so it is not easy to estimate σ_i through (19, 20, 21).

Estimation of Peak Height: Finally, we develop a way to estimate real height of Gaussian peak. With Gaussian peak $f_i(t) = A_i \exp(-\frac{(t-\mu_i)^2}{2\sigma_i^2})$, we have

$$A_i = \frac{f_i(\mu_i) - f_i(\mu_i - \sigma_i)}{0.3935}. \quad (22)$$

We can use (22) to estimate height of Gaussian peak after knowing μ_i and σ_i .

An Example: To demonstrate the above theory, we assume we have a Gaussian peak as follows:

$$x(t) = A_x \exp(-\frac{(t - \mu_x)^2}{2 \times \sigma_x^2}), \quad (23)$$

where $A_x = 10$, $\mu_x = 5$ and $\sigma_x = 0.5$. This peak is added Gaussian noise and baseline as follows:

$$f(t) = x(t) + G(\sigma, \mu) + b, \quad (24)$$

where b is constant, a representation of base line, $\mu = 0$ and $\sigma = [0.25; 0.5; 0.75; 1]$. With each σ value, 200 signals $f(s)$ have been created. One sample $f(t)$ is shown in Figure 1(a). We will estimate μ_x , σ_x and A_x using above zero-crossing theory. Error rate which is defined in (25) will be used to compare accuracy of different estimation methods:

$$\text{error rate} = \frac{|\text{true value} - \text{estimated value}|}{\text{true value}} \times 100. \quad (25)$$

Figure 1(b), (c), and (e) show zero-crossing lines in 128 scales using *Gaus1*, *Gaus2* and *Gaus3*. These zero-crossing lines will be used to estimate μ_x , σ_x and A_x . Table 2 lists error rates

of four methods to estimate peak position μ_x . With *Gaus1*, *Gaus2*, and *Gaus3* methods, μ_x values are calculated by (13, 14, 15) correspondingly. The term ‘‘with denoise’’ means bivariate shrinkage estimator is used to denoise Gaussian noise in signal $f(t)$. The Mexh, Mexican hat wavelet corresponding to *Gaus2*, is used as core part to detect peak in CWT method (Du *et al.*, 2006) and peak tree method (Zhang *et al.*, 2009). Based on Table 2’s results, the error rate when using Mexh wavelet (Du *et al.*, 2006) is the largest. We note that we use package ‘‘MassSpecWavelet’’ (Du *et al.*, 2009) which uses denoising with DWT and finds peak position using ridge lines (Du *et al.*, 2006) with Mexh wavelet. ‘‘Gaus1 with denoise’’ has the smallest error rate. However, error rates in *Gaus1* without denoising and in *Gaus2* are still acceptable and much better than in Mexh wavelet.

We can estimate σ_x by (18) or (21). However, with *Gaus3*, zero-crossing lines are not continuous lines (see Remark in section 2.2.1). Thus, estimation of zero-crossing in 128 scales of *Gaus3* is a problem. This problem causes a larger error in calculating the σ_x . From result of Table 1, we can conclude that *Gaus2* with denoising should be used to estimate σ_x because its’ error rate is the smallest.

By using (22) with zero-crossing lines of both *Gaus2* and *Gaus3*, the height of Gaussian peak is estimated. In this case, baseline b which is used in (24) is a constant. From Table 3, *Gaus2* with denoising gives the smallest error rate and should be used to calculate A_x .

From above example, the best way to estimate peak position μ_x is from the first Gaussian derivative wavelet, *Gaus1*. The second Gaussian derivative wavelet, *Gaus2*, should be used to estimate standard deviation σ_x and height A_x of a Gaussian peak. Figure 1(d)(f) show Ridge lines which correspond to zero-crossing

Table 2. Error of Peak Position Estimation. Using zero-crossing lines in multi-scale of Gaussian derivative wavelet, there are three ways to estimate peak position as in (13, 14, 15). We compare errors of these estimations and CWT's estimation (Du *et al.*, 2006). The error rate is defined by (25). In each Gaussian noise level, σ , we created two hundred signals. Error value shown in this table is average value.

σ in (24)	Gaus1 without Denoise	Gaus1 with Denoise	Gaus2 without Denoise	Gaus2 with Denoise	Gaus3 without Denoise	Gaus3 with Denoise	Mexh (Du <i>et al.</i> , 2006)
0.25	0.0519	0.0365	0.1533	0.1434	0.4890	0.2652	1.979
0.50	0.1319	0.0809	0.2253	0.1943	0.6918	0.3851	2.0170
0.75	0.1658	0.1034	0.3382	0.2353	0.7008	0.4855	2.1137
1.00	0.2118	0.1469	0.4630	0.2672	0.8681	0.5874	2.1618

Table 1. Error of Peak's Standard Deviation Estimation. σ_x can be estimated by (18) with *Gaus2* or (21) with *Gaus3*. Error rate here is defined by (25). These error values are average values which get from 200 signals with each added Gaussian noise level, σ .

σ in (24)	Gaus2 without Denoise	Gaus2 with Denoise	Gaus3 without Denoise
0.25	1.6560	1.3829	2.3371
0.50	2.5626	2.3392	3.7318
0.75	3.3841	2.5087	4.7881
1.00	3.9726	2.8529	5.9220

Table 3. Error of Peak's Height Estimation. Peak height A_x can be calculated by (22). Error rate here is defined by (25). These error values are average values which get from 200 signals, with each added Gaussian noise level, σ .

σ in (24)	Gaus2 without Denoise	Gaus2 with Denoise	Gaus3 without Denoise
0.25	4.1032	1.7544	4.8886
0.50	7.8084	2.6869	8.2126
0.75	11.0612	2.8954	14.3860
1.00	13.6141	3.0502	16.9405

lines in Figure 1(c)(e). It is clearly that detecting Ridge-lines is more difficult than detecting zero-crossing lines. Ridge-lines in Du *et al.* (2006) are similar to Ridge-lines in Figure 1(f), corresponding to zero-crossing line in *Gaus3* which should not be used because of its high error in calculating parameters of peaks.

2.3 Saving small energy peaks by Envelope analysis

Envelope analysis was introduced by (Nguyen *et al.*, 2009). Any finite energy signal $y(t)$ can be analyzed into three envelope signals including *MAX*, *MIN*, and *MED* envelopes at the first level. Each of these envelopes can be considered as a signal and will be decomposed into three envelopes. In this paper, we use *MAX* and *MED* envelopes to detect peaks because *MIN* envelopes contain no peak. We decompose the original signal into one *MAX* envelop at level 1, one *MAX* and one *MED* envelopes at level 2 and four envelopes which comprise two *MAX* envelopes and two *MED* envelopes at level $n > 2$. Empirically, 5 – 7 are recommended as the number of levels to get significant peaks.

2.4 Proposed Gaussian derivative wavelet based method (GDWavelet)

The framework of our proposed GDWavelet method is shown in Figure 2. First, raw MS data is smoothed by bivariate shrinkage estimator (2) in SWT domain to keep true signal and reduce noise. Note that, the lowest frequency detail scale and approximate scale

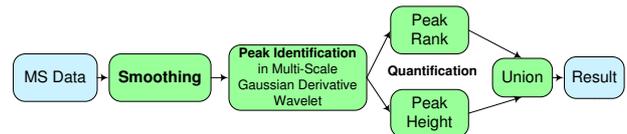


Fig. 2. GDWavelet Method's Flowchart. Raw MS data is smoothed by bivariate shrinkage estimator in SWT domain to keep true signal and to reduce noise. Without removing baseline, smoothed signal is used to estimate parameters of peaks by zero-crossing lines in multi-scale Gaussian derivative wavelet domain. After removing peaks with frequency response and width less than a threshold, we get all peak candidates. All peak candidates are quantified by peak rank in envelop analysis and peak height. Union results are final output.

which may include true signal should not be applied with any estimator, so that true signal isn't removed. As a result, noise in signal is reduced and smoothed signal still has a little noise. Second, without applying baseline removal that often discards true peaks and creates new peaks, smoothed signal is used to estimate frequency response, position, height, and standard deviation of Gaussian peak by zero-crossing lines in multi-scale Gaussian derivative wavelet domain. Frequency response of Gaussian peak is proportional to the length of zero-crossing line if the first derivative Gaussian (*Gaus1*) is used. Peak position, μ_i , is estimated by (13). Standard deviation, σ_i , of Gaussian peak is calculated by (18). Result of (22) with *Gaus2* gives heights of peaks. Using the first and the second derivative Gaussian wavelet, we can estimate all parameters of a Gaussian peak. After removing peaks with frequency response and standard deviation less than a threshold, we get all peak candidates. Third, in peak quantification step, we use two rules to remove false peaks: 1) All peak candidates are quantified by peak rank (PR) (Nguyen *et al.*, 2009) in envelop analysis. Peaks with $PR = 1$, even small peaks, are important peaks. 2) Peak height is used to remove peaks with height smaller than threshold. We use peak height to substitute SNR that was used by Morris *et al.* (2005) and Du *et al.* (2006), because noise can not be exactly estimated in either time domain or wavelet domain. Finally, the union results of two quantifying rules are the final detected peaks.

We randomly select 19th sample of CAMDA, 2006 to illustrate how GDWavelet method detects MS peaks. In Figure 3(a), blue signal represents raw signal and red one is signal smoothed by (2). A zoom in subfigure draws the peak which is used to show its' zero-crossing lines in Figure 3(b). Using one zero-crossing line in multi-scale of the *Gaus1* and two zero-crossing lines in multi-scale of the *Gaus2*, position, height, standard deviation, and frequency response of this peak are estimated. In Figure 3(c), we quantify peaks by two rules: peak height and peak rank (in

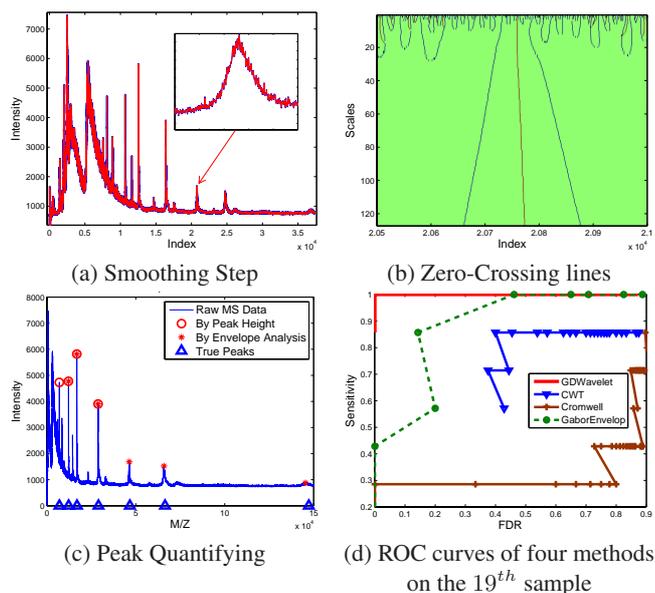


Fig. 3. An Illustration of GDWavelet. The 19th sample of (CAMDA, 2006) dataset is selected to illustrate how GDWavelet method detects MS peaks. (a) Blue signal is raw signal and red one is signal smoothed by bivariate shrinkage estimator in wavelet domain. (b) Parameters of peaks are estimated by zero-crossing lines. This figure shows zero-crossing lines of one zoomed peak in (a). (c) Peaks are quantified by peak height and peak rank. Union results include peaks with heights larger than a threshold or with peak ranks as one. (d) ROC curves of four methods’ performance on the 19th sample of CAMDA, 2006 dataset. GDWavelet gives the best performance.

envelope analysis). The circles are results from peak height based quantification. The stars are from peak rank based quantification. Union results include all peaks with heights larger than a threshold or peak rank one. Figure 3(d) shows ROC curves of four related methods. GDWavelet gives the best performance.

3 EXPERIMENTS AND DISCUSSIONS

3.1 Experimental setup

Cruz-Marcelo *et al.* (2008) and Emanuele *et al.* (2009) presented the extensive studies to compare the performance of state-of-the-art methods for SELDI data pre-processing, including CWT (Du *et al.*, 2006), Cromwell (Morris *et al.*, 2005; Coombes *et al.*, 2005), PROcess (Li *et al.*, 2006), Ciphergen and SpecAlign (Wong *et al.*, 2005). They concluded that CWT (Du *et al.*, 2006) has the best performance. Another method which also works well is Cromwell (Morris *et al.*, 2005; Coombes *et al.*, 2005). In this section, our GDWavelet method will be compared to the Cromwell (Morris *et al.*, 2005; Coombes *et al.*, 2005), the CWT (Du *et al.*, 2006), and our previous method, GaborEnvelop (Nguyen *et al.*, 2009). Cromwell method is implemented by MATLAB which can be downloaded from (UTMD Anderson, 2002). The CWT method (Du *et al.*, 2006) was implemented in R (called ‘MassSpecWavelet’) and Version 1.12 can be downloaded from (Du *et al.*, 2009). GaborEnvelop (Nguyen *et al.*, 2009) is implemented in MATLAB.

We evaluate the performance of above methods by the ROC curve. Both simulated and real data are used. The first simulated data was

proposed by Morris *et al.* (2005) and Coombes *et al.* (2005) and is available for download at (Simulated proteomics spectra, 2005). In this data, hundreds of mean spectrum samples with hundreds of proteomics datasets are generated.

Based on the simulation engine developed by Morris *et al.*, 2005 and code (R and MATLAB) to generate simulated data proposed by Cruz-Marcelo *et al.* (2008) and Zhang *et al.* (2009), we also create two new simulated datasets to investigate noise affection on different algorithms. The 100 spectrums with 20 – 30 true peaks are created first, and Gaussian and real noise are added separately to get two datasets. When Gaussian noise is added, each sample includes 20% of protein peaks which are below three time of SNR. Real noise is extracted from real data in which there is not any true peaks. There is only noise from 26000 (index) to end in 1st sample of CAMDA, 2006. Real noise probability density function is built. Using this function, noise with different standard deviation will be created. Based on this configuration, we create about 20 – 30 true peaks and more small energy peaks in simulated data.

The CAMDA dataset (2006) of all-in-1 Protein Standard II (Ciphergen Cat. # C100 – 007) is the real dataset. Because we know polypeptide composition and position in this dataset, we can estimate the sensitivity and the false discovery rate (FDR). There are seven polypeptides which create seven true peaks at 7034, 12230, 16951, 29023, 46671, 66433, and 147300 of the *m/z* values.

The sensitivity and FDR of four methods are calculated for 60 real MS signals and three simulated MS datasets with 100 signals each. The SNR thresholding values are increased gradually to calculate the ROC curves of Cromwell and CWT methods. The SNR thresholding values are chosen from 0 to 20 for Cromwell method and from 0 to 120 for CWT method. In our GDWavelet method, the peak height ratio, which is defined as the ratio of current peak height over average height of peaks, is changed from zero to ten to build the ROC curve. We plot the average ROC curves in Figure 4 and Figure 5. We should notice that we take average of all ROC points with the same SNR threshold (for Cromwell and CWT) and the same peak height rate (for our GDWavelet method).

3.2 Experimental results

Three simulated datasets and one real SELDI-TOF dataset are used to create ROC curves in Figure 4 and Figure 5. In all four datasets, the performance of Cromwell is not stable and gets worse than CWT and GDWavelet. Between GaborEnvelop which used envelope analysis and CWT which used ridge lines and SNR in peak quantification, GaborEnvelop is better than CWT in real data in Figure 4 (b) and CWT is better than GaborEnvelop in simulated data. In all cases, our GDWavelet method has much better performance than GaborEnvelop and CWT methods. At the same FDR, the sensitivity of our method is consistently higher than GaborEnvelop’s and CWT’s. It is clear that utilizing both of envelope analysis and Gaussian derivative wavelet in peak quantification made significant contributions to detect both high energy and small energy peaks. Bivariate shrinkage estimator in wavelet domain guarantees that denoising step in our method saves more true signal than (Morris *et al.*, 2005). Zero-crossing lines based peak parameters estimations in our paper is more accurate and robust to noise than ridge lines in (Du *et al.*, 2006). Envelope analysis is more efficiently used in GDWavelet than in

GaborEnvelop. Therefore, the GDWavelet has better peak detection results than Cromwell, GaborEnvelop, and CWT. Thus, it is an efficient and accuracy method to detect peaks in both real and simulated MS data. In Figure 4 and 5, CWT's ROC curves is limited in small FDR because two thresholds of the length of ridge lines and the scale corresponding to the maximum amplitude on the ridge line are used as default in MassSpecWavelet (Du *et al.*, 2009). Finally the runtime of GDWavelet algorithm is comparable to CWT method, because both methods need more computational time to decompose a signal to many scale using continuous wavelet transform.

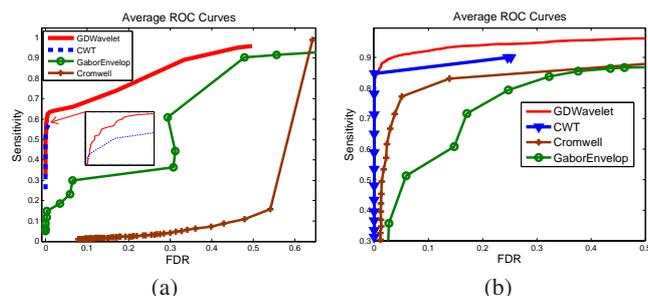


Fig. 4. ROC Curves - Simulated data with Gaussian noise. Average ROC curves of four methods (Cromwell, CWT, GaborEnvelop, and GDWavelet). (a) Obtained from 100 mean simulated MS signals which can be downloaded from (Simulated proteomics spectra, 2005). There are 149 true peaks in this data. (b) Obtained from 100 simulated MS signals in which Gaussian noise is added. There are 20 – 30 true peaks in this data.

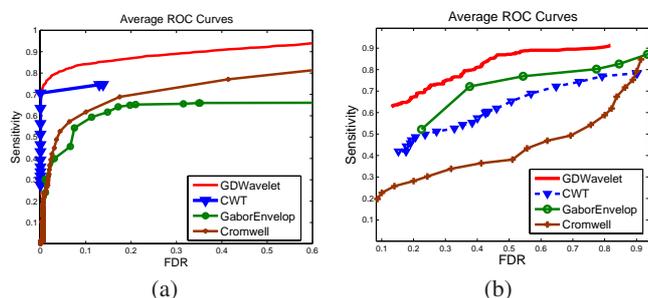


Fig. 5. ROC Curves - Simulated data with real noise and real data. Average ROC curves of four methods (Cromwell, CWT, GaborEnvelop, and GDWavelet). (a) Obtained from 100 simulated MS signals in which real noise is used. There are 20 – 30 true peaks in this data. (b) Obtained from 60 MS signals (CAMDA, 2006). There are seven true peaks in this data.

4 CONCLUSION

In this paper, we proposed new zero-crossing line theory in multi-scale of Gaussian derivative wavelet to estimate parameters of peaks in mass spectrometry which has been assumed as a mixture of Gaussian. A novel GDWavelet method was proposed to efficiently and accurately detect MS peaks by integrating bivariate shrinkage model, Gaussian derivative, and envelope analysis. The bivariate shrinkage estimator in SWT domain was used to reduce noise and still keep true peaks. All parameters of a Gaussian peak, estimated by multi-scale in Gaussian derivative wavelet and envelope analysis, have been used to remove false peaks. The peak height and peak rank were introduced as a nice substitution of the previous SNR

method to identify true peaks. Both simulated data and real MS data are used to evaluate our GDWavelet method. Simulated data was created with both Gaussian noise and real noise. Our GDWavelet method gave out a much better performance in the ROC curves than three other state-of-the-art peak detection methods. GDWavelet algorithm will be extended and test with other kinds of MS (such as MALDI-TOF) as future work. Based on GDWavelet method, many MS data related applications will be improved, such as protein identification, biomarker discovery, cancer classification, *etc.*

ACKNOWLEDGEMENT

Funding: The project is supported by NSF-CCF 0830780, NSF-CCF 0939187, NSF-CCF 0917274, NSF-DMS 0915228, NSF-CNS 0923494, UTA-REP.

REFERENCES

- CAMDA, (2006), Conference contest. Available at <http://camda.duke.edu/camda06/datasets/index.html> (last accessed data July 12, 2010).
- Coombes,K. *et al.*, (2005), Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform, *Proteomics*, 5, 16, 4107-4117.
- Cruz-Marcelo,A. *et al.*, (2008), Comparison of algorithms for pre-processing of seldi-tof mass spectrometry data, *Bioinformatics*, 24, 2129-2136.
- Du,P. , (2009), Mass spectrum processing by wavelet-based algorithms. Available at <http://bioconductor.org/packages/2.5/bioc/html/MassSpecWavelet.html> (last accessed date July 12, 2010)
- Du,P. *et al.*, (2006), Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching,*Bioinformatics*, 22(17) 2059.
- Donoho,D. *et al.*, (1995),De-Noiseing by soft-thresholding, *IEEE Trans. on Inf. Theory*, vol. 41, no. 3, pp. 613–627.
- Emanuele,V. and Gurbaxani,B. ,(2009),Benchmarking currently avaiable seldi-tof ms preprocessing techniques, *Proteomics*, vol. 9, pp. 1754–1762.
- Huang,H. *et al.*, (2008),Array CGH data modeling and smoothing in stationary wavelet packet transform domain, *BMC Genomics*, vol. 9, p. S2:S17.
- Li,X. *et al.*, (2006), SELDI-TOF mass spectrometry protein data, chapter 6, Springer New York, 91109, 2006.
- Myers,C. *et al.*, (2004), Accurate detection of aneuploidies in array CGH and gene expression microarray data,*Bioinformatics*, vol. 20, pp. 3533–3543.
- Morris,J. *et al.*, (2005), Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum, *Bioinformatics*, 21, 9, 1764-1775.
- Nguyen,N. *et al.*, (2009), Peak detection in mass spectrometry by gabor filters and envelope analysis, *JBCB*, vol. 7, pp. 547–569.
- Nguyen,N. *et al* (2010), Stationary Wavelet Packet Transform and Dependent Laplacian Bivariate Shrinkage Estimator For Array-CGH Data Smoothing, *JCB, Journal of Computational Biology*, 17, 139–152.
- Selesnick,I. W.,*et al.*,(2001),Hilbert transform pairs of wavelet bases,*IEEE Signal Processing Letter*, vol. 8, no. 6, pp. 170–173.
- Sendur,L. *et al.*,(2002), Bivariate shrinkage function for wavelet-based denoising exploiting interscale dependency, *IEEE Trans. on Signal Processing*, vol. 50, no. 11, 2744-2756.
- Simulated proteomics spectra, (2005). [Online]. Available: Available at <http://bioinformatics.mdanderson.org/Supplements/Datasets/Simulations/index.html> (last accessed date July 12, 2010).
- UT-MD-Anderson-Cancer-Center, The new model processor for mass spectrometry data. Available at <http://bioinformatics.mdanderson.org/cromwell.html> (last accessed date July 12, 2010).
- Vo,A. *et al.*, (1996), Scaling theorems for zero crossings of bandlimited signals,*IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 18, pp. 309–320.
- Mallat,S., (2009) *Wavelet Tour of Signal Processing - The Sparse Way.*, Academic Press, 2009, Elsevier.
- Wong,J. *et al.*, (2005), Specialignprocessing and alignment of mass spectra datasets, *Bioinformatics*, 21, 2088-2090.

Yuille,A. *et al.*, (1986), Scaling theorems for zero crossings, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 8, pp. 15-25.

Zhang,P. *et al.*, (2009), Peak tree: A new tool for multi-scale hierarchical representation and peak detection of mass spectrometry data, *IEEE/ACM TCCB*.