Statistical Computing Schemes for

Proteomics Data Processing and Insurance Solvency Modeling

by

Lu Xiong

A Dissertation Presented to the Faculty of the Computational Science Program

In Partial Fulfillment of the Requirements for the Degree of

Ph.D. in Computational Science

Middle Tennessee State University

December 2014

Thesis Committee:

Dr. Don Hong, Chair

Dr. William Robertson

Dr. Xiaoya Zha

Dr. Qiang Wu

Dr. John Wallin

This thesis is dedicated to my parents, who have been teaching me, encouraging me and supporting me in my life. Thanks for all your patience, love, and unconditional support.

ACKNOWLEDGMENTS

I would like to thank my adviser, my thesis committee, my classmates and Computational Science (COMS) PhD Program at Middle Tennessee State University. I learned a lot in my four and half years' PhD study in COMS program. Dr. Don Hong, my adviser, has spent lots of time on my work and I benefited much from discussions with him. His profound knowledge and experience in mathematics, statistics and computational science directed me in correct direction on my research. Thank you Dr. Hong! I am also grateful for Vanderbilt Mass Spectrometry Research Center for providing data sets for this research. I appreciate the COMS Ph.D. program for providing support to my participation of conferences and organizing research seminars so that we can access cutting edge ideas and receive discussion opportunities with invited speakers. I thank Dr. Wu for discussing chapter 4. and giving me valuable comments. I also thank SIGMA Actuarial Consulting Group providing me internship opportunity so that I was able to access to captive insurance solvency rating problem. I especially thank Mr. Timothy Coomer for discussing captive solvency rating project and giving valuable suggestions.

ABSTRACT

The accumulating of big-data such as medical data and insurance data requires more advanced computational statistical data analysis methods. As an interdisciplinary computational science research, we study mathematical methods of multiresolution analysis (MRA), statistical techniques of Bayes classifiers and Markov Random Field (MRF), computing tools of pyramid imaging matching and Markov Chain Monte Carlo (MCMC) and develop new statistical computing schemes in the applications of Imaging Mass Spectrometry (IMS) proteomic data analysis and insurance solvency modeling.

IMS technique is an important and useful tool to discover biomarkers and detect early cancer. However, the high-dimensionality of IMS data makes IMS data processing a difficult task and the development of computational methods for IMS data analysis is lagging behind its technological progress. To overcome high-dimensionality difficulty in IMS data analysis, we propose the MRA method to reduce the dimensionality of IMS data. By transforming IMS data onto wavelet coefficients space and analyze it from low resolution scale to high resolution scale using the idea inspired by pyramid imaging matching technique, the computational complexity can be reduced, while important biomarkers are still selected. For better IMS classification results, we select feature variables from wavelet coefficients and use Bayes classifier to classify IMS pixels based on its feature variables. To incorporate spatial information of IMS data, we consider the Markovianity in cancer growth that the state (cancer or non-cancer) of a sample point (pixel) is highly determined by the configuration of its neighboring system and use MRF to incorporate spatial information of IMS data. This algorithm is implemented using MCMC sampling and the result is probabilistic which provides more information than a deterministic result. We also tested different

neighborhood definitions.

As another application of statistical computing techniques, we study insurance solvency modeling. Insurance solvency is one of the most important measurements of insurance companies' financial health. It is directly related to the financial security of an insurance company and the benefits of insurance policyholders. The current solvency prediction methods are more deterministic rather than probabilistic. However, the deterministic method can not provide information such as percentiles and probabilities as a probabilistic method provides. In this application, we design an innovating model to predict captive insurance solvency using a probabilistic method with Monte Carlo simulation. Based on a pre-built financial report for captive insurance, we simulate future losses according to loss distribution to predict solvency scores in coming years. We score solvency from 0 to 1. This solvency score measures the probability that any of the future Insurance Regulatory Information System (IRIS) ratios breaks its upper and lower bounds. These bounds can be defined by users according to their business situations.

The data experiment shows MRA methods in proteomic data analysis are able to select important biomarkers and also achieve a higher classification accuracy with less computation complexity. The data experiment for the MCMC-MRF method shows that the MCMC-MRF method can improve classification accuracy significantly. Also, the captive insurance solvency model designed in this research can be a useful tool for captive managers to use and give more probabilistic information than the traditional deterministic IRIS models.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

With the development of computer storage and data collecting techniques, the amount of data we are accumulating in different areas, such as Proteomics, insurance risk management and so on, is experiencing an explosive growth. In this big-data time, we need more advanced mathematical, statistical, and computational data analysis techniques to process the data we have in order to discover new knowledge. As an interdisciplinary computational science research project, we study Imaging Mass Spectrometry proteomics data analysis and captive insurance solvency modeling by using wavelet method, Multi-resolution analysis, Bayes classifier, Metropolis-Hasting algorithm [29], Monte-Carlo Markov Chain (MCMC), Insurance regulation information system (IRIS) ratios and Monte-Carlo Simulation.

1.1 Imaging Mass Spectrometry (IMS) data analysis

Imaging mass spectrometry (IMS) is a technique developed from mass spectrometry to visualize the spatial distribution of moieties such as proteins, peptides, metabolites and lipids ([30], [25]). Currently, IMS is one of the few biochemical technologies able to establish the spatial biochemical composition of a sample in the full molecular range [38]. It can be used to map biomolecules in biological tissues and has attracted a great deal of attention in the analyses of drug effects, screening of drugs, and support for medical diagnoses [35]. However, the development of computational methods for IMS is lagging behind its technological progress [42].

IMS data set can be treated as a hyper-spectral imaging type data cube, see Figure 1. The value at each entry of the IMS data cube shows the abundance of corresponding

molecule. For a fixed m/z value (mass-to-charge ratio) in an IMS data cube, the corresponding intensity values make up an image that shows the distribution of that specific biochemical component in sample associated with this m/z value. Also, for a fixed pixel in the image cube, there is a mass spectrum (MS) corresponding to this pixel.

The main tasks for IMS data analysis are biomarker selection and classification. A biomarker is a biological molecule found in blood, other body fluids, or tissues that is a sign of normal or abnormal processes, or of a condition or disease [27]. In IMS data analysis, one usually finds biomarkers in terms of m/z values associated with proteins or peptides. Current popular analysis methods for IMS data include Principle Component Analysis (PCA) ([40], [16]), Support Vector Machine (SVM) [14] and Clustering methods [11]. With the development of IMS techniques, the amount and resolution of IMS data has increased. This requires faster and more accurate data analysis algorithms.

In chapter 2, our first concern is to reduce the high dimensionality of IMS data. We use wavelet transform to achieve this goal. wavelet transform has multi-resolution property and the combination of low and high resolution coefficients can greatly reduce the high dimensionality of original data. When we search for biomarkers, we use the idea inspired by pyramid image matching. At the low resolution level we can detect the m/z intervals which contain potential biomarkers and at the higher resolution level we only search in these m/z intervals with more details. In this way, we can reduce IMS data dimensionality in searching biomarkers because at lower resolution levels we only select those m/z data intervals contain potential biomarkers. To do classification, we select feature variables from wavelet coefficients of IMS data. These features representing IMS data at different resolution levels will be more robust

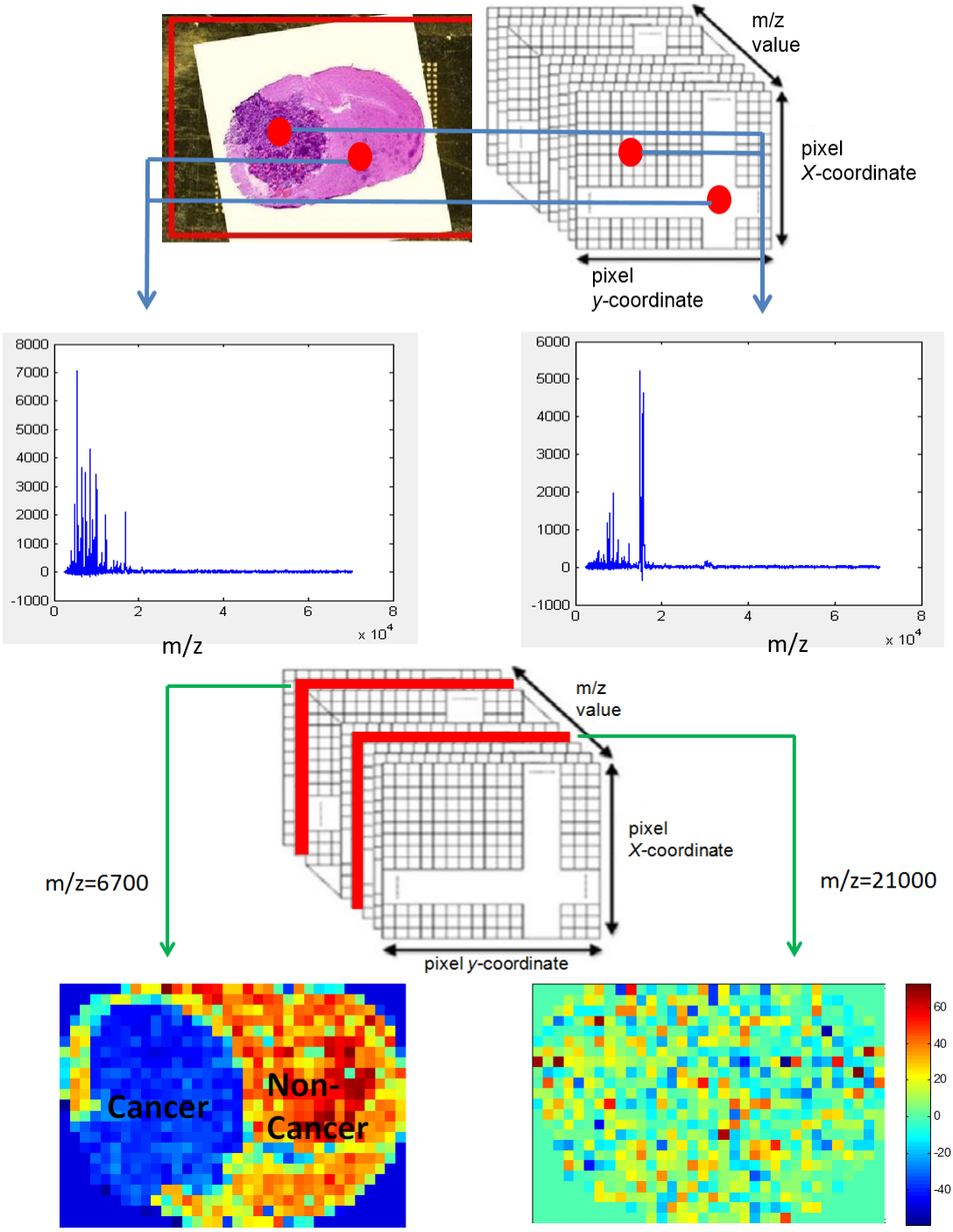

Figure 1: *An illustration of IMS data. (Top) For a fixed IMS pixel, there is a corresponding mass spectrum (MS). (Bottom) If an m/z value (mass-to-charge ratio) is fixed, the corresponding MS intensities for all pixels that make up an image shows the spatial intensity of that protein.*

than feature variable selected as original m/z data points because wavelet coefficients describe data at different resolution levels of data amplitudes. The feature variables we selected are those wavelet coefficients that are significantly different between cancer training data and non-cancer training data because they can classify cancer group and non-cancer group apart. We use Bayes classifier to classify for each IMS pixel according to its feature variables.

Chapter 3 is based on the fact that MRA method descriibed in chapter 2 without incorporating spatial information for IMS data processing. However, IMS data not only provide Mass Spectrum information but also image information, which is a type of spatial information. To further utilize IMS data information, one needs to consider the spatial relation between data pixels. Because the nature of tumor growth is spatially continuous, the class (cancer or non-cancer) of an IMS data pixel is closely related to its spatial neighboring pixels. We find Markov Random Field (MRF) is an ideal tool to describe such spatial relations, because in MRF the value of a pixel can be determined by the values of its neighboring pixels with a probability. The framework of chapter 3 is as follows: First, based on the result of MRA method, we consider the classification result from the MRA method presented in chapter 2 and denote it as an observation class $\mathbf{y}$. Our goal is to estimate its true class θ using observation class $\mathbf{y}$. This is a posterior estimation problem. Then, we use Metropolis-Hasting algorithm to implement this posterior estimation. Posterior is the product of a prior and a likelihood. We use Ising model (binary format of MRF) as the prior to incorporate spatial information and the posterior probability $P(\mathbf{y}|\theta)$ in training data as the likelihood. By examining different types of neighboring systems to describe different spatial impact mechanisms in tumor growth, we found that in IMS data classification, an 8-points neighboring system works better than 4-points neighboring

system and the higher-order neighboring works better than the 1st order neighboring.

1.2 Captive insurance solvency prediction problem

Continuing with the stochastic simulation idea in MCMC-MRF in menthiond in previous subsection, we choose captive insurance solvency prediction as another application topic, where stochastic Monte Carlo simulation and model design technique will be applied. This research is motivated by my internship experience in SIGMA Actuarial Consulting Group, where I got opportunity to know captive solvency rating problem and its needs in insurance practice.

Solvency is the ability of a company to meet its long-term financial obligations. Usually, we say an insurance is solvent if its financial situation is healthy enough to be able to pay future claims. Captive insurance is an insurance that is wholly owned and controlled by its insureds. Captive insurance is non-profit. Solvency is key for the survival of captive insurance. In practice, a professional captive insurance manager usually manage dozens of small size captive funds. They need a good tool to measure their solvencies. In chapter 4, we apply Monte Carlo simulation based on captive insurance company's historical data distribution to develop a predictive model for captive insurance solvency.

Existing popular methods that evaluate insurance solvency include Insurance Regulatory Information System (IRIS) [4], Financial Analysis Solvency Tools (FAST), solvency II. These methods have shortcomings when it comes to captive insurance solvency. First, they are not specially designed for captive insurance, therefore can not predict captive insurance well. Second, current popular solvency evaluation methods are deterministic, for example, in IRIS method, a deterministic rating will be given without probability distribution. Third, these popular methods focus on cur-

rent stage solvency evaluation, not future prediction. Knowing future solvency trend provides captive insurance manager confidence about decision makings. Last, IRIS based method have been used since 1970s for solvency regulation, but its range from the lower bound to the upper bound are fixed, which is not suitable for every situation. Based on these shortages, we aim to design a model that gives a probabilistic result instead of deterministic one, dynamically predict future solvency scores instead of current static solvency evaluation, and allows the user to define IRIS upper lower bounds with flexibility instead of fixed IRIS upper and lower bounds.

Retention is the maximum risk an insurance organization is willing to take. The loss above the retention cap will be covered by reinsurance companies. For captive insurance managers, choosing the right retention level is important for future captive insurance solvency. If the retention is too low, then captive insurance need to pay high premiums to reinsurance, and future solvency score will be lower because captive insurance may not have enough fund to pay claims. If the retention is too high, then future solvency score can also be low because of long-tail effect in loss distribution. Based on these discussions, when we design the model, we associate retention with future solvency in consideration. In our model, we allow the user to select several retention levels they are interested in. Under each retention level we use Monte Carlo simulation to approximate future years' solvency scores. The result will be a matrix of solvency scores under different retention levels for different future years. With this result, the user is able to not only estimate future years' solvencies, but also compare different retention levels when making retention selection decision.

CHAPTER 2

MULTI-RESOLUTION ANALYSIS METHOD FOR IMS PROTEOMIC DATA BIOMARKER SELECTION AND CLASSIFICATION

Even though imaging mass spectrometry (IMS) technique is evolving rapidly, its data analysis capability lags behind. Especially with the improving of IMS data resolution, faster and more accurate data analysis algorithms are required. To meet such challenges in IMS data analysis, an effective and efficient algorithm for IMS data biomarker selection and classification using multi-resolution (wavelet) analysis method is proposed. We first applied wavelet transform [2] to IMS data denoising. The idea of wavelet pyramid method for image matching was then applied for biomarker selection, in which Jaccard similarity is used to measure the similarity of wavelet coefficients. Last, the Naive Bayes classifier [26] was used for classification based on feature vectors in terms of wavelet coefficients. Performance of the algorithm was evaluated in real data applications. Experimental results show that this multi-resolution method has advantages of fast computing and accuracy.

2.1 Motivation of this study

To meet challenges and needs in IMS data analysis, we have developed a mathematical and statistical model using the wavelet method for IMS cancer data analysis in biomarker selection and classification. The motivations for introducing the wavelet method to IMS data analysis are based on the following. First, the multi-resolution property of wavelets allows us to analyze IMS data on different resolution levels to obtain accurate results with less computation. The low resolution analysis can decrease analysis time because we can represent the whole data set with less wavelet

coefficients. Also, over-fitting can be reduced and noise can be lessened at low resolution analysis. The high resolution analysis can improve biomarker selection accuracy by analyzing data without losing detailed information. The wavelet method combines the aforementioned advantages of low and high resolution analysis together. Second, wavelet pyramid method in image matching [49] can be applied to identify biomarkers from low resolution to high resolution. Note that in cancer IMS studies, biomarkers are identified by comparing cancer IMS data and non-cancer IMS data. This process is similar to image matching. Hence, the wavelet method, which is essential in the pyramid imaging matching process, can also be expected to be useful in IMS data analysis. Third, wavelet transform can reduce the high dimensionality of IMS data. By transforming IMS data to wavelet coefficient space, we can represent IMS data sparsely at low resolution while still keeping the necessary detail information at high resolution. Last, only few studies have applied the wavelet method to IMS data analysis, though there are some work in mass spectrometry (MS) applying the wavelet method, for example work in [10]. We would like to apply the wavelet method to IMS data to determine if this method has some advantages compares with other current methods. Successful application in MS data analysis would show that the wavelet method can also be promising in dealing with IMS data.

The main contributions of this research include: combining the advantages of both low resolution and high resolution analysis in IMS data processing to achieve fast and accurate biomarker selection algorithm; providing a new perspective of IMS data by transforming the original IMS data to wavelet coefficient space and can find those patterns not easy to see in original data; introducing probabilistic calcification instead of traditional binary classification to obtain not only a classification result but also a confidence level.

The remaining of this chapter is organized in the following manner: Section 2.2, we propose a wavelet based de-noise algorithm for IMS data; Section 2.3, a wavelet based IMS biomarker selection algorithm using the idea of pyramid matching is proposed; Section 2.4, we propose an IMS data classification algorithm using feature variables selected from wavelet coefficients combined with Naive Bayes classifier.

2.2 Wavelet method for IMS data de-noising

Before we start biomarker selection, we need to pre-process IMS data by data de-noising. Denoising is based on the wavelet method [9]. Figure 2 presents an example of wavelet coefficients (discrete Haar wavelet coefficients) for a pixel in the IMS data, with false color representation of the coefficient value. The coefficients on the top of Figure 2 are low frequency wavelet coefficients, which describe the data on a large scale and show the outline. The coefficients on the bottom of Figure 2 are high frequency wavelet coefficients, which describe the data on a smaller scale and show the details. In N-level decomposition, one signal is decomposed into N detailed components and one approximation component. We can de-noise the signal by keeping the large coefficients while setting the small coefficients to be 0 based on a threshold level. By applying this method, we can remove the majority of the noise. Here are the basic steps for de-noising,

- **Step 1**: Decompose the signal f. Compute the wavelet decomposition of the signal f from resolution level 1 to N.
- **Step 2**: Threshold detail coefficients. For each level from 1 to N, set the detail coefficients less than threshold to be 0. In illustrative Figure 3, the yellow broken line is the threshold level.

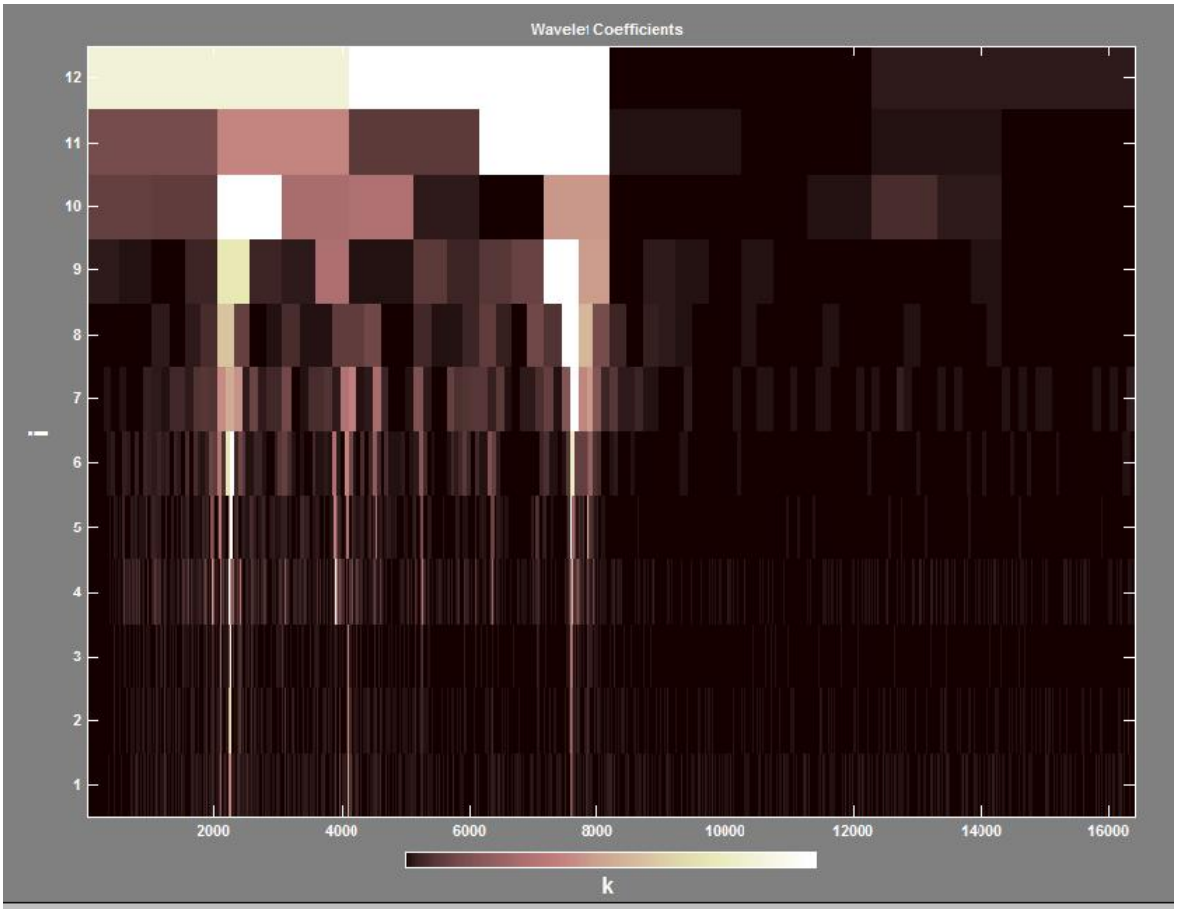

Figure 2: *An example of wavelet coefficients of the mass spectrometry for one IMS pixel. Those top coefficients cover wide intervals are low resolution coefficients which describe data on large and rough scale. Those bottom coefficients cover narrow intervals are high resolution coefficients which describe data on small and precise scale.*

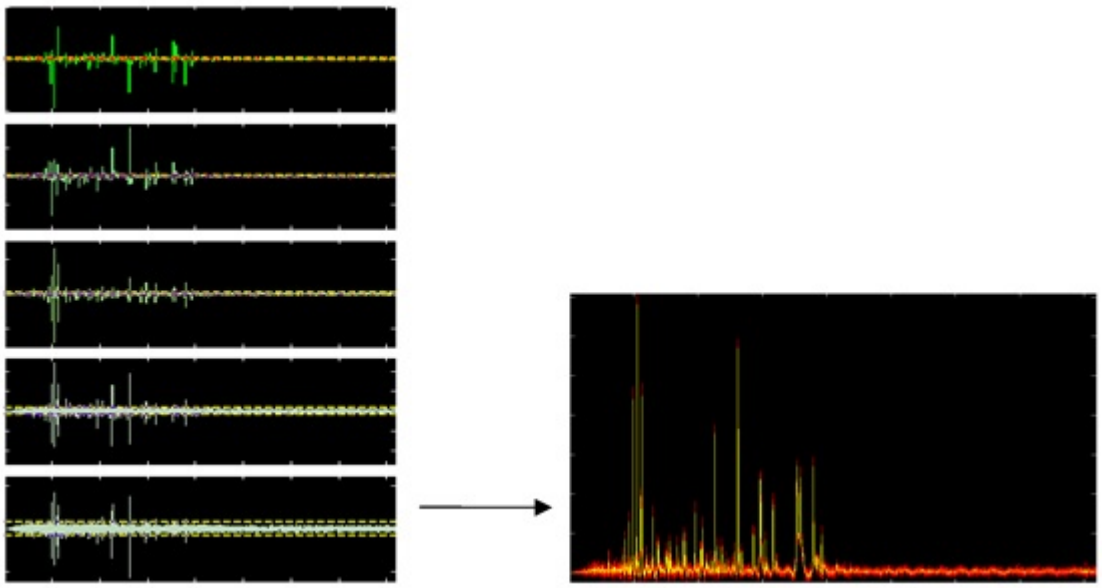

Figure 3: *Illustration of IMS data de-noising by using the wavelet method. (Left) Apply wavelet transform to mass spectrum. For each resolution level, set a threshold line, the yellow broken lines as shown in left figure. Only keep large coefficients, greater than threshold, and set small coefficients smaller than threshold to zero. Then apply wavelet inverse transform to the modified coefficients and the result is de-noised data. (Right) The yellow data is de-noised. The red data is original data.*

- **Step 3**: Reconstruction of the signal. Compute wavelet reconstruction using the modified coefficients to recover the de-noised signal.

We apply this process to all the original IMS data to obtain the de-noised IMS data. All sequent analysis is based on the de-noised data. See [10] for more details on MS data preprocessing.

2.3 Biomarker selection

2.3.1 Algorithm idea

Biomarkers in IMS cancer studies are proteins whose intensities differ between cancer area tissue and non-cancer area tissue, therefore allowing them to be used as markers to tell the cancer status of the specimen. The biomarker selection problem in IMS data analysis is very similar to the image matching problem. In image matching, people find objects that are similar between images using wavelet pyramid method [48]. Here in IMS data analysis, we find those proteins whose intensities are different between sample data. We just need to define a variable to measure the difference instead of similarity, and biomarker selection problem can be handled in a similar way as wavelet pyramid method applied in image matching.

The basic idea for image matching based on wavelet pyramid multi-resolution analysis can be briefly described as following [49].

- **Step 1**: Compare sub-images at the low resolution level.
- **Step 2**: Amplify the matched area and compare images at higher resolution level.

- **Step 3**: Repeat step 2 until to full resolution to find out the matched object in compared two images and purpose of image matching achieved.

We apply this idea to wavelet multi-resolution IMS cancer data analysis to select biomarkers.

- **Step 1**: Compare cancer data and non-cancer data at low resolution level to select the m/z ranges whose intensities are statistically significantly different between cancer and non-cancer data. Those selected m/z ranges can be treated as "suspicious" m/z data ranges because their data difference in statistics may be caused by the existence of cancer biomarkers.

- **Step 2**: Increase the resolution level of those suspicious m/z data ranges to compare them between cancer and non-cancer data at a higher resolution and select those smaller suspicious m/z data sub-ranges with intensity statistically different between two data groups.

- **Step 3**: Repeat step 2 until to full resolution level. Those m/z values selected at full resolution level are the biomarkers we selected from this algorithm.

2.3.2 Algorithm detail

In this study, we use two IMS data sets as shown in Figure 4. They are generated from the Vanderbilt Mass Spectrometry Research Center using two different mouse brains from same species implanted with the same type of cancer cells. Data set-1 has resolution 24*34, which contains 816 MS pixels. Data set-2 has resolution 64*44, which contains 2816 MS pixels. We use one data set as training data and another as test data. We illustrate this biomarker selection algorithm using the data experiment

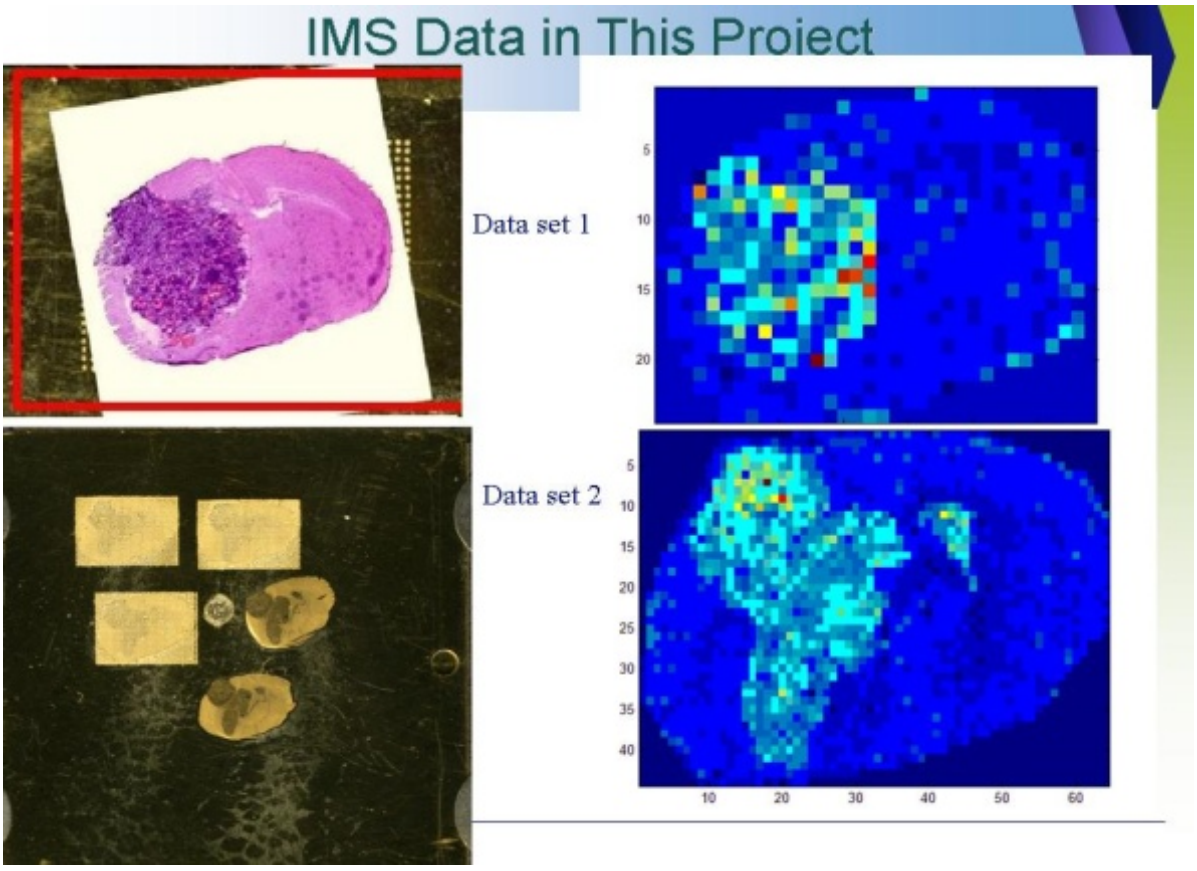

Figure 4: *IMS data sets used in this study. (Top left) The brain tissue slice picture where the IMS data set 1 is generated from. (Top right) IMS data set 1 snap for a specific m/z channel. (Bottom left) The brain tissue slice picture where the IMS data set 2 is generated from. (Bottom right) IMS data set 2 snap for a specific m/z channel.*

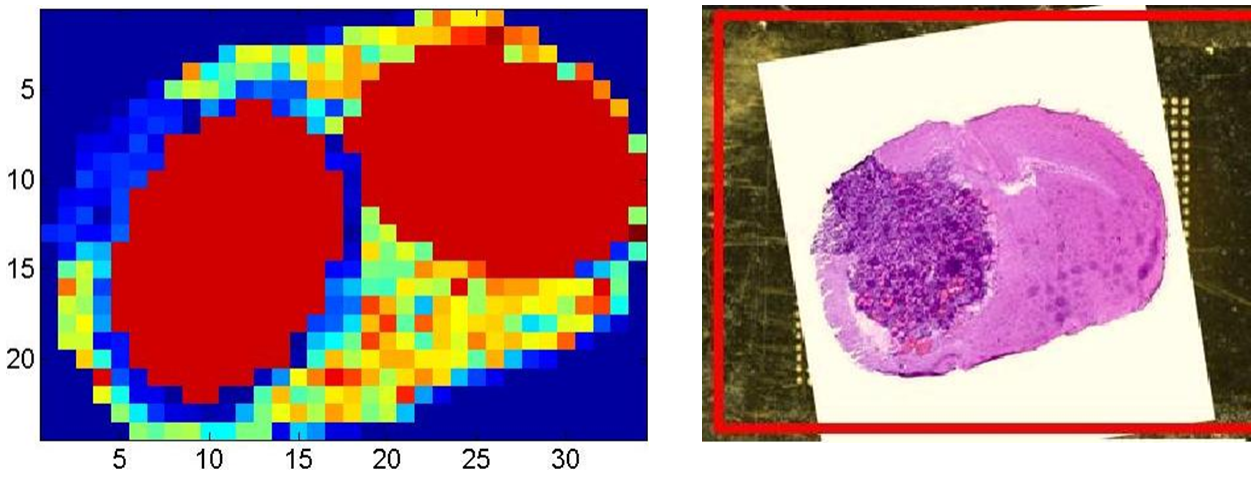

Figure 5: *Training data set. (Left) The left marked round area is the selected cancer pixels and right marked round area is the selected non-cancer pixels; (Right) Slide picture of the mouse brain with a tumor where the data in left was generated.*

we did on data set-1. From data set-1 (Figure 4), we select two round IMS data areas with radius of $r = 6$ (6 pixels distance) which are symmetrical to each other by the symmetric line of the mouse brain slice. Because of their symmetrical positions, these two areas contain the very same biological structure so that we can better emphasize the differentiation of cancer and non-cancer in IMS intensities. The data in these two selected areas are used as training data. Each round area contains 109 IMS pixels, i.e. 109 mass spectra (MS).

For each selected training MS, compute its 12-level discrete wavelet decomposition. Figure 6 shows wavelet coefficients space for a cancer training pixel MS and a non-cancer training pixel MS. Applying wavelet transform [2]to each mass spectrum turns a spectrum data cube into a wavelet coefficient data cube. Originally, each pixel is associated with a mass spectrum, but after transformation, each pixel is associated

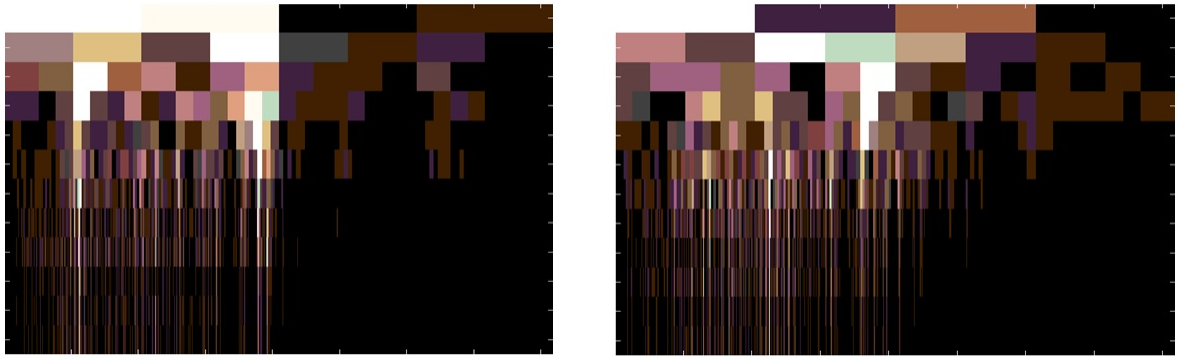

Figure 6: *Wavelet coefficients space for a cancer MS and a non-cancer MS. (Left) Wavelet coefficients space for a cancer MS. (Right) Wavelet coefficients space for a non-cancer MS. For each cancer IMS pixel, there is a corresponding wavelet coefficients space like left picture. For each non-cancer IMS pixel, there is a corresponding wavelet coefficients space like right picture. The difference table as shown in Figure 10 is computed by comparing statistical difference of wavelet coefficients from cancer MS group and non-cancer MS group.*

with a wavelet coefficient vector space. Since the MS intensities of cancer biomarkers vary dramatically from cancer pixels to non-cancer pixels and wavelet coefficient is a description of MS on wavelet space, we can to locate biomarkers by measuring the difference between cancer wavelet coefficients and non-cancer wavelet coefficients. The difference of wavelet coefficients can indicate the difference between cancer MS and non-cancer MS at different resolution levels. Analyzing it from low resolution to high resolution, we can quickly locate the biomarkers. This idea was inspired by the wavelet pyramid method in image matching [49].

We measure the difference using a method analogous to Jaccard similarity [36],

[37]. It measures difference by measuring how much two groups data are overlapped (Figure 7). Statistically, the more two group of data overlap, the more different they are (Figure 8).

The following is the mathematical definition of the difference described above. We denote the set of selected training cancer pixels as S_c, the set of selected training non-cancer pixels as S_n. For a fixed wavelet resolution level $j \in J$ (in our experimental data we define $J = \{1, 2, ..., 12\}$) , a fixed wavelet window position $k \in K$ (in our experimental data we define $K = \{1, 2, ..., 2^{j+2}\}$) and a selected pixel $i \in S_c$ or $i \in S_n$, we denote the corresponding cancer wavelet coefficient as $c^c_{j,k,i}$ and its empirical distribution along the selected training cancer pixels set S_c as $f^c_{j,k}$, and the corresponding non-cancer wavelet coefficients group as $c^n_{j,k,i}$ and its empirical distribution along the selected training non-cancer pixels set S_n as $f^c_{j,k}$. The similarity of wavelet coefficients between cancer data group $\{c^c_{j,k,i}\}_{i \in S_c}$ and non-cancer data group $\{c^n_{j,k,i}\}_{i \in S_c}$ is defined as

$$S_{j,k} = \int_{-\infty}^{+\infty} \min\{f^c_{j,k}(x), f^n_{j,k}(x)\}dx \tag{1}$$

The intuitional meaning of $S_{j,k}$ is the overlapping area of the histogram of the two groups to be compared. We can approximately calculate this integral using the histogram of the empirical distribution. Finally, we define the difference between the wavelet coefficients $\{c^c_{j,k,i}\}_{i \in S_c}$ and $\{c^n_{j,k,i}\}_{i \in S_n}$ as:

$$d_{j,k} = 1 - S_{j,k} \tag{2}$$

An illustration of $d_{j,k}$ is given in Figure 9.

We define $D = \{d_{j,k}\}_{j \in J, k \in K}$, the difference table that describes the difference of the corresponding wavelet coefficients between cancer group and non-cancer group at different wavelet resolution level as the false-color map shown in Figure 10. These

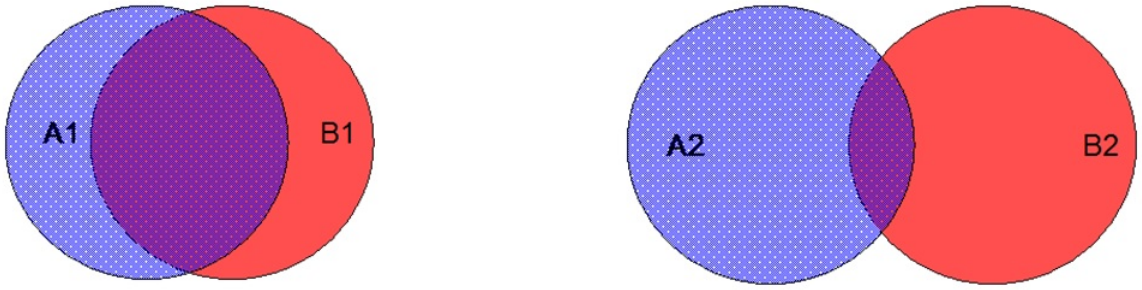

Figure 7: *An example of Jaccard similarity. According to definition of Jaccard similarity* $J(A,B) = \frac{|A \cap B|}{A \cup B}$*, the similarity between A1 and B1 are greater than the difference between A2 and B2. Hence the difference between A1 and B1 is smaller than the difference between A2 and B2.*

large difference value areas contain potential biomarkers. "Large difference" means the corresponding $d_{j,k}$ greater than the threshold, i.e., there is statistically significant difference exist here between two groups of data. Similar to the wavelet pyramid method applied in image matching, we take advantage of the multi-resolution property of wavelet analysis to locate the biomarkers from low resolution wavelet coefficients to high resolution wavelet coefficients using the idea we described in Section 2.3.1. We analyze difference table D from lower resolution level $j = 8$ to highest resolution level $j = 12$. From level $j = 8$ to level $j = 12$, if $d_{j,k}$ is greater than the threshold (we set it as 0.85 for the experiment data we use here), that means the contrast between cancer MS and non-cancer MS on the corresponding chemical protein is noticeable at the jth wavelet resolution level as well as the kth wavelet window position. Thus, there is a good chance that biomarkers exist in the corresponding m/z intervals. We then further analyze the wavelet coefficients on the next higher wavelet resolution (i.e.

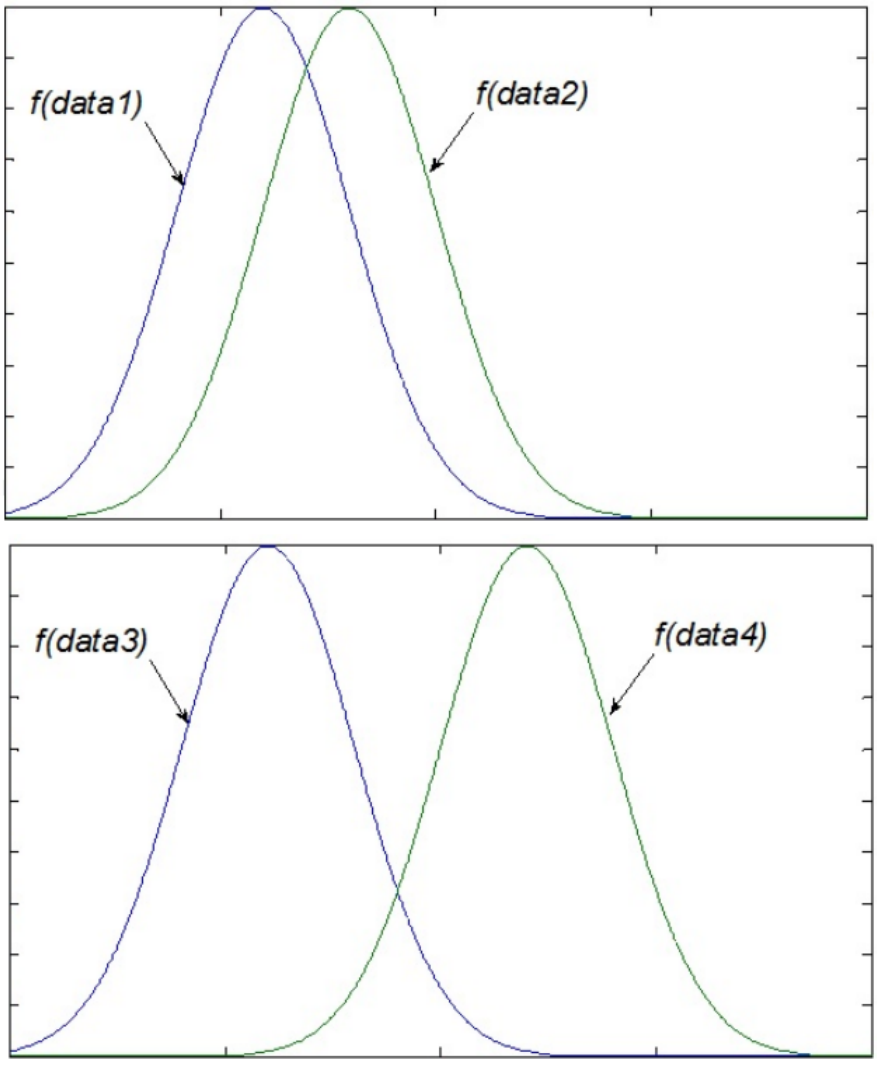

Figure 8: *An illustration of statistical Jaccard similarity. Statistically, according to definition of difference defined in formula (1) and (2), the difference between data*1 *and data*2 *(Top) are smaller than the difference between data*3 *and data*4 *(Bottom), since data*1 *and data*2 *have more overlap values. f is the distribution of data.*

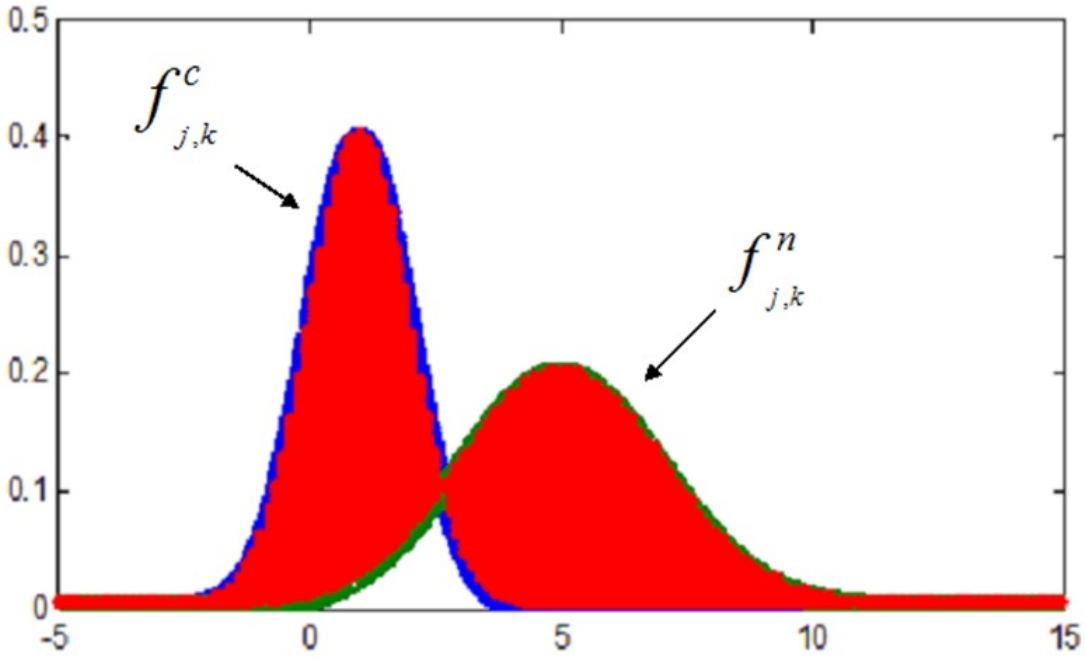

Figure 9: *Illustration of the definition of difference $d_{j,k}$, area of red shadow. Since Jaccard distance measure similarity, the difference (un-similarity) should be the complement value of Jaccard distance.*

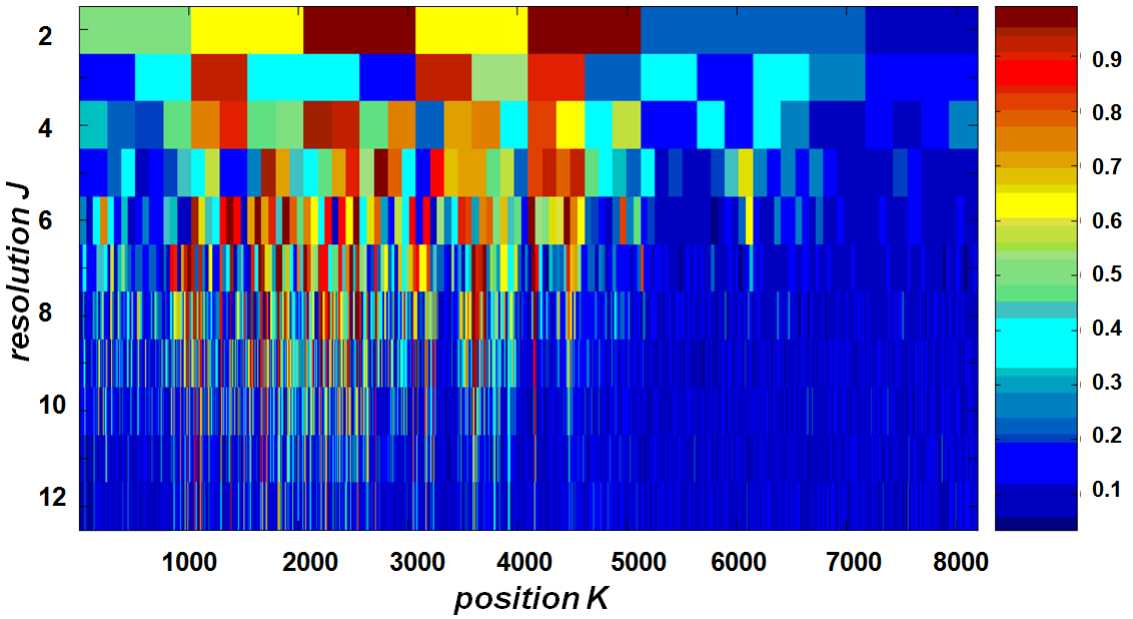

Figure 10: *Difference table* $D = \{d_{j,k}\}_{j \in J, k \in K}$. *The color represents value. It measures the difference between cancer data and non-cancer data. The MRA (multi-resolution analysis) biomarkers selection from low resolution to high resolution is done based on this table.*

amplify it) level in the same wavelet window position. Otherwise if $d_{j,k}$ is not greater than the threshold, we stop and shift our analysis to the adjacent wavelet window position $k+1$. We repeat this process until we reach the highest resolution level, level $j = 12$ in the data we used, and determine the specific m/z value whose intensities difference are greater than the threshold. These m/z values selected at highest level $j = 12$ are the m/z values of the biomarkers selected by this algorithm. The threshold can be changed in order to select the corresponding number of biomarkers. Algorithm 1 shows this algorithm's pseudo-code.

Table 1 is the list of the m/z values of the biomarkers selected by this multi-

Algorithm 1 MRA biomarker selection for IMS data

Input: IMS selected data

Output: biomarkers

Compute difference table D ▷ use formula (1), (2)
$lowest\_resolution\_level = 8$ ▷ selected by user
$highest\_resolution\_level = 12$ ▷ selected by user
$desired\_biomarker\_number = 30$ ▷ selected by user
$initial\_threshold = 0.7$ ▷ selected by user
$decrement\_size = 0.01$ ▷ selected by user
$threshold = initial\_threshold$
$biomarkers = []$ ▷ to record biomarkers
while length($biomarkers$) > $desired\_biomarker\_number$ **do**
    for $j = lowest\_resolution\_level \rightarrow highest\_resolution\_level$ **do**
        for $k = 1 \rightarrow 2^{(j+1)}$ **do**
            if $(2^{j+1} * (k-1) + 1, 2^{j+1} * k)$ is in marked interval **then**
                if $D_{j,k} > threshold$ **then**
                    Mark $(2^{j+1} * (k-1) + 1, 2^{j+1} * k)$ to be marked interval
                end if
                if $j == highest\_resolution\_level$ **then**
                    $biomarkers = [biomarkers, 2^{j+1} * (k-1) + 1]$
                end if
            end if
        end for
    end for
    $threshold = threshold - Decrement\_size$
end while

resolution analysis method (MRA) algorithm described above along with the lists of biomarkers selected by some other popular methods for IMS data biomarker selection.

According to the biological study [23], the biomarkers whose $m/z = 6700$ and $m/z = 8380$ are widely confirmed as the key cancer biomarkers for GL26 IMS data sets that we used in this research. Compared with other methods, the MRA method discovered both biomarkers while including a relatively shorter biomarkers list. Figure 11 is the intensity distributions of these two biomarkers ($m/z = 6702.2$, $m/z = 8374.9$). Its intensity differences between cancer and non-cancer area are significant at this two m/z channels. These are biomarkers that have already been proven in a previous cancer study [22]. Two such biomarkers include cytochrome c oxidase copper chaperone and cytochrome c oxidase subunit 6c. They are related to the growth, division, and expansion of tumor cells. These facts support the results of this MRA algorithm.

Additionally, based on our computing experiment, we determined that the MRA method for IMS data biomarker selection has high algorithm computing speed. We tested the algorithm speed using MATLAB 7.0 installed on a DELL laptop to run EN4IMS proposed by D. Hong and F. Zhang in 2010 [18] and the MRA method discussed in this research with the same data set (data set-1). Here is the hardware information of the computer used for this test: `Intel(R) Core(TM)2 Duo CPU T7250 @2.00GHZ 778 MHz, 2.00 GB.` The test showed that the CPU time for EN4IMS to select biomarkers is 49.265 seconds. The CPU time for MRA method is only 26.562 seconds. The shorter running time of MRA method comes from the advantage of multi-resolution. In MRA method, we saved computing time by avoiding analyzing every m/z data point one by one. We exclude those m/z intervals whose data difference is not as large as the threshold we set. The amount of m/z data points

Table 1: *A comparison of biomarker lists generated by the Multi-Resolution Analysis Method (MRA) and by currently major methods [17] for IMS data analysis. MRA method generates a shorter list while still contains major biomarkers ($m/z = 6702$, $m/z = 8375$).*

EN4IMS list	SAM list	EN list	PCA list	MRA list
4664	2791 3434 8337	4476 13562	4934 8567	4599
4667	3010 3764 8366	4664 14327	4936 10257	4607
4670	3056 4011 8380	4670 14336	4937 10259	4757
4812	3734 4076 8395	4812 14343	4938 10261	4759
5446	3800 4271 8492	4884 14781	4939 10263	4762
5753	3920 4538 8672	5425 14786	4960 14969	4767
5754	4206 4566 8945	5429 14805	4962 14971	4770
5756	4341 4665 8982	5446	4963 14974	4892
5757	4605 4676 9327	5753	4964 14976	4895
6165	4734 4899 9343	5754	4966 14979	4903
6702	4767 5106 9531	5756	5439 14981	5438
6706	4921 5120 9602	6165	5441 14983	5446
7799	4936 5428 9619	6702	5442 14986	5449
8019	4964 5444 10238	6706	5444 15603	5714
8024	4981 5707 10267	6794	5445 15606	6244
8384	5001 5753 10466	7799	5446 15608	6248
8386	5024 6166 10662	8019	5448 15611	6312
9344	5170 6186 12434	8024	5449 15613	6702
10172	6225 6251 13560	8028	5451 15616	6705
10261	7706 6310 14525	8384	6571 15618	8375
10263	8420 6574	8386	6572 15620	8400
10265	8603 6700	8495	6574 15623	8403
10267	8709 6719	8524	6575 15625	8572
10282	8747 6780	9344	6577 16780	8978
10366	9062 7099	9553	7749 16782	9332
10374	9736 7118	10172	7751 16785	9613
10825	9956 7297	10261	7752 16787	9616
10949	10167 7315	10263	7792	9624
13562	10952 7338	10267	7794	11632
14336	11388 7357	10282	7795	
14343	11640 7751	10366	7797	
14781	12203 7776	10374	8560	
14786	14865 7795	10811	8562	
14805	14927 8025	10825	8564	
	14978 8107	10949	8566	

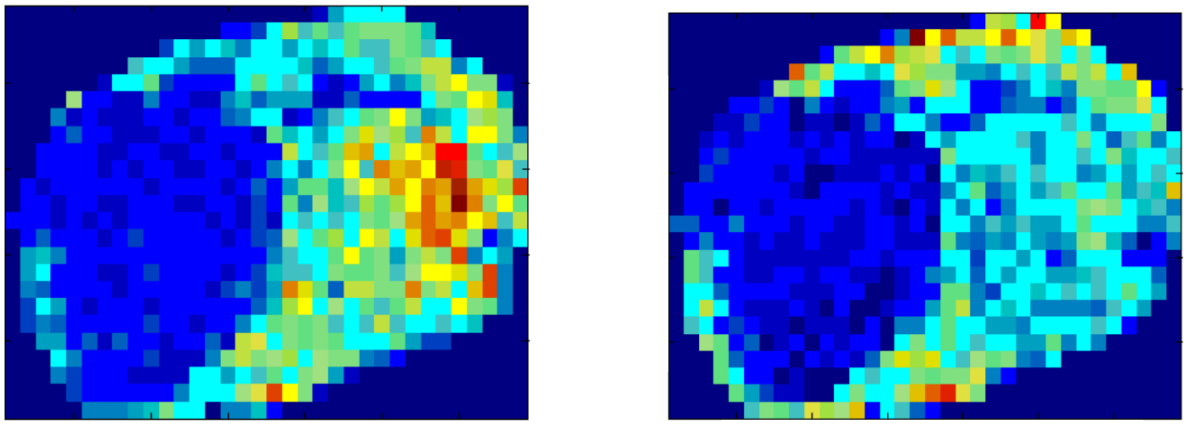

Figure 11: *Two important biomarkers selected. (Left) Intensity distribution for biomarker of* $m/z = 6702.2$ *selected by MRA method. (Right) Intensity distribution for biomarker of* $m/z = 8374.9$ *selected by MRA method. These two biomarkers have been confirmed by biology study and also selected out by MRA method.*

that still remain at higher resolution levels are much less than the total amount of whole m/z data points. In this way, the amount of data we need to analyze is reduced. Thus, MRA method can achieve high computing efficiency in IMS data analysis.

2.4 Classification

In this section, we will use the Naive Bayes classifier [12] to do classification on wavelet coefficient space. Bayes classifier is an appropriate tool to deal with IMS data classification problem. It classifies data based on its probabilities in each class and chooses the class with the highest probability to be data's class. Compared with non-probability classification [28] method, Bayes classifier not only tells us a classification result but also the probability to be classified in each class, so we can

measure the confidence of results. This advantage is also helpful if we want know the cancer stage or the degree of cancer for each pixel, because more serious degree of cancer corresponds to higher probability to be classified as cancer class in Bayes classifier.

As shown in Figure 12, we use data set-1 as training data and data set-2 as test data for classification study. We train a model from data set-1 and test the trained model using data set-2 to see its performance. Normalization is a necessary step before we start classification, this is because the scale in training data and the scale in test data are different. For normalization purposes, we divide each mass spectrum with its average intensity. After normalization, the scale will be the same in all data sets.

Classification is based on feature variables. We select 10 feature variables from the wavelet coefficients of each pixel's mass spectrum. These feature variables are selected from training data's wavelet coefficients whose values are significantly different between cancer data group and non-cancer data group. We can identify them using the difference table D as shown in Figure 10. Those large entries $d_{j,k}$ in the difference table D correspond to the wavelet coefficients whose difference is large between the cancer group and non-cancer group. Therefore, we chose those wavelet coefficients with large $d_{j,k}$ in D as feature variables. For the data sets used in this study, we chose wavelet coefficients from level 6 to level 12 whose difference is greater than the difference threshold we set. We can ignore the detail coefficients from level 1 to level 5, since most of the noise exists in high frequency coefficients, if our data contains too much detail, the amount of noise will influence the classification accuracy. With the threshold we set, 10 feature variables are selected from wavelet coefficients space for the mass spectrum of each pixel in cancer and non-cancer training data. These 10

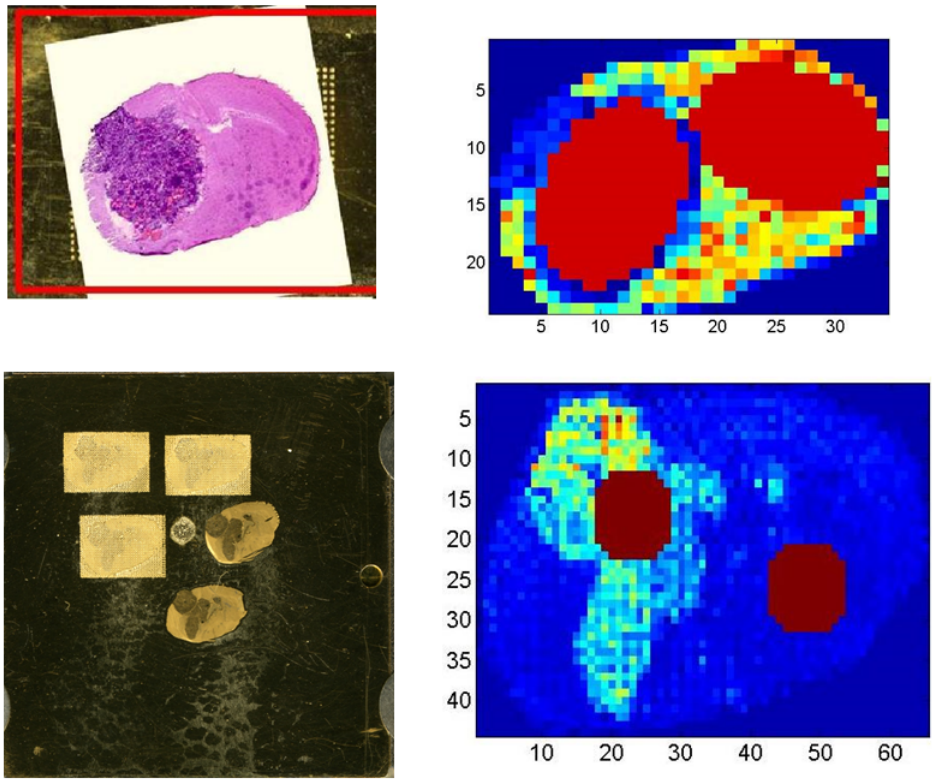

Figure 12: *Training date set and test data set. We use two different IMS data set to train and test model. (Top left) Picture of the mouse brain tissue slice from which training IMS data were generated. The dark area is cancer area. (Top right) A snapshot of training IMS data set. The round red areas in left side and right side are cancer training data and non-cancer training data respectively. Resolution for this IMS data set is* 24×34 *pixels. (Bottom left) Picture of the mouse brain tissue slice where test IMS data was generated from. The dark areas on brain slice are cancer areas. (Bottom right) A snapshot of test IMS data set. The round red areas in left side and right side are cancer test data and non-cancer test data respectively. Resolution for this data set is* 44×64 *pixels.*

feature variables, as components, form a feature vector, denoted by $\boldsymbol{e}$, corresponding to a pixel MS. The classification process of each pixel is based on its feature vector which made up by its selected feature variables.

In the next step, we use the Naive Bayes classifier to classify the cancer and non-cancer pixels based on pixels feature vector. We denote the probability that an unknown testing pixel i, which has a feature vector $\boldsymbol{X}$, is a cancer pixel as

$$P(i \in C|\boldsymbol{e} = \boldsymbol{X}),$$

where i denotes the testing pixel, C denotes the set of cancer pixels, $\boldsymbol{e}$ denotes the feature vector of this testing pixel, $\boldsymbol{X}$ denotes the value of its feature vector. Similarly, the probability that an unknown testing pixel i, which has a feature vector $\boldsymbol{X}$, is a non-cancer pixel is defined as

$$P(i \in N_c|\boldsymbol{e} = \boldsymbol{X}),$$

where N_c denotes the set of non-cancer pixels. If

$$P(i \in C|\boldsymbol{e} = \boldsymbol{X}) > P(i \in N_c|\boldsymbol{e} = \boldsymbol{X}),$$

the chance of this testing pixel being in the cancer group is greater than its chance in the non-cancer group. If this is the case, then we classify this pixel as a cancer pixel. Otherwise, we classify it as a non-cancer pixel. We can calculate the above conditional probabilities using Bayes formula:

$$P(i \in C|\boldsymbol{e} = \boldsymbol{X}) = \frac{P(\boldsymbol{e} = \boldsymbol{X}|i \in C)P(i \in C)}{P(\boldsymbol{e} = \boldsymbol{X})} \tag{3}$$

$$P(i \in N_c|\boldsymbol{e} = \boldsymbol{X}) = \frac{P(\boldsymbol{e} = \boldsymbol{X}|i \in N_c)P(i \in N_c)}{P(\boldsymbol{e} = \boldsymbol{X})} \tag{4}$$

Then we compare these two probabilities and $P(\boldsymbol{e} = \boldsymbol{X})$ can be canceled, thus leading to:

$$\frac{P(i \in C|\boldsymbol{e} = \boldsymbol{X})}{P(i \in N_c|\boldsymbol{e} = \boldsymbol{X})} = \frac{P(\boldsymbol{e} = \boldsymbol{X}|i \in C)P(i \in C)}{P(\boldsymbol{e} = \boldsymbol{X}|i \in N_c)P(i \in N_c)} \tag{5}$$

If $\frac{P(i \in C|\boldsymbol{e}=\boldsymbol{X})}{P(i \in N_c|\boldsymbol{e}=\boldsymbol{X})} > 1$, that means $P(i \in C|\boldsymbol{e} = \boldsymbol{X})$ is greater than $P(i \in N_c|\boldsymbol{e} = \boldsymbol{X})$. We then classify the testing pixel as cancer pixel, since the chance being cancer is larger than the chance being non-cancer. Otherwise, we classify this testing pixel as non-cancer pixel. Therefore, the classification criterion can be defined as:

$$\frac{P(\boldsymbol{e} = \boldsymbol{X}|i \in C)P(i \in C)}{P(\boldsymbol{e} = \boldsymbol{X}|i \in N_c)P(i \in N_c)} > 1 \iff i \in C \tag{6}$$

$$\frac{P(\boldsymbol{e} = \boldsymbol{X}|i \in C)P(i \in C)}{P(\boldsymbol{e} = \boldsymbol{X}|i \in N_c)P(i \in N_c)} < 1 \iff i \in N_c \tag{7}$$

To calculate values in formula (6) and formula (7), we need to determine the likelihood probability $P(\boldsymbol{e} = \boldsymbol{X}|i)$ and find prior probability $P(i)$. Figure 13 shows the distributions of cancer feature variables as well as non-cancer feature variables. They are mostly in normal distributions. Since the feature vector is made up by these 10 feature variables, we can assume that the distribution of feature vector in cancer or in non-cancer is a 10-dimensional normal distribution.

Thus the likelihood $P(\boldsymbol{e} = \boldsymbol{X}|i \in C)$ and $P(\boldsymbol{e} = \boldsymbol{X}|i \in N_c)$, which is a probability density, can be calculated by a 10-dimensional normal distribution. The mean value for the cancer data group can be obtained by computing the average value of the feature vectors of all cancer pixels in training data. The standard deviation for the cancer group can be obtained by computing the covariance matrix of the feature vectors of all cancer pixels in the training data. The same idea applies for the non-cancer group. Then, the likelihood for the testing feature vector X can be determined by the remaining of the distributions,

$$(\boldsymbol{e}|i \in C) \backsim N_{10}(\mu_{cancer}, \Sigma_{cancer}) \tag{8}$$

$$(\boldsymbol{e}|i \in N_c) \backsim N_{10}(\mu_{noncancer}, \Sigma_{noncancer}) \tag{9}$$

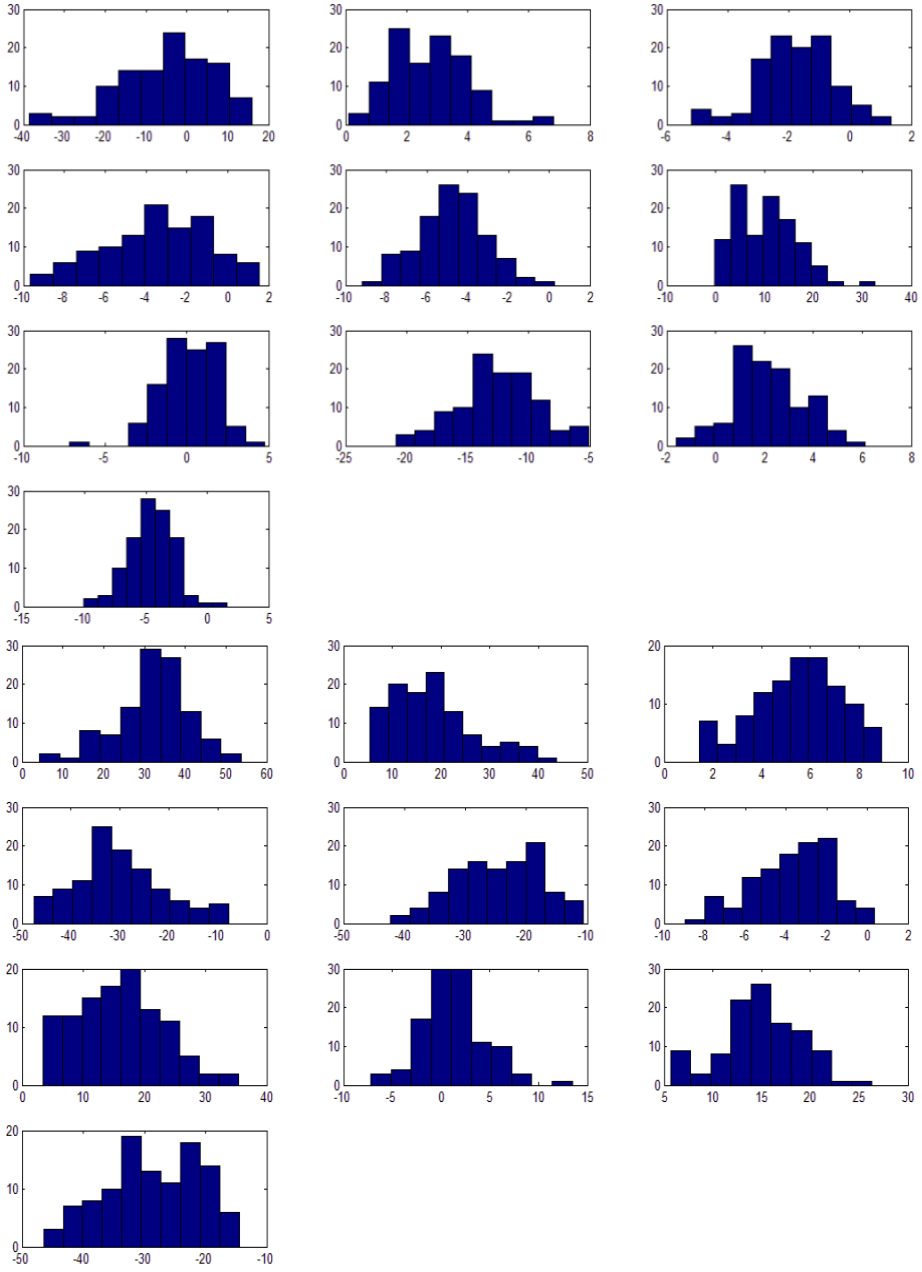

Figure 13: *Likelihood estimation for Bayes classifier. (Top 10 subfigures) Distribution of 10 selected feature variables from cancer data group. (Bottom 10 subfigures) Distribution 10 selected feature variables from non-cancer data group. They are mostly approximately normal distributed. Hence it's rational to use 10-dimensional normal distribution to approximate the distribution of feature vector.*

Table 2: *Classification algorithm performance of Multi-resolution Analysis Method (MRA) and other popular methods for IMS data analysis, where accuracy represents the rate of correct classification, sensitivity represents the rate that cancer is classified correctly as cancer and specificity represents the rate that non-cancer is classified correctly as non-cancer.*

	Accuracy	Sensitivity	Specificity
PCA+LDA	78.64%	100%	57.27%
PCA+SVM	71.82%	84.56%	59.09%
MRA	99.5%	99.08%	100%

where μ, Σ are mean and covariance of the 10-dimensional normal distributions.

To calculate the prior probability $P(i \in C)$ and $P(i \in N_c)$, we count the percentage of each type of pixels in training data,

$$P(i \in C) = \frac{|C|}{|C| + |N|} \tag{10}$$

$$P(i \in N_c) = \frac{|N|}{|C| + |N|} \tag{11}$$

Where $|C|$, $|N|$ are the number of cancer pixels and number of non-cancer pixels in training data respectively. With the above calculations, we can develop a classification model from training data. After we have developed this model using the training data, we test its performance using data set-2 (Figure 12). We select the two rounded marked areas as shown in Figure 12 as test data. Each area contains 109 pixels. The pixels in the left side rounded area are cancer pixels. Those in the right side rounded area are non-cancer pixels.

Figure 14 is the classification result. This graph shows the exponent value of the left side part of formula (6) and (7) for each pixel. Red points are results for cancer

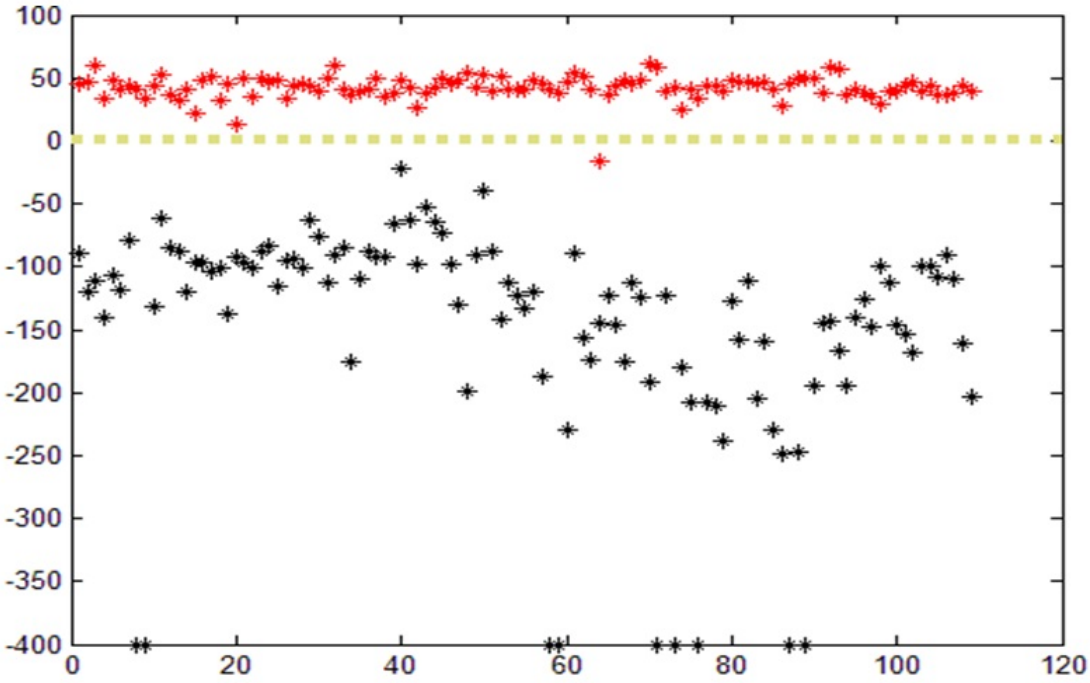

Figure 14: *Classification result of MRA method. This figure shows the value of* $log_{10}\frac{P(\boldsymbol{e}=\boldsymbol{X}|i\in C)P(i\in C)}{P(\boldsymbol{e}=\boldsymbol{X}|i\in N_c)P(i\in N_c)}$ *for each test data pixel. According to the classification criteria defined in formula (6) and (7), 0 is classification boundary (the broken yellow line in this figure). Red points are cancer pixels. Black points are non-cancer pixels. According to classification criteria, those above the yellow broken line should be classified as cancer pixels and those bellow the yellow broken line should be classified as non-cancer pixels.*

pixels and black points are for non-cancer pixels.

According to the classification criteria defined in formula (6) and (7), threshold should be 0, since $log_{10}(1) = 0$. Thus, those red points above the threshold and those black points below threshold are classified correctly. According to the result in Figure 14, there is only one pixel in cancer data that is misclassified as non-cancer. Thus, the performance for this classification algorithm is: 99.5% for accuracy, which represents the rate of correct classification; 99.08% for sensitivity, which represents the rate that cancer is classified correctly as cancer; and 100% for specificity, which represents the rate that non-cancer is classified correctly as non-cancer. Table 2 is a comparison of the performance of Multi-resolution Analysis Method with several other methods.

2.5 Conclusion

We proposed a multi-resolution analysis (MRA) method for IMS data analysis in biomarker selection and classification. According to data experiment results in table 1 of Section 2.3 and table 2 of Section 2.4, MRA method has advantages in effectiveness and accuracy in biomarker selection and classification comparing with other popular methods. The multi-resolution property of wavelet space saves computation time in finding biomarkers. The data experiment has shown that the CPU computing time of MRA method took only 54% of the computing time using EN4IMS method ([18], [17]). This work has been summarized in a paper [46] published recently.

Though it is challenge to incorporate spatial information for IMS data analysis using MRA method, we will tackle this important problem and report corresponding results in Chapter 3.

CHAPTER 3

AN ALGORITHM FOR INCORPORATING SPATIAL INFORMATION IN IMS DATA PROCESSING

To fully utilize IMS data, it is desirable to not only identify the peaks of the mass spectrum within individual pixels but also to study relations between pixels using the spatial information for the entire image cube. In fact, the state (cancer or non-cancer) of a pixel is highly determined by the configuration of its neighboring system. Because such locality and Markovianity property in space exists in IMS data, Markov Random Field (MRF) is an ideal tool that can describe this fact well. In this work, we will incorporate spatial information in IMS data analysis using MRF and optimize classification accuracy with Markov chain Monte Carlo (MCMC) sampling [7]. Firstly, we introduce the necessity of incorporating spatial information in IMS data analysis. Secondly, we give a brief introduction to MRF. Then, we will discuss the computation framework using MCMC sampling and Ising model, which is the simplest MRF, as prior information to optimize IMS data classification accuracy. The method to estimate parameters using training data is also discussed. Finally, we use test data to test the performance of this developed model under different assumptions of neighboring system definition. The experiment results show that this model can improve IMS data classification accuracy at more than 6%, and the more realistic the neighboring system is defined, the better classification result will be.

3.1 Motivation of this study

Even though IMS data provides us spatial information, when doing IMS data classification, most of current IMS data analysis methods like Principle Component Analysis

(PCA) [40], Support Vector Machine (SVM) [14], Multi-Resolution Analysis Method (MRA) [47] and Wavelet-Based Procedures for Proteomic MS Data Processing [10] do not consider the interactions between pixels depending on their spatial relations. However, such spatial relations do exist. Hong and Zhang proposed Weighted Elastic Net (WEN) model for IMS data processing [18], it's based on elastic net model proposed by Zou, et al [50] but a standard deviation weight is added to describe spatial relation between IMS pixels. But still WEN has certain limitations in the robustness of this method that the result accuracy may depend on data structure. We consider the fact that the growth of tumor is a continuous process. A tumor usually starts from one spot and spans to its neighboring area. If a cell is spatially surrounded by cancer cells, then this cell should have a high probability to be cancer. In other words, the class of a cell (cancer or non-cancer) is highly determined by the class configuration of its neighboring cells. Such spatial property can be described as locality or Markovianity in 2-D space. Markov random field (MRF) [31] is a mathematical tool that describes such Markovianity in a 2-D space. Therefore, MRF is an ideal tool to incorporate spatial information in IMS data. We can use MRF as the prior distribution of pixel classes and use MCMC framework to estimate the true class label of each pixel based on the initial classification result from current analysis methods [47] that do not consider spatial information. The classification accuracy can be expected to be improved in this way since we fully utilized spatial information of IMS data with MRF to describe pixels classes' spatial relations.

The following part of this chapter is organized as follows. In Section 3.2, we will give a brief introduction to MRF and Ising model, which is a specific type of MRF. In Section 3.3 and 3.4, we will talk about the MCMC computation framework for IMS data classification and the parameters estimation methods. In Section 3.5,

we will implement a data experiment using one data set to train parameters and another different data set to test the model. We will also discuss how the way the neighborhood system is defined matters the classification result.

3.2 Introduction to Markov Random Field

3.2.1 Definition of Markov Random Field

Markov random field (MRF) is n-dimensional random process defined on a discrete lattice. Usually the lattice is a regular 2-dimensional grid in the plane, finite or infinite. In the $2-D$ setting, assume that $S = \{1, 2, ..., N\} \times \{1, 2, ..., M\}$ is the set of $N \times M$ points [6], called sites. For a fixed site s define a neighborhood ∂s. For example for site (i, j) the neighborhood could be $\partial(i, j) = \{(i-1, j), (i+1, j), (i, j-1), (i, j+1)\}$. Markov Property (Markovianity) of $X(S)$ is defined via local conditions,

$$P(x_s|x_r, r \neq s) = P(x_s|x_{\partial r}) \tag{12}$$

where S is defined as set of lattice points, s is a lattice point, $(s \in S)$, X_s is the value of X at s, ∂s are the neighboring points of s. The random field X which has Markov property defined in formula (12) is called Markov random field (MRF). Formula (12) shows that in MRF, the probability of the value at any point is only determined by the values configuration of its neighboring points. For example, in the MRF shows in Figure 15, the value of $X_{2,2}$ is only determined by values of its neighborhood (green pixels).

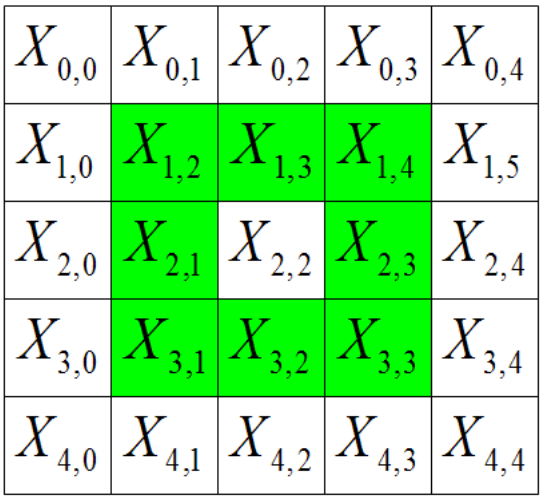

Figure 15: *In MRF, the value of a pixel is only determined by values of its neighborhood (green pixels). Such Markov property fits the reality of cancer tissue classification problem. In a tissue with cancer and non-cancer area, if a position is surrounded by cancer areas, then this position has a high probability to be cancer. The same rule applies for non-cancer area. Therefore, MRF can be an ideal tool to deal with the cancer tissue classification problem.*

3.2.2 A simplest MRF - Ising Model

Ising Model is the simplest type of MRF where there are only two possible values at any site: +1 and -1. It was proposed by a German physicist named Ernst Ising from his magnetic substance research. Originally in magnetic substance research context, +1 represents the north polarity of a particle is up and -1 represents the north polarity of a particle is down, while the polarity direction of particles are interacted with each other [21]. Such interaction is similar to the interaction of cancer and non-cancer cells depending on the spatial position. Here in IMS data analysis, we can represent

a cancer pixel by -1 site, and a non-cancer pixel by +1 site. Figure 16 is an example of Ising model.

The density function of Ising model is

$$P(X = x) \propto \frac{1}{Z}\exp(\beta \sum_{i \sim j} x_i x_j) \propto \frac{1}{Z}\exp(-2Jd_x) \tag{13}$$

where J is a constant parameter needs to be estimated using training data, d_x is the number of disagree edges (see Figure 17), $i \sim j$ means pixel X_i and X_j are neighboring to each other, Z is scale constant which will be cancelled in Ising prior ratio.

3.3 MCMC computation framework for IMS data classification

We denote θ as true classification which is our goal to approximate, y as observed classification which we already obtained from a previous existing classification algorithm [47] without incorporating spatial information. Our task is to estimate true classification θ while the observed classification y is given. Here we accomplish this task in a probabilistic way. We first obtain the distribution of true classification θ in condition that the observed classification y is given, and then we estimate θ by its posterior distribution probability in each class. Therefore, we can not only get a classification result, but also a probability value being classified to each class. According to Metropolis-Hasting theorem [29], the key theory of MCMC sampling, if we use Metropolis-Hasting algorithm to simulate $\theta|y$, the simulated data will finally have a distribution that converges to $f(\theta|y)$, the true distribution of $\theta|y$. In this way, we can estimate the true classification according to its probability distribution when only the observed classification is available. Here is the sampling rule of Metropolis-Hasting

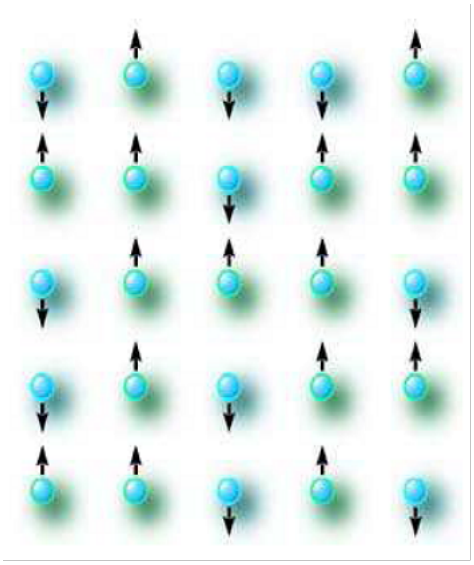

Figure 16: *Ising model is originally proposed from the research of magnetic substance to describe the macroscopic change of particle spin direction configuration with the microcosmic interaction of individual particle existing. +1 represents spin up and -1 represents spin down. We can use this model to describe the interaction between cancer and non-cancer pixels, with -1 representing cancer and +1 representing non-cancer.*

algorithm [29]:

$$\alpha[(\theta'|y)|(\theta|y)] = \min\left(1, \frac{f(\theta'|y)}{f(\theta|y)}\right) = \min\left(1, \frac{L(y|\theta')P(\theta')}{L(y|\theta)P(\theta)}\right) \tag{14}$$

where $\theta|y$ is the original value θ given observed classification y, $\theta'|y$ is the new proposed value θ' during MCMC sampling given observed classification y. The specific expressions for likelihood ratio $\frac{L(y|\theta')}{L(y|\theta)}$ and prior ratio $\frac{P(\theta')}{P(\theta)}$ in this formula will be given later in formula (15) and (16). Here are the steps of MCMC sampling algorithm:

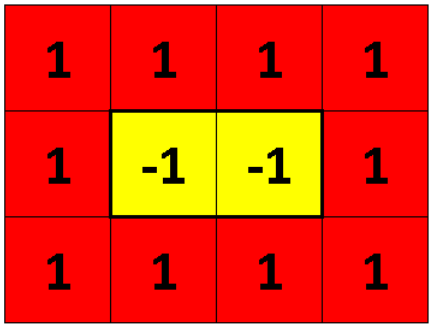

Figure 17: *Illustration of disagree edges number. The edge between two pixels with different values is a disagree edge, which is marked as a bold black line here. In this figure, the number of disagree edges is 6.*

Algorithm 1 [41]

Start with the space of all configuration $\boldsymbol{C}$ in which each configuration $\boldsymbol{\theta}|\boldsymbol{y}$ is represented as a vector:

$$\boldsymbol{\theta}|\boldsymbol{y} = (\theta_1, \theta_2, ..., \theta_{n-1}, \theta_n, \theta_{n+1}, ..., \theta_{M\times N})$$

with the indexing $(i, j) \mapsto n = (i-1) \times N + j$. The MCMC sampling algorithm would have following steps:

Step 1 Start with $\boldsymbol{\theta}|\boldsymbol{y} \in \boldsymbol{C}$. Usually, initially assign $\boldsymbol{\theta_0} = \boldsymbol{y}$.

Step 2 Randomly select a pixel from $\boldsymbol{\theta}|\boldsymbol{y}$, for example θ_n.

Step 3 Propose new value $\boldsymbol{\theta'}|\boldsymbol{y}$ as $\boldsymbol{\theta'}|\boldsymbol{y} = (\theta_1, \theta_2, ..., \theta_{n-1}, -\theta_n, \theta_{n+1}, ..., \theta_{M\times N})$ by changing the sign of the selected pixel's value.

Step 4 Generate a uniform random number $u \sim U(0, 1)$. If $u < \alpha[(\boldsymbol{\theta'}|\boldsymbol{y})|(\boldsymbol{\theta}|\boldsymbol{y})]$, then accept $(\boldsymbol{\theta'}|\boldsymbol{y})$ as new configuration. Otherwise, keep $(\boldsymbol{\theta}|\boldsymbol{y})$ as current configuration.

Iterate above process until converges.

Then we will discuss the specific expressions for likelihood ratio and prior ratio in formula (14). Since at each site, the conditional variable is independent to each other (this is because at each site, the observed class is completely determined by its true class and observation noise), then likelihood ratio can be written as:

$$\frac{L(y|\theta')}{L(y|\theta)} = \frac{\prod_{m,n} L(y_{m,n}|\theta'_{m,n})}{\prod_{m,n} L(y_{m,n}|\theta_{m,n})} \tag{15}$$

The prior is an Ising MRF. Then the prior ratio can be obtained using the distribution formula of Ising model in formula (13)

$$\frac{P(\theta')}{P(\theta)} = \frac{\frac{1}{Z}\exp(-2Jd_{\theta'})}{\frac{1}{Z}\exp(-2Jd_{\theta})} = \exp[-2J(d_{\theta'} - d_{\theta})] \tag{16}$$

The parameter J in the above formula can be estimated by training data. We will discuss the details about how to estimate parameter J in next section. $d_{\theta'}$ and d_{θ} are numbers of disagree adages for estimation θ' and θ respectively. The definition of d_{θ} was introduced in Section 3.2.2.

3.4 Parameter estimation of MRF prior and likelihood

3.4.1 Ising MRF prior parameter estimation using Maximum Pseudo Likelihood (MPL)

In 1986, Geman and Graffigne [13] proved that the following pseudo likelihood method can be used to approximate the likelihood of Gibbs distribution, which is the general format of the distribution of MRF. This method converges to Gibbs distribution with

probability 1. Here is how the pseudo likelihood defined:

$$PL(\mathbf{X}) \triangleq \prod_{i \in S} P(x_i | x_{N_i}) = \prod_{i \in S} \frac{P(x_i, x_{N_i})}{P(x_{N_i})} = \prod_{i \in S} \frac{P(x_i, x_{N_i})}{\sum\limits_{x_j \in L_S} P(x_j, x_{N_j})} \tag{17}$$

where L_S is labeling space (or class space). For instance, the label space for Ising model, a binary system, is $L_S = \{-1, 1\}$. i is pixel index, S is the set of all pixels, x_{N_i} are neighboring pixels of x_i.

For the Ising model, pluging in its distribution from formula (13) to formula (17), we can obtain its pseudo likelihood:

$$PL(\mathbf{X}) = \prod_{i \in S} \frac{\frac{1}{Z}\exp(-2Jd_{x_i})}{\sum\limits_{x_j \in L_S} \frac{1}{Z}\exp(-2Jd_{x_j})} \tag{18}$$

To find its maximum, we take its natural log. Then

$$\ln[PL(\mathbf{X})] = \sum_{i \in S} \{-2Jd_{x_i} - \ln[\sum_{x_j \in L_S} \exp(-2Jd_{x_j})]\} \tag{19}$$

Therefore, the MPL estimation of J is:

$$\widehat{J} = \arg\max_{J} \{\ln[PL(\mathbf{X})]\} \tag{20}$$

Using one dimensional optimization method, we can find the specific value of J that maximizes the pseudo likelihood in formula (19). This value is the estimated value for parameter J.

3.4.2 Likelihood estimation

Using the training data whose observed classification and true classification are both known, we can estimate the likelihood. Here, we are doing binary classification.

Therefore, there are only 4 types of likelihoods. They are

$$L(y = -1|\theta = -1)$$

$$L(y = 1|\theta = -1)$$

$$L(y = -1|\theta = 1)$$

$$L(y = 1|\theta = 1)$$

For example, $L(y = 1|\theta = -1)$ denotes the probability that a cancer pixel is observed as a non-cancer pixel. We count the frequency for each instance in the training data. Then we divided these frequencies with the total sample size to obtain the estimation of the above 4 likelihoods. For example, if the case $y = 1|\theta = -1$, which means the case that a cancer pixel is observed as a non-cancer pixel by an initial algorithm that does not incorporate spatial information, happens 30 times, and the total number of computed pixel is 200, then the likelihood probability $L(y = 1|\theta = -1) = \frac{30}{200} = 0.15$.

3.5 Data experiment

3.5.1 Data introduction

We have two IMS data sets. They are from two different mouse brains from the same species implanted with the same type of tumor. These two IMS data sets are produced from Vanderbilt Mass Spectrometry Research Center. One IMS data set has 24×34 pixels resolution; another has 44×64 pixels resolution. We use one data set to train model parameters and test the model performance on another data set. To reduce the mistakes made by the boundary pixels, we select only the central part of cancer and non-cancer area as training and test data so that the selected pixel class

is easy to see to judge performance of tested model. The red round areas shown in Figure 12 are selected training and test data.

3.5.2 Model parameter estimations using training data

First, we use MPL method discussed in Section 3.4.1 to estimate the parameter J in Ising MRF model using formula (20). To compute this, we take consideration of the class label configuration for all the pixels and their neighboring relations. Figure 18 shows the true configuration of training data class label. The blue pixels are cancer pixels ($X_i = -1$). Red pixels are non-cancer pixels ($X_i = +1$). Green pixels are margin blank space. There are 634 valid pixels for training data shown in Figure 18. We plug in values for training data into formula (19), (20) and use one dimensional optimization method, we obtained that when $J = 1.0266$ the $PL(X)$ is maximized. Therefore, the Maximum Pseudo Likelihood estimation value for J is 1.0266. Hence, for the experiment IMS data we use here, the prior ratio in acceptance probability for MCMC sampling in formula (14) can be written as:

$$\frac{P(\theta')}{P(\theta)} = \exp[-2J(d_{\theta'} - d_\theta)] = \exp[-2 \times 1.0266(d_{\theta'} - d_\theta)] \tag{21}$$

Second, we use the idea discussed in Section 3.4.2 to estimate likelihood. We count the frequencies of 4 cases happening in selected training data computed by initial algorithm: the case that cancer pixel is classified as cancer pixel (corresponding to $L(y = -1|\theta = -1)$); the case that cancer pixel is classified as non-cancer pixel (corresponding to $L(y = 1|\theta = -1)$); the case that non-cancer pixel is classified as cancer pixel (corresponding to $L(y = -1|\theta = 1)$); the case that non-cancer pixel classified as non-cancer pixel (corresponding to $L(y = 1|\theta = 1)$). For example, we selected two round areas in training data, totally 323 IMS data pixels. According

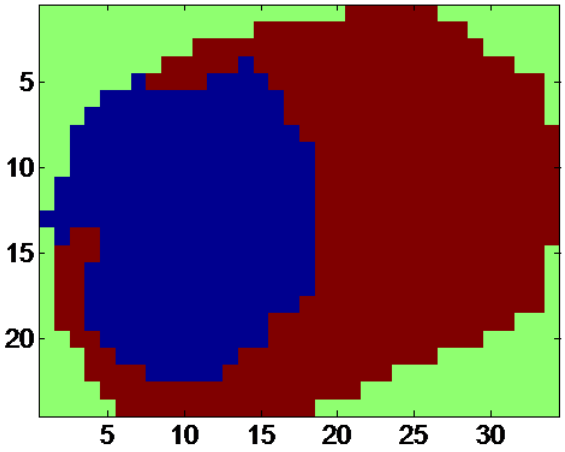

Figure 18: *True configuration of training data class label. The blue pixels are cancer pixels. Red pixels are non-cancer pixels. Green pixels are margin blank space. This configuration can be used to estimate parameter J in Ising MRF model using MPL method discussed in Section 3.4.1.*

to the classification result by initial algorithm [47] which did not incorporate spatial information, 122 cancer pixels are classified as non-cancer pixel, then likelihood for this case is $L(y = 1|\theta = -1) = \frac{37}{323} = 0.1146$. Here are the likelihoods we computed from training data using the initial algorithm before optimization:

$$\begin{cases} L(y = -1|\theta = -1) = \frac{122}{323} = 0.3777 \\ L(y = 1|\theta = -1) = \frac{37}{323} = 0.1146 \\ L(y = -1|\theta = 1) = \frac{130}{323} = 0.4025 \\ L(y = 1|\theta = 1) = \frac{34}{323} = 0.1053 \end{cases} \tag{22}$$

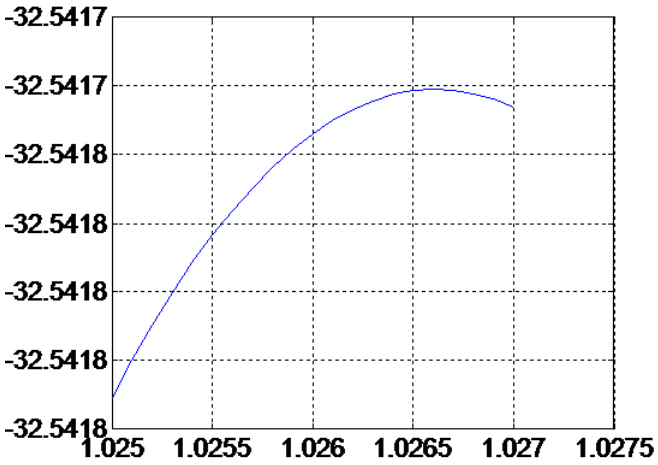

Figure 19: *Ising prior paramter estimation. When* $J = 1.0266$*, formula (19) for training data is a maximum. Therefore,* $J = 1.0266$ *is the estimated value for corresponding parameter of Ising MRF prior ratio in formula (16).*

3.5.3 Computation and result on test data

We already estimated prior and likelihood for acceptance probability in formula (14). Now we can start MCMC simulation discussed in Section 3.3 for test data to estimate its true classification θ with its observed classification y, which is firstly computed by initial algorithm that did not consider spatial information for IMS data. The initial algorithm we use here is modified from MRA (Multi-resolution Analysis) method for IMS discussed in Chapter 2, which did not incorporate IMS data spatial information. Usually, for MRA method getting good classification accuracy, there should be 10 feature variables selected. But here, to leave some potential for optimization, we only select 3 feature variables so that the classification accuracy turns out to be 86%. We take this classification result as observed classification y as shown in Figure 20. The

acceptance probability in MCMC simulation for test data is

$$\begin{aligned}\alpha[(\theta'|y)|(\theta|y)] &= \min(1, \frac{f(\theta'|y)}{f(\theta|y)}) = \min(1, \frac{L(y|\theta')P(\theta')}{L(y|\theta)P(\theta)}) \\ &= \min\{1, \frac{\prod\limits_{m,n} L(y|\theta')}{\prod\limits_{m,n} L(y|\theta)} \exp[-2 \times 1.0266(d_{\theta'} - d_{\theta})]\}\end{aligned} \tag{23}$$

The value of likelihood $L(y_{m,n}|\theta'_{m,n})$ is estimated in formula (22).

Then we follow MCMC simulation steps described in **Algorithm 1**. We start with initial estimation $\boldsymbol{\theta_0} = \boldsymbol{y}$. Then propose a change of one pixel's class estimation; generate a uniform random number and compare it with acceptance probability in formula (23) to determine whether accept this proposal or not. Iterate this process for a certain amount of times until it converges. Then gather statistics of the simulated data to compute the posterior probability $f(\theta_{m,n} = -1|y_{m,n})$ for each pixel. Then set 0.5 as probability threshold. If $f(\theta_{m,n} = -1|y_{m,n}) > 0.5$, this means pixel $X_{m,n}$ has more chance to be cancer and we classify it as a cancer pixel. Otherwise, we classify $X_{m,n}$ as non-cancer pixel. Figure 20 shows the result before and after applying **Algorithm 1**. We can see the classification accuracy is improved from 86.2385% to 92.6606%.

In the above computation process, the neighborhood system in Ising MRF prior is defined as 4-points neighborhood system. That is, for site (i, j) the neighborhood is

$$\partial(i, j) = \{(i-1, j), (i+1, j), (i, j-1), (i, j+1)\} \tag{24}$$

Only the up, down, left, right adjacent pixels are considered as neighboring pixels in 4-points neighborhood system. However, it is more reasonable to also consider diagonal adjacent pixels as neighboring pixels because they also have impacts on pixel at site (i, j). Therefore, we can define the neighborhood for site (i, j) in 8-

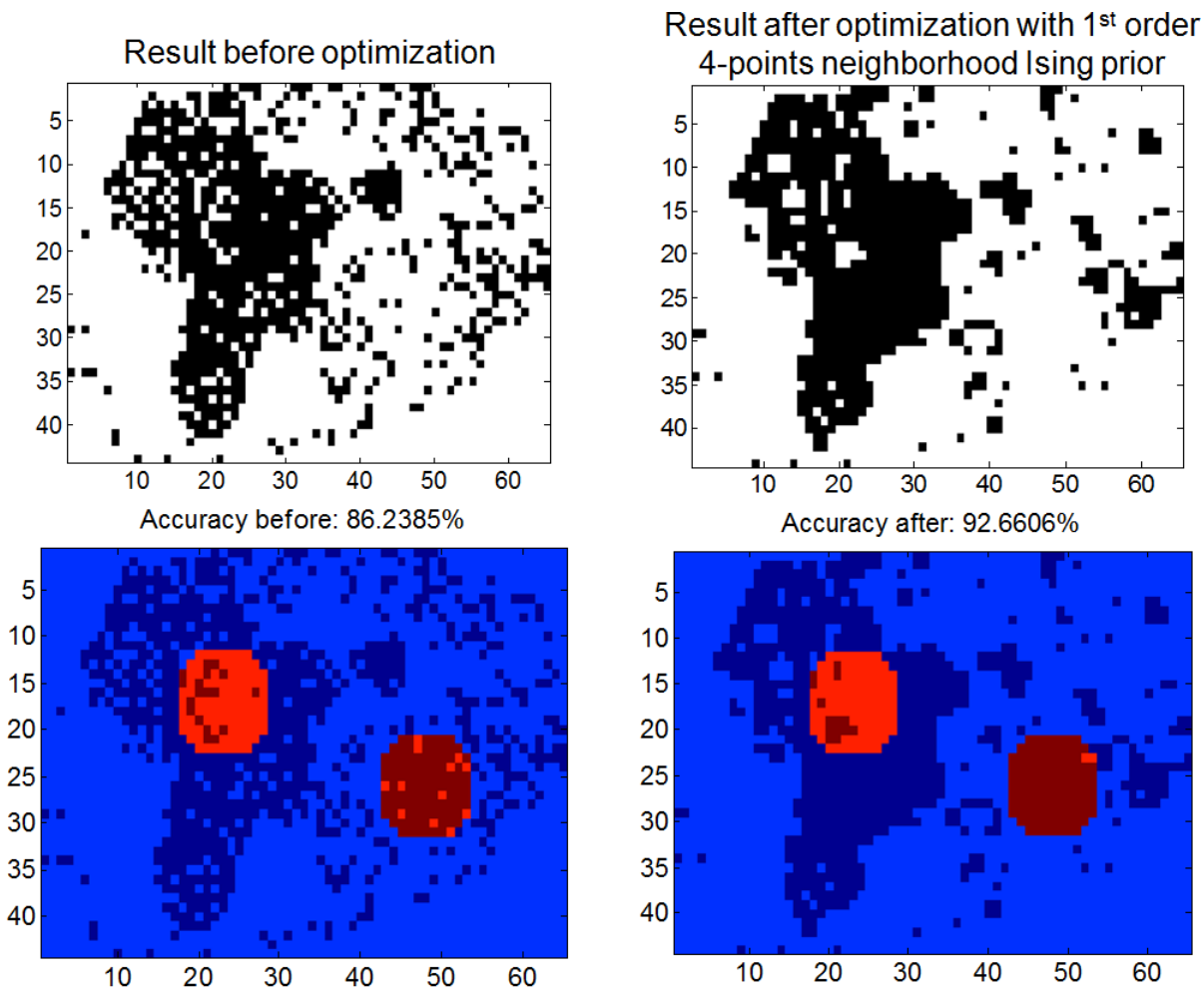

Figure 20: *Result after optimization with the 1st order 4-points neighborhood Ising prior. (Top left) Classification result using an algorithm without incorporating spatial information. The black pixels (-1 in Ising MRF) are classified as cancer pixels and white pixels (+1 in Ising MRF) are classified as non-cancer pixels. (Top right) Classification result of optimized algorithm using MCMC-MRF to incorporate spatial information. Lots of misclassifications (noise) have disappeared. (Bottom left) The classification accuracy for selected test area using an initial algorithm without incorporating spatial information is only 86.2385%. (Bottom right) The optimized classification accuracy for selected test area using MCMC-MRF to incorporate spatial information is improved to 92.6606%.*

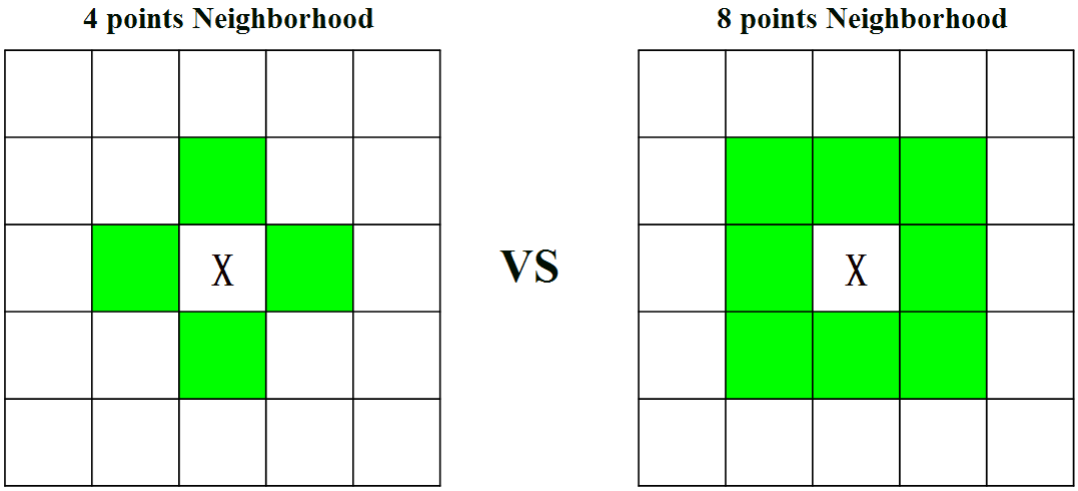

Figure 21: *4-points neighborhood system and 8-points neighborhood system in Ising MRF. As left figure shows, in 4-points neighborhood system, only up, down, left, right adjacent green pixels are considered to be neighboring to pixel X. But in 8-points neighborhood system as shown in the right figure, the diagonal adjacent pixels are also considered as neighboring to X.*

points neighborhood system as

$$\partial(i,j) = \{(i-1,j),(i+1,j),(i,j-1),(i,j+1),(i-1,j-1),(i-1,j+1),(i+1,j-1),(i+1,j+1)\} \tag{25}$$

After we modified the definition of neighborhood system in Ising MRF prior from 4-points neighborhood to 8-points neighborhood, we apply MCMC simulation described in **Algorithm 1** to test the performance again. The result shows the performance under 8-points neighborhood assumption is better than 4-points neighborhood assumption. The classification accuracy is improved to 94.9541%. This shows that a more realistic definition of neighborhood system which describes the spatial impacts

between cancer cells leads to a better classification result.

To take one step further, we can define even higher order neighborhood system for Ising MRF prior. Because in the reality, not only the directly adjacent cells have impacts on the surrounded central cell, but the cells close while not directly adjacent also have impacts on the central cell even though the impacts can be weaker. Therefore, it is more reasonable to consider neighborhood like Figure 23 shows: the closest points are the 1st order neighboring points; the second closest points are the 2nd order neighboring points, etc. Then the Ising MRF prior probability for nth order neighborhood system is:

$$\begin{aligned} P(x) \propto \frac{1}{Z}\exp(\beta \sum_{i\sim j} x_i x_j) \propto \frac{1}{Z}\exp(c_1 \sum_{i_1\sim j_1} x_{i_1}x_{j_1} + ... + c_n \sum_{i_n\sim j_n} x_{i_n}x_{j_n}) \\ \propto \frac{1}{Z}\exp[-2J(c_1 d_{x,l=1} + ... + c_n d_{x,l=n})] \end{aligned} \tag{26}$$

where $d_{x,l=i}$ is number of disagree edges between the ith order neighboring pixels and the central pixel, c_n is the impact coefficient from nth neighboring pixels. It is reasonable to assume that further pixels have less impact. Therefore C_n is defined as inverse proportional to Euclidean distance:

$$c_n \propto \frac{1}{D(x_{i_n}, x_{j_n})} \tag{27}$$

Here we define 5-orders neighborhood system as shown in Figure 23. Then we plug formula (26), formula (27) to acceptance probability in formula (14) to update the corresponding computation in **Algorithm 1** and retest the model on test data. It turns out that the result is even better than 8-points neighborhood assumption. As Figure 24 shows, the classification accuracy is improved to 95.4128%, better than the 1st order 4-points, 8-points neighborhood system. This shows again that the

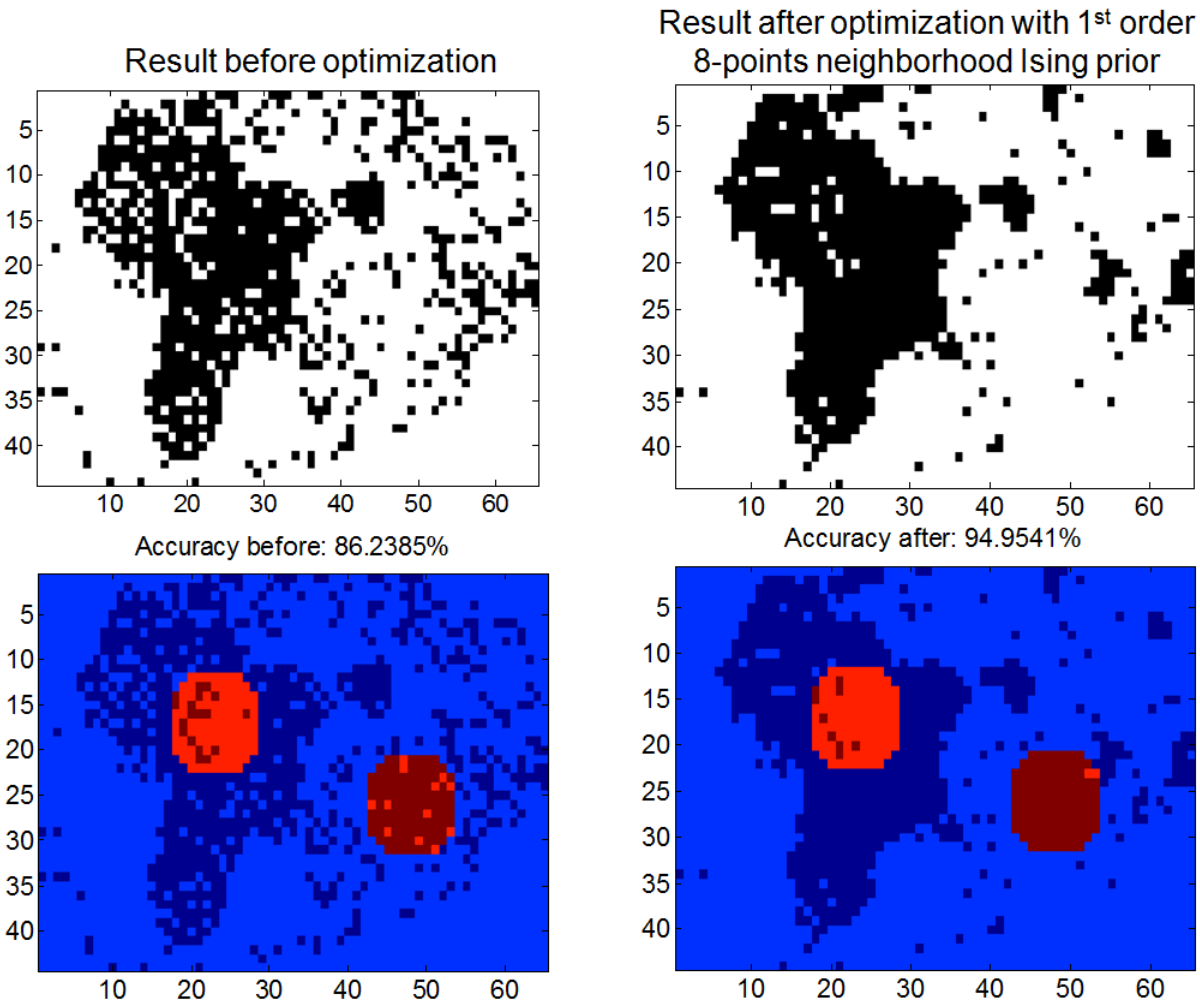

Figure 22: *Result after optimization with the 1st order 8-points neighborhood Ising prior. Classification result of 8-points Neighborhood. (Top left) The initial classification result before optimization using an algorithm without incorporating spatial information. The black pixels (-1 in Ising MRF) are classified as cancer pixels and white pixels (+1 in Ising MRF) are classified as non-cancer pixels. (Top right) Classification result of optimized algorithm using MRF-MCMC to incorporate spatial information with 8-points neighborhood system. (Bottom left) The classification accuracy for selected test area before optimization is only 86.2385%. (Bottom right) The optimized classification accuracy for selected test area using 8-points neighborhood MRF is improved to 94.9541%, better than the accuracy under 4-points neighborhood assumption.*

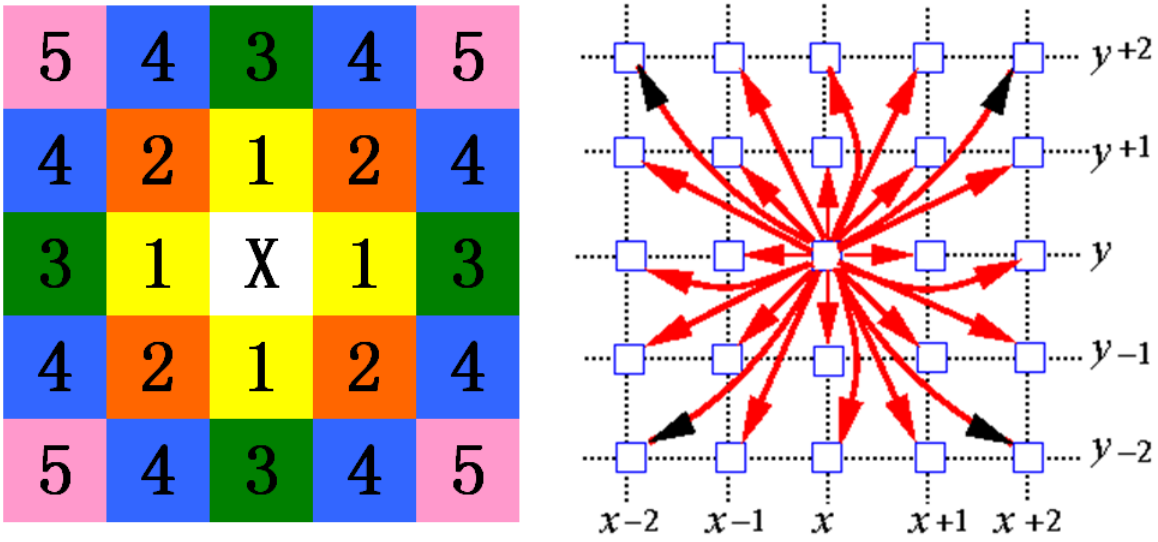

Figure 23: *Higher orders neighborhood system. (Left) 5-orders neighborhood system. The closest pixels to X in the figure are 1st order pixels. The furthest pixels of X in this figure are 5th order pixels. (Right) The impact between neighboring pixels is defined in formula (15), inverse proportional to their Euclidean distance. In other words, the closer pixel has stronger impact, while further pixel has weaker impact to the central pixel.*

more realistic neighborhood system definition which describe the spatial interactions between different areas in cancer tissue more precisely, leads to better classification result.

3.6 Conclusion and future work

The main idea of this work is using MRF as a prior knowledge to describe the spatial relationships between different parts of cancer tissue and using MCMC to estimate

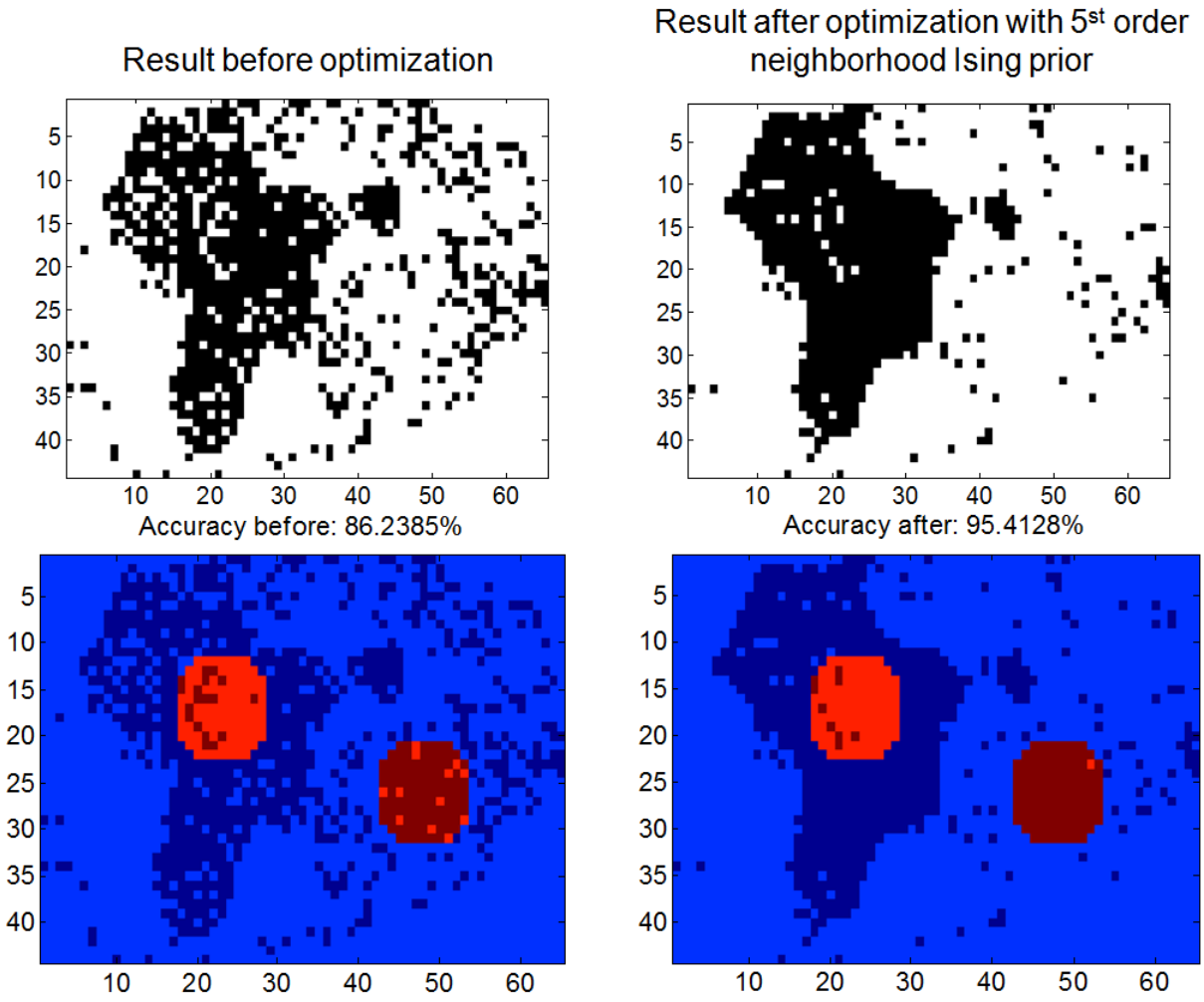

Figure 24: *Comparison of classification result of 5-orders neighborhood system before and after MRF-MCMC optimization. (Top left) The initial classification result before optimization. (Top right) Classification result after MRF-MCMC optimization to incorporate spatial information with 5-orders neighborhood system. (Bottom left) The classification accuracy for selected test area before optimization is only 86.2385%. (Bottom right) The optimized classification accuracy for selected test area using 5-orders neighborhood MRF is improved to 95.4128%, better than the accuracy under 1st order neighborhood system of 4-points or 8-points neighborhood.*

the true classification based on the observed classification (initial classification before optimization) by approximating the probability distribution of true classification using MCMC sampling. We estimated the MCMC-MRF model parameters using training data and tested the model using another test data set. The data experiment shows that this method can improve the classification accuracy at more than 6% compared with traditional IMS data algorithm like PCA, SVM, MRA method that have not incorporate spatial information. Also, the test result shows that the more realistic we define the neighborhood which precisely describes the interactions mechanism between different parts in cancer tissue, the better classification result we can obtain. This work was summarized in a manuscript [46] submitted for consideration of publication in a statistical computing journal.

The future work can be considered in three aspects. Firstly, we need to consider faster computing method, either coding wise or mathematical algorithm wise, since we experienced the time consuming of this MCMC-MRF simulation during test, especially when the neighborhood system is defined as high order. Secondly, we can apply some statistical analysis to the simulation result to obtain variables such as confidence interval, standard deviation so that we will have a better evaluation of the simulation. Instead of getting just one class label for each pixel, we can obtain more information using statistical analysis to the simulated data. Finally, we can consider defining a more complicated and more precise neighborhood system with impacts coefficients estimated using more training data so that the model can describe the reality even better.

CHAPTER 4

USING MONTE CARLO SIMULATION TO PREDICT CAPTIVE INSURANCE SOLVENCY

Computational statistics has very broad applications. In this Chapter, I would like to present a very recent research project based on my internship at SIGMA Actuarial Consulting Group regarding solvency of captive insurance.

The solvency [32] of captive insurance [34] fund is a main thing captive manager cares. The challenges of captive insurance rating come from the following aspects. First, the high-dimensionality of factor space. For example, for work compensation captive insurance, there are factors such as age, gender, education level of participants, policy retention (deductible), premiums, investment income, asset, liability of captive insurance company, etc. All these factors influence solvency of captive insurance companies and make captive solvency rating a complex high-dimensional modeling problem. Which factors to choose for rating solvency? How to integrate them together? These questions need to be answered well. The second challenge comes from the definition of solvency itself. There are different standards to say an insurance company is solvent or not, for example, whether its asset is greater than its liability, or whether it can pay claims in time, etc. Which standard should we use for captive insurance solvency rating? This is another difficult question. However, nowadays the research for captive insurance solvency rating is few. In the summer of 2014, I got a chance to intern at SIGMA Actuarial Consulting Group and got to know this problem. The company pre-designed a spreadsheet describing the important financial ratios and their relations for captive insurance. It needs to develop a robust solvency rating model for captive insurance based on this spreadsheet. I have spent 2

months for checking the accuracy of the pre-designed financial spreadsheet, then built a solvency prediction model for captive insurance fund using Monte Carlo simulation [15] with the fund's current financial data and setups. This model can tell captive managers the solvency score of the current fund using the fund survival probability in the next several years as a measurement of solvency. Standard financial reports will also be generated in each year including the income statement, the balance sheet and the summary of financial ratios. Based on this captive financial model, we design a captive solvency rating model by generating random numbers following the distribution of historical loss to simulate future losses. If the solvency ratios break the upper and lower bounds in a simulated case, we count it as an insolvent case; otherwise, it is a survival (or solvent) case. After large sampling, we can approximate the future survival probability of the current captive fund. We use a heat-map to visualize the solvency score of each setup choice so that it will be easier for captive insurance managers to compare their decision choices.

The preliminary results of this work was presented on the 49th Actuarial Research Conference (ARC) in the poster session on July 14th, 2014 [45].

4.1 Introduction

4.1.1 Captive insurance

Captive insurance was initially created because people could not find a good enough insurance provider in the traditional public commercial insurance market, either the price is too high, or the specific type of risk cannot be covered. Then people think about creating their own insurance company as a subsidiary company to manage their risks instead of buying traditional commercial insurance. This type of insurance company established with the specific objective of insuring risks emanating from their

parent group or groups is called captive insurance company [43]. Captive insurance, as an alternative risk financing program, has several advantages over traditional insurance. First, parent group is very familiar with the risks they have and can estimate potential future loss very accurately according to their experience. Second, the data of some specific type of risks in the parent group(s) are confidential, therefore captive insurance, which is owned by the parent group(s), is a better choice than public insurance providers. Third, a captive insurance company can be used as a tax shelter [24]. Fourth, captive insurance can avoid the boom and bust cycle [19] of the insurance industry. When insurance industry is in a hard market, it is difficult to find an insurance provider offering the risk protection at an expected price, but captive insurance can be independent of the market cycle since it is not open to public market. Fifth, captive insurance can offer insurance protection at a lower cost. Because in traditional commercial insurance, 40% of premiums are additional fees, including advertisement fees, management fees and profit; however, captive insurance, a company owned by the parent group to insure the parent group, does not have these fees. Sixth, captive insurance company and its parent group(s) have a coincidence of interests, therefore there is no risk of indirect selecting and moral risk. Also, captive insurance can improve cash flow stability of its parent group(s), because the premiums can be paid in a more flexible way. Because of these advantages, captive insurance market have been continuously growing in the last 20 years (see figure 25) [8].

4.1.2 Solvency of captive insurance

Solvency is the ability of a company to meet its long-term financial obligations. It can be viewed from different angles. For example, solvency can be viewed as a "ruin theory" that if assets are greater than liabilities then it is considered to be solvent.

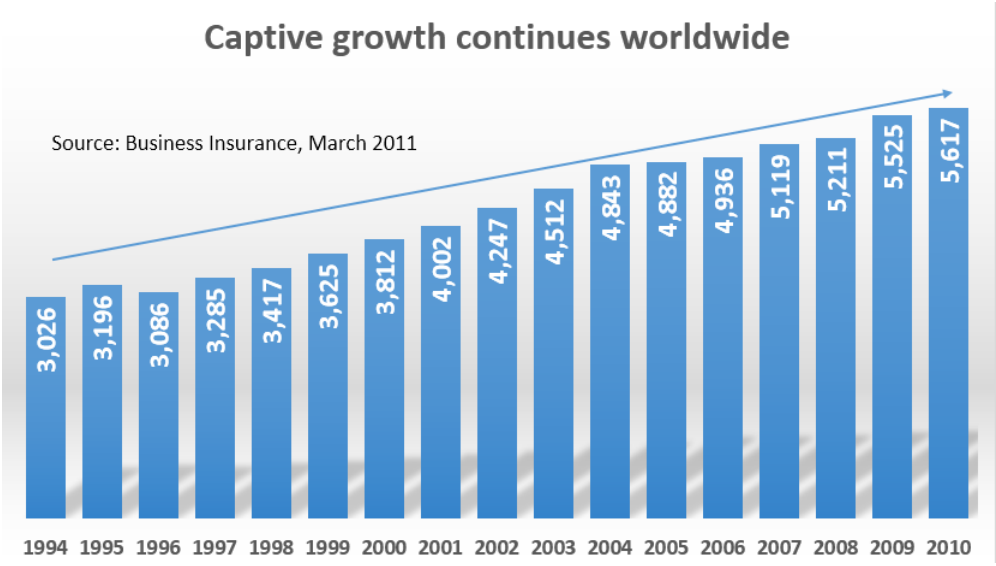

Figure 25: *Captive growth continues worldwide [8].*

Also solvency can be viewed as a "liquidity theory", if a insurance company can meet its current liabilities then it is considered to be solvent [44]. Solvency is not measured by just one ratio, it is a systematic measurement of a insurance company's financial health.

The current popular solvency rating methods include Insurance Regulatory Information System (IRIS), Financial Analysis and Solvency Tracking (FAST) and solvency II. IRIS was developed by the National Association of Insurance Commissioners (NAIC) in 1970s based on the Early Warning System. Usually, there are 11 IRIS ratios for property\casualty insurance and 12 IRIS ratios for life insurance. These ratios, combined with their normal range, are used to rate insurance companies' solvencies to prevent solvency crises. Based on IRIS, since 1995 NAIC applied an additional analysis to large insurance companies (the company that has annual premium greater than $50 million for life\health insurance and $30 million for property\casualty insur-

ance), the so called Financial Analysis and Solvency Tracking (FAST) system. The purpose of the FAST system is to prevent large insurance companies from having a solvency crisis. The FAST system is more complex than IRIS. FAST is working together with IRIS, not replacing it. In 2002, the European Commission passed Solvency 1. However, with the financial integration development within the European Union and the change in the insurance industry, the existing regulation framework cannot work well any more. since Solvency 1 was passed, the European Commission started the Solvency II project, aiming to develop a new solveny regulation system which works closely to the risk management of insurance companies.

Solvency rating is especially important to captive insurance because captive insurance does not receive strict regulations as traditional commercial insurance companies do and is easy to have solvency problem if solvency situation is not evaluated often or well. The solvency problem will lead a captive insurance company to collapse and affect the benefits of its insureds. But current popular solvency rating methods are not specially designed for captive insurance, therefore they do not work best for captive insurance solvency rating.

4.2 Motivation of this study

The motivation of this study is to overcome the shortages of current popular IRIS-based solvency prediction methods.

First, current popular solvency evaluation methods using IRIS ratios are deterministic, where a deterministic rating will be given without probability distribution. For example, in [4], 1 point is given for each of 12 IRIS ratios going outside the usual range and solvency is scored from 0 to 12 points. However, future solvency is a random variable since future losses and other business variables are not determinis-

tic. Therefore, it makes more sense to use a probabilistic method to measure future solvency to figure out the probability of solvency and insolvency in future years.

Second, current popular methods using IRIS ratios focus on the current solvency evaluation [39], not future prediction. But knowing the future solvency trend gives captive insurance managers more confidence about decision makings. We use Monte Carlo simulation to simulate future losses according to experienced loss distribution to simulate future IRIS ratios to predict future years' solvencies.

Third, instead of fixed lower bound and upper bound of IRIS ratios used in most current deterministic solvency prediction methods, we modify it to allow users define the IRIS lower and upper bound according to their business situations while referring to the recommended IRIS "usual range" so that this model will have more flexibility compared with the traditional model to describe different business situations.

We also visualize the results using a heat-map matrix to make users compare results easily when they make retention decisions.

4.3 Methodology

We use Monte Carlo simulation to simulate future losses according to the experience loss distribution as shown in Table 3. Once any of the IRIS ratios hit the red lines set by user (see Figure 26), we count it as a failure case. Otherwise, it is a survival case. For example, in the simulation illustrated in Figure 27, the 7th simulated IRIS ratio breaks normal bounds. Therefore, this simulation will be counted as an insolvency case.

Solvency is measured by fund survival probability which can be approximated by as:

$$P(\mathbf{S}) \approx \frac{|\mathbf{S}|}{|\mathbf{F}| + |\mathbf{S}|} \tag{28}$$

where $\mathbf{F}$ is the set of failures in Monte Carlo simulation, $\mathbf{S}$ is the set of fund survival (solvent) cases, $|\mathbf{F}|$ is the size of the set.

Future losses are simulated based on experience loss distribution as shown in Table 3. The distribution data that users input are discrete data points (see table 3). To get a continuous distribution function so that the simulated loss can be continuously sampled from, we use log-normal distribution to fit the discrete data we have. The reason we choose log-normal distribution is that it is positive skewed and widely used within the insurance industry. Formula (30) is the cumulative distribution function (CDF) of log-normal distribution. The future loss is simulated using inverse method as formula (29). In inverse method,

$$x_i = F^{-1}(u_i), \tag{29}$$

where $u_i \sim U[0, 1]$ is generated random number which has a uniform distribution from 0 to 1. Therefore, using inverse method, we can map each uniform random number u_i to simulated future loss x_i that has a log-normal distribution. If discrete distribution is used, for example the input discrete distribution in Table 3, then $x_i = \dot{F}(u_i)$ is a piecewise inverse function.

In formula (30), there are 2 parameters, μ and σ, that we need estimate from fitted data points. We use least squared error as the measurement to find the values of μ and σ which can minimize the squared error between fitting CDF and fitted data points. Therefore, μ and σ can be estimated using formula (31).

	Must be Above	Must be Below
Gross Premium to Surplus	0%	900%
Net Premiums to Surplus	0	300%
Per Occurrence Risk Retention to Surplus	0	25%
Paid Losses to Liquid Assets	0	120%
Change in Premiums	-33%	33%
Change in Surplus	-10%	50%
One Year Reserve Development to Surplus	0	20%
Investment Yield on Invested Assets	3%	6.50%
Investment Yield on Underwriting Income	-5%	5%
One Year Change in Reserve to Surplus	0%	10%
Liabilities to Liquid Assets	0%	110%
Loss and LAE to Net Premiums Earned	0%	70%
Underwriting Expense to Net Premiums Earned	0%	15%
Policyholder Dividend to Net Premiums Earned	0%	10%
Combined Ratio	0%	100%
Net Investment Income to Net Premiums Earned	-5%	5%
Operating Ratio	0%	50%
Overall Liquidity	100%	900%
Underwriting Cash Flow	User define	User define
Operating Cash Flow	User define	User define
Net Income to Net Premiums Written	User define	User define

Figure 26: *IRIS ratios industry recommended usual range [4]. Users can modify the values of red lines according to their situations. For example, if user think only Combined Ratio matters when predict solvency, then user can set other ratios' usual ranges as* $(-\infty, +\infty)$*, so that those ratios will not affect the solvency prediction result.*

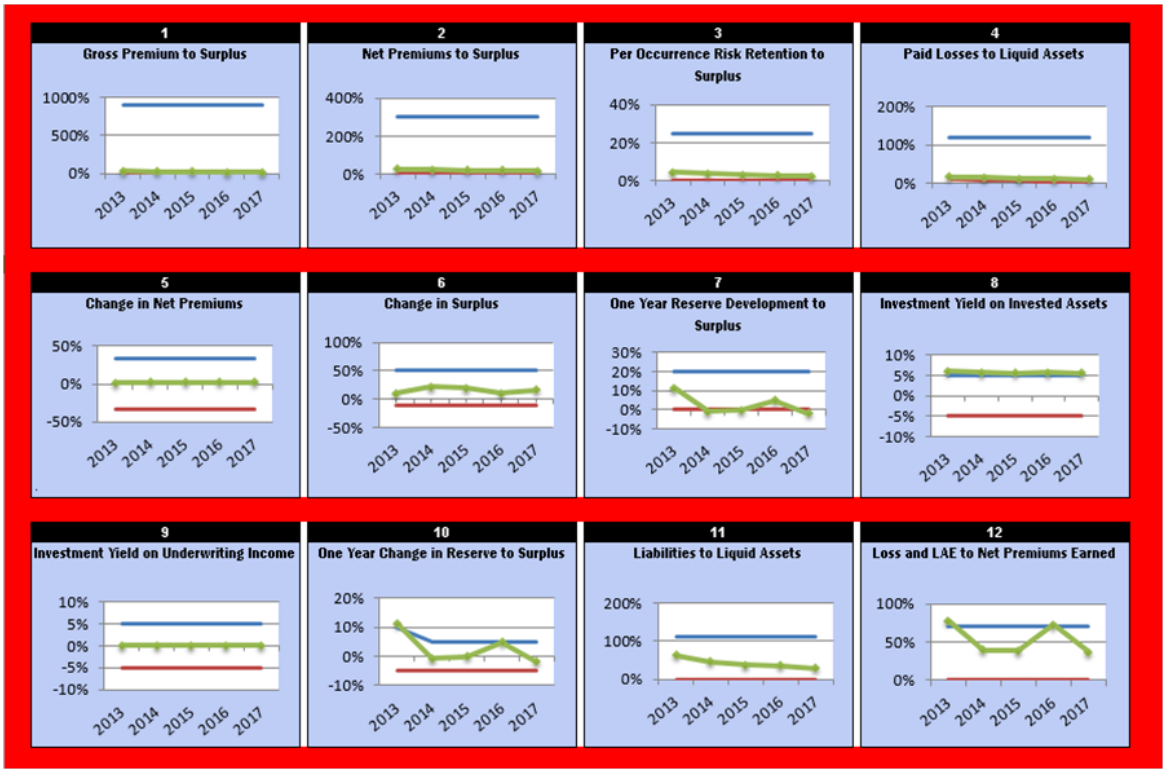

Figure 27: *One run of simulation. In a simulation, each of IRIS ratios will be compared with upper bound (the blue line if this figure) and lower bound (the red line in this figure) defined by user. Any of IRIS ratios break its bounds defined by user will be counted as an insolvency case, otherwise counted as a solvency case. This figure shows an insolvency case since the* 7*th,* 10*th and* 12*th simulated IRIS ratios break their bounds.*

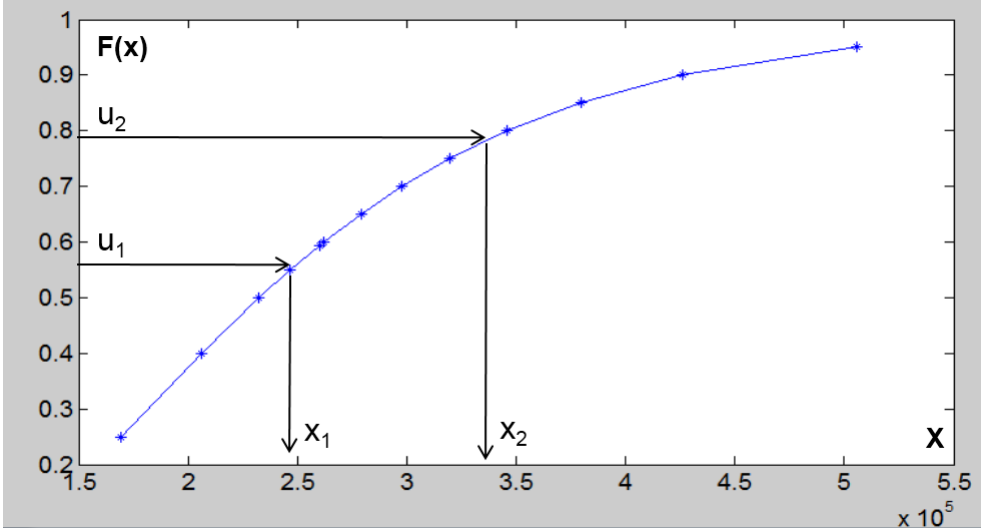

Figure 28: *Log-normal distribution is used to fit the experience loss distribution data points input by user. Then random number $u_i \in [0, 1]$ will be generated to simulate future loss x_i using inverse method $x_i = F^{-1}(u_i)$.*

$$F(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int_0^x \frac{\exp[\frac{-(\ln(t)-\mu)^2}{2\sigma^2}]}{t} dt \tag{30}$$

$$(\hat{\mu}, \hat{\sigma}) = \arg\min_{\mu,\sigma} \sum_i [(F(x_i; \mu, \sigma) - \dot{F}(x_i)]^2 \tag{31}$$

To implement the computing of formula (31), we use grid search method to find best μ and σ. Grid searching is a global search method. In grid search method, we divide searching area into fine grid and compute squared error between fitting CDF and fitted data at each grid. In other word, for each value of μ and σ on grids, we compute corresponding squared error $|F(x_i) - \dot{F}(x_i)|^2$ to find the μ, σ which can minimize this squared error.

To reduce the size of searching area, before we start grid searching, we do an initial

estimation of μ and σ using the moment formula of log-nomal distribution. We can use data mean $\overline{X_i}$ to approximate the first moment $E(x)$ and mean of sqaured data $\overline{X_i^2}$ approximate the second moment $E(x^2)$. Therefore we obtain equations system 32.

$$\begin{cases} E(x) = \exp(\mu + \sigma^2/2) \approx \overline{X_i} \\ E(x^2) = \exp(2\mu + 2\sigma^2) \approx \overline{X_i^2} \end{cases} \tag{32}$$

Therefore, we can solve μ and σ in above system as an initial estimation. The solution for this questions system is

$$\begin{cases} \sigma^2 \approx \ln(\overline{X_i^2}) - 2\ln(\overline{X_i}) \\ \mu \approx [4\ln(\overline{X_i}) - \ln(\overline{X_i^2})]/2 \end{cases} \tag{33}$$

Using the data points input by user in Table 3, the initial estimate of log-normal parameters is $\mu = 12.3419$, $\sigma = 0.4134$. The searching area should be around this initial estimation. For example, the searching boundary can be 30% above and below the initial estimation. To find the minimum, we need make sure the searching area contains the local minimum point (we believe $\sum_i |F(x_i;\mu,\sigma) - \dot{F}(x_i)|^2$ is a 2D convex function for μ and σ, therefore a local minimum is a global minimum). We can try and adjust the different searching boundaries until a local minimum point is included and the convex shape is shown. since we are searching for 2 parameters, the searching area is a 2D squared area. In figure 31, the area is $\mu \in [12, 13]$, $\sigma \in [0, 1]$. The squared fitting error is minimized to 7.067×10^{-4} when $\mu = 12.36$ (see Figure 29), $\sigma = 0.45$ and this is the best estimation of log-normal parameters using least squared fitting error as fitting measurement.

Since our model only ask user to input experience loss distribution under one specific retention level, we need obtain other loss distributions if retention level is different. According to experience, we assume there exists a linear relation between

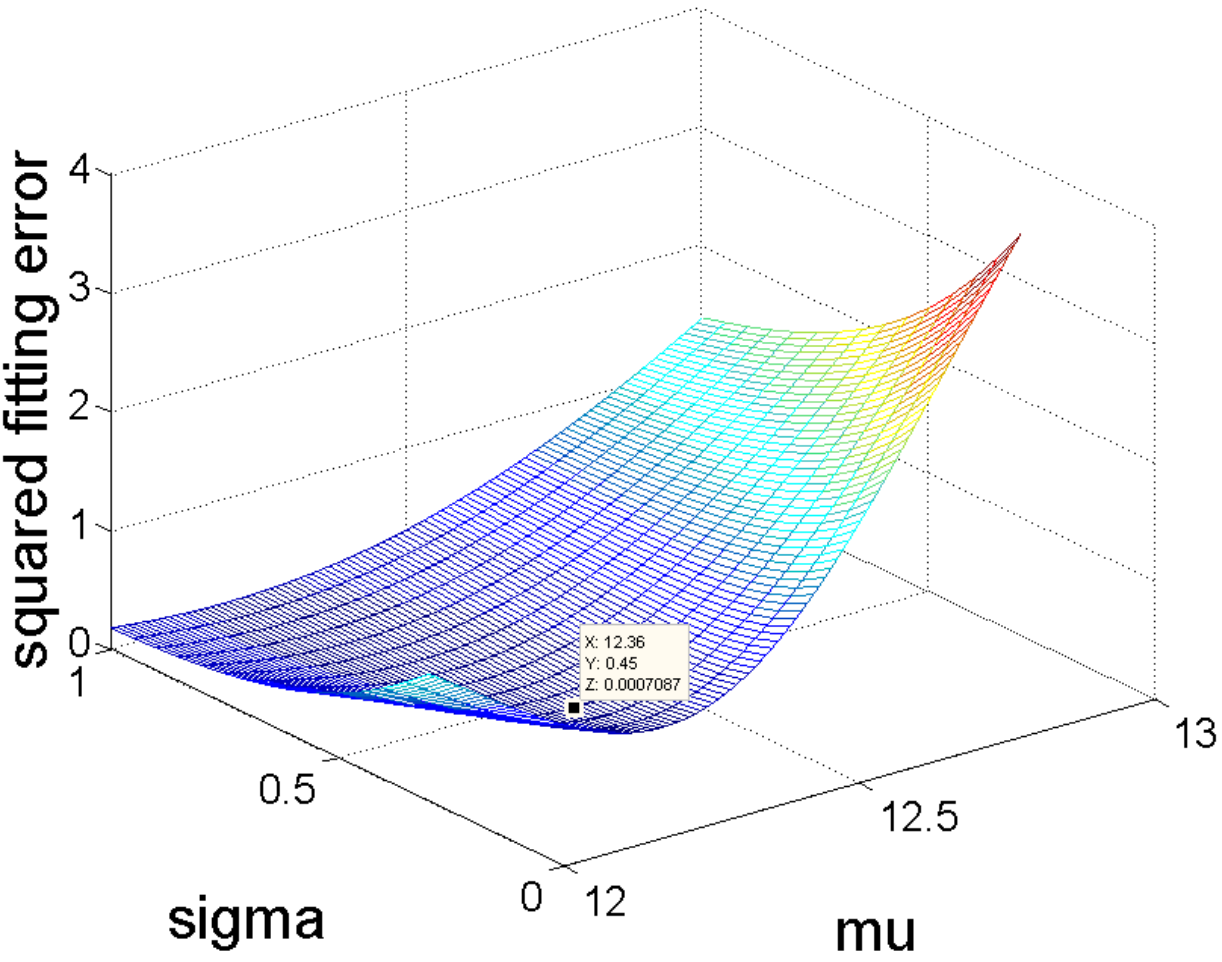

Figure 29: *Parameters estimation of log-normal distribution. When* $\mu = 12.36$, $\sigma = 0.45$, *sqaured fiiting error is minimized. Therefore,* $\mu = 12.36$, $\sigma = 0.45$ *is the best estimation of log-normal parameters.*

Table 3: *Experience loss distribution data points input by user (at retention \$100,000), where $\dot{F}(x)$ is cumulative probability.*

$\dot{F}(x)$	Loss X
0.25	\$169101.40
0.40	\$206320.40
0.50	\$232551.80
0.55	\$246773.80
0.59336	\$260000.00
0.60	\$262116.40
0.65	\$278977.40
0.70	\$297921.00
0.75	\$319810.40
0.80	\$346080.80
0.85	\$379441.40
0.90	\$426020.40
0.95	\$505783.20

the loss x_i and retention R_j selected as shown in table 4. For example, if loss is x_i when retention R_j = \$100000, then when retention R_j = \$250000 the loss will be $\frac{1.4}{1.0}x_i$. This relation is summarized in formula (34), where R_m is mth retention and x_m is loss under this retention R_m.

$$\frac{R_m}{R_n} = \frac{x_m}{x_n} \tag{34}$$

Based on above assumptions and equstion, we propose **Algorithm 3.1** to predict future solvencies.

Table 4: *The relation between retention and expected loss (data from insurance experience). We assume there exists a linear relation between the loss x_i and retention R_j selected. This linear relation is summarized in formula (34).*

Retention (R_j)	Increased Limits Factor (L_j)	Expected loss (X_j)
\$25,000	0.65	\$169,000
\$50,000	0.80	\$208,000
\$100,000	1.00	\$260,000
\$250,000	1.40	\$364,000
\$500,000	1.75	\$455,000
\$1,000,000	2.20	\$572,000
\$Unlimited	3.00	\$780,000

Algorithm 3.1

Step 1 User inputs recent 2-years' financial report (Income Statement and Balance Sheet), additional information, additional data required and define n, the number of future years to simulate, $\mathbf{B}_u$ and $\mathbf{B}_l$, the upper bound and lower bound of IRIS ratios.

Step 2 Initialize $|\mathbf{F}|$, the number of insolvency cases, to be 0, and $|\mathbf{S}|$, the number of solvency cases, to be 0. Based on user's input, for each retention level R_j, generate a random number u_1 and use inverse method to experience loss distribution to map u_1 to x_1, the coming first year's simulated loss and calculate first year's simulated financial report $\mathbf{Fin}_1$ and IRIS ratios $\mathbf{I}_1$.

Step 3 Based on previous year's simulated financial report $\mathbf{Fin}_{i-1}$, generate a random number u_{i-1} and use inverse method to experience loss distribution to map u_{i-1} to x_{i-1}, the coming ith year's simulated loss and calculate ith year's simulated financial report $\mathbf{Fin}_i$ and IRIS ratios $\mathbf{I}_i$ using simulated loss and previous year's simulated financials.

Step 4 Repeat Step 3 until $\mathbf{Fin}_n$ and $\mathbf{I}_n$ is simulated.

Step 5 For each retention level R_j, compare each year's simulated IRIS $\mathbf{I}_i$ with interval $[\mathbf{B}_u, \mathbf{B}_l]$. If any value in vector $\mathbf{I}_i$ is outside $[\mathbf{B}_u, \mathbf{B}_l]$, we count it as a insolvency case and add 1 to $|\mathbf{F}|$; otherwise, add 1 to $|\mathbf{S}|$;

Step 6 For each retention level R_j and year future year i, its solvency score $P(\mathbf{S}) = \frac{|\mathbf{S}|}{|\mathbf{F}|+|\mathbf{S}|}$.

4.4 Results and discussion

The survival probability $P(\mathbf{S})$ is approximated using formula (28), by dividing the number of survival cases to the total number of simulated cases. There is a survival probability computed for each future year under each retention level. It measures the probability of solvency in a specific future year under specific retention level. We use the most recent 2 years' historical data as the input to simulate its future solvency using Algorithm 3.1. We visualize the result (see Figure 30) using a heat map. These results are generated using a discrete loss distribution input by the user.The darker color corresponds to the lower survival probability of the fund and the lower solvency score, indicating a higher chance of insolvency. The lighter color corresponds to higher survival probability of fund and higher solvency score, indicating lower chance of insolvency. Higher solvency score is better. The number in each cell is fund survival probability (or solvency score) approximated using Monte Carlo simulation we discussed before. For example, the number in row 1, column 5 is 0.36, which means the solvency score (from 0 to 1 measured by fund survival probability) 5 years later with \$25,000 retention is 0.36, and solvency problem is more serious than its previous year (since it is 0.424 solvency score for the 4th year). This result gives each year's solvency score from 0 to 1 at each retention level. it is nice to see a result like this which changes continuously from one to another. it is interesting to find that at \$25000 retention for example, even though solvency score is high in the beginning, it decreases in future. Maybe it is because too less profit they make since they take so less risk. However, the unlimited risk is too risky and the solveny score turns to be low in this case. Retention \$250,000 for example, makes solvency score increasing in future and keep in a related high solvency level, can be a good choice.

	Year 1	Year 2	Year 3	Year 4	Year 5
Retention=$25000	0.73	0.42	0.51	0.42	0.36
Retention=$50000	0.89	0.73	0.73	0.73	0.51
Retention=$100000	0.85	0.85	0.89	0.89	0.63
Retention=$250000	0.63	0.71	0.75	0.75	0.78
Retention=$500000	0.48	0.57	0.59	0.63	0.63
Retention=$1000000	0.26	0.26	0.26	0.48	0.48
Retention=$unlimited	0.00	0.00	0.00	0.00	0.00

Figure 30: *A heat map style visualized result.*

CHAPTER 5

CONCLUSIONS

We proposed a Multi-Resolution Analysis (MRA) method using wavelet transform for IMS data biomarker selection and classification. By transforming IMS data onto wavelet coefficients, we reduced the high dimensionality of IMS data and by analyzing IMS data taking advantage of the multi-resolution property of wavelet coefficients, we saved computing time to achieve a faster algorithm speed than other popular methods. The biomarkers list selected by MRA method is shorter than other popular methods while the important biomarker are selected but less noise are selected. The two biomarkers has already been confirmed by cancer study. Based on wavelet transformed IMS data, we select its wavelet coefficients as feature variable and use Bayes classifier to classify each IMS pixel. Data experiment shows this algorithm can get higher classification accuracy than other popular methods.

To incorporate spatial information of IMS data, we consider the local property (Markovianity) in 2D space of tumor growth and use Markov Rand Field (MRF) to describe this spatial relation. We use Ising model (a binary format of MRF) as prior and training data as likelihood to estimate the posterior probability using Metropolis-Hasting algorithm. In this posterior estimation, the result generated by any algorithm that not consider spatial information is the conditional variable. We use this as simulation staring point to estimate the true classification. Data experiment shows this MCMC-MRF algorithm can improve the IMS data classification accuracy at least 6% from other method not consider spatial information. We also test different neighboring rules, and data experiment show 8 points neighboring works better than 4 points neighboring, higher-order neighborhood system works better than first-order

neighborhood for IMS pixel classification. We can deduce that a good definition of neighboring in MRF for IMS data must describe the biology mechanism of spatial impact in tumor growth.

Continued with the stochastic simulation idea in MCMC-MRF mentioned in previous subsection, we applied stochastic Monte Carlo simulation to captive insurance solvency prediction. We combined a pre-build financial model for captive insurance, IRIS ratios and Monte Carlo simulation to design a flexible, robust and easy to use solvency prediction model specially for captive insurance to meet its solvency prediction and retention selection need. This model shows some advantage compared with other similar models: First, this is specially designed of captive insurance while open research for captive insurance solvency is few; second, the solvency score has an obvious probabilistic meaning that it ranges from 0 to 1 measure the probability that it breaks IRIS bounds, while other IRIS based methods is deterministic without probability meaning. Third, this model allows the user modified the IRIS bounds values according to their business situation and this gives them more flexibility to build a model fit their unique solvency prediction problem compared with classic fixed IRIS industry bounds.

BIBLIOGRAPHY

[1] Abdollahi A., S. Akbari, H. R. Maimani, *Non-commuting graph of a group*, Journal of Algebra, **298**, (2006), 468-492.

[2] Addison P., *The illustrated wavelet transform handbook: Introductory theory and applications in science, engineering, medicine and finance,* CRC Press, 2010.

[3] Alexandrov T., J. H. Kobarg, *Efficient spatial segmentation of large imaging mass spectrometry datasets with spatially aware clustering*, Bioinformatics, **27**(13), (2011), i230–i238.

[4] Balbin E. P., *Insurance regulatory information system (IRIS)*, USAID, EISA, Cairo, Egypt, 21 May 2008.

[5] Biomedical, F. A. S. T. *NIH definition of Biomarker.* Clin Pharmacol Ther, **69**, (2001), 89-95.

[6] Bouman C., K. Sauer, S. Saquib, *Markov random fields and stochastic image models*, 1995 IEEE International Conference on Image Processing.

[7] Brooks S., A. Gelman, G. Jones, X. Meng, *Handbook of markov chain monte carlo*, Chapman and Hall/CRC, 2011.

[8] Bulkowski J., A. Rhodes, *An introduction to captives and conducting the annual reserve analysis*, CAS 2012, 2012.

[9] Chen C. F., C. H. Hsiao. *Haar wavelet method for solving lumped and distributed-parameter systems*, Control Theory and Applications, IEE Proceedings-, **144**(1), IET, 1997.

[10] Chen S., D. Hong, Y. Shyr, *Wavelet-based procedures for proteomic MS data processing*, Computational Statistics and Data Analysis, **52**(1), (2007), 211–220.

[11] Deininger S. O., M. P. Ebert, A. Ftterer, M. Gerhard, *MALDI imaging combined with hierarchical clustering as a new tool for the interpretation of complex human cancers*, Journal of Proteome Research, **7**(12), (2008), 5230–5236.

[12] Domingos P., M. Pazzani, *On the optimality of the simple Bayesian classier under zero-one loss*, Machine Learning, **29**(2-3), (1997), 103–130.

[13] Geman S., C. Graffigne, *Markov random field image models and their applications to computer vision*, Proceedings of the International Congress of Mathematicians, **4**(5), (2011), 1496–1517.

[14] Gerhard M., S. Deininger, F. M. Schleif, *Statistical classification and visualization of MALDI imaging data*, In Computer-Based Medical Systems, CBMS'07. Twentieth IEEE International Symposium. IEEE , (2007), 403-405.

[15] Glasserman P., *Monte Carlo methods in financial engineering,* Vol. 53. Springer, 2004.

[16] Hanselmann M., M. Kirchner, B. Y. Renard, E. R. Amstalden, K. Glunde, R. M. Heeren, F. A. Hamprecht, *Concise representation of mass spectrometry images by probabilistic latent semantic analysis*, Analytical Chemistry, **80**(24), (2008), 9649-9658.

[17] Hong D., F. Zhang, *Elastic netbased framework for imaging mass spectrometry data biomarker selection and classification*, Statistics in Medicine, **30**, (2011), 753-768.

[18] Hong D., F. Zhang, *Weighted elastic net model for mass spectrometry imaging processing*, Mathematical Modelling of Natural Phenomena, **5**(3), (2010), 808-814.

[19] Lane M., O. Mahul, *Catastrophe risk ricing : An empirical analysis,* World Bank, Washington, DC, (2008). (http://hdl.handle.net/10986/6900)

[20] Leydold J., W. Hormann, *UNU.RAN user manual, version 1.8.0*, Institut fuer Statistik, WU Wien, 25 October 2010 (Available on line: http://statmath.wu.ac.at/unuran/doc/unuran.pdf)

[21] Lieb E., T. Schultz, D. Mattis, *Two-dimensional ising model as a soluble problem of many fermions*, Rev. Mod. Phys, **36**, (1964), 856–871.

[22] Matoba S., J. G. Kang, et al., *P53 regulates mitochondrial respiration*, Science, **312**(5780), (2006), 1650-1653.

[23] Mayevsky A., *Mitochondrial function and energy metabolism in cancer cells: Past overview and future perspectives*, Mitochondrio, **9**(3), (2009), 165–179.

[24] Mazerov M., *State corporate tax shelters and the need for Combined Reporting*, State Tax Notes, **46**, (2007), 621-638.

[25] McDonnell L. A., R. Heeren, *Imaging mass spectrometry*, Mass Spectrometry Reviews, **26**(4), (2007), 606-643.

[26] Morris J. S., et al. *Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models.* Biometrics, **64**(2), (2008), 479-489.

[27] National Cancer Institute, *NCI dictionary of cancer terms: biomaker*,

http://www.cancer.gov/dictionary?cdrid=45618, last accessed on 12/20/2013 at 15:36.

[28] Ren X., J. Malik, *Learning a classification model for segmentation,* Computer Vision, Proceedings. Ninth IEEE International Conference on. IEEE, 2003.

[29] Robert C., G. Casella, *Introducing monte carlo methods with R*, Springer New York, 2010. (Available online: http://dx.doi.org/10.1007/978-1-4419-1576-4_6)

[30] Rohner T., D. Staab, M. Stoeckli, *MALDI mass spectrometric imaging of biological tissue sections*, Mechanisms of Ageing and Development, **126**(1), (2005), 177-185.

[31] Rozanov Y., *Markov random fields*, Springer, New York, 1982.

[32] Sandstrom A., *Solvency: models, assessment and regulation.* CRC Press, 2005.

[33] Schmidt A., I. Forne, A. Imhof. *Bioinformatic analysis of proteomics data.* BMC Systems Biology, **8**(2), (2014), 1-7.

[34] Scordis N. A., M. M. Porat, *Captive insurance companies and manager-owner conflicts,* Journal of Risk and Insurance, **65**(2), (1998), 289-302.

[35] Shimma S., M. Setou, *Review of imaging mass spectrometry*, Journal of the Mass Spectrometry Society of Japan, **53**, (2005), 230–238.

[36] Tan P., M. Steinbach, V. Kumar, *Introduction to data mining*, Addison-Wesley, 1st edition, 2005.

[37] Toldo R., F. Andrea, *Robust multiple structures estimation with j-linkage,* Computer Vision–ECCV, Springer Berlin Heidelberg, 2008. 537–547.

[38] Trede D., J. H. Kobarg, K. Steinhorst, T. Alexandrov, *Mathematical methods for imaging mass spectrometry*, Cancer Research, **4**, (2011), 5.

[39] United States General Accounting Office (GAO), *The insurance regulatory information system needs improvement*, GAO/GGD-91-20 Insurance Regulation, 1990.

[40] Van de Plas R., B. De Moor, E. Waelkens, *Imaging mass spectrometry based exploration of biochemical tissue composition using peak intensity weighted PCA*, In Life Science Systems and Applications Workshop, LISA 2007, IEEE/NIH. IEEE , (2007), 209–212.

[41] Wang L., J. Liu, S. Li, *MRF parameter estimation by MCMC method*, Pattern recognition, **33**(11), (2000), 1919–1925.

[42] Watrous J. D. , T. Alexandrov, P. C. Dorrestein, *The evolving field of imaging mass spectrometry and its impact on future biological research*, Journal of Mass Spectrometry, **46**(2), (2011), 209-222.

[43] Wikipedia contributors, "*Captive insurance*", Wikipedia: The Free Encyclopedia, http://en.wikipedia.org/wiki/Captive_insurance (accessed October 17, 2014).

[44] Wthrich M. V., M. Merz., *Financial modeling, actuarial valuation and solvency in insurance*. Springer Berlin Heidelberg, 2013. 261-336.

[45] Xiong L., *Using Monte Carlo simulation to predict captive insurance solvency* , 49th Actuarial Research Conference (ARC), University of California, Santa Barbara, California, in July 13-16, 2014. (http://www.pstat.ucsb.edu/ARC/ARC2014/poster.html)

[46] Xiong L., D. Hong, *An MCMC-MRF algorithm for incorporating spatial information in IMS data processing*, submitted manuscript, 2014.

[47] Xiong L., D. Hong, *Multi-resolution analysis method for IMS proteomic data biomarker selection and classification*, British Journal of Mathematics and Computer Science, **5**(1), (2015), 65–81. (Available online: http://www.sciencedomain.org/issue.php?iid=707id=6).

[48] Xu Y., X. Yang, H. Ling, H. Ji, *A new texture descriptor using multifractal analysis in multi-orientation wavelet pyramid,* In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on (pp. 161-168). IEEE.

[49] Zavorin I., M. J. Le, *Use of multiresolution wavelet feature pyramids for automatic registration of multisensor imagery*, Image Processing, **14**(6), (2005), 770-782.

[50] Zou H., T. Hastie, *Regularization and variable selection via the elastic net.* Journal of the Royal Statistical Society: Series B (Statistical Methodology), **67**(2), (2005), 301-320.