2004-11-30

IP Algorithm Applied to Proteomics Data

Christopher Lee Green
Brigham Young University - Provo

IP Algorithm Applied to Proteomics Data

By

Christopher L. Green

A project submitted to the faculty of

Brigham Young University

In partial fulfillment of the requirements for the degree of

Master of Science

Department of Statistics

Brigham Young University

December 2004

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a project submitted by

Christopher L. Green

This project has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

_____	_____
Date	H. Dennis Tolley Chair
_____	_____
Date	Gilbert W. Fellingham
_____	_____
Date	Scott D. Grimshaw

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the project of Christopher Green in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

_____ _____

Date H. Dennis Tolley
Chair, Graduate Committee

Accepted for the Department

_____

G. Bruce Schaalje
Graduate Coordinator

Accepted for the College

_____

G. Rex Bryce
Associate Dean, College of Physical and Mathematical Sciences

ABSTRACT

IP ALGORITHM APPLIED TO PROTEOMICS DATA

Christopher L. Green

Department of Statistics

Master of Science

Mass spectrometry has been used extensively in recent years as a valuable tool in the study of proteomics. However, the data thus produced exhibits hyper-dimensionality. Reducing the dimensionality of the data often requires the imposition of many assumptions which can be harmful to subsequent analysis. The IP algorithm is a dimension reduction algorithm, similar in purpose to latent variable analysis. It is based on the principle of maximum entropy and therefore imposes a minimum number of assumptions on the data. Partial Least Squares (PLS) is an algorithm commonly used with proteomics data from mass spectrometry in order to reduce the dimension of the data. The IP algorithm and a PLS algorithm were applied to proteomics data from mass spectrometry to reduce the dimension of the data. The data came from three groups of patients, those with no tumors, malignant or benign tumors. Reduced data sets were produced from the IP algorithm and the PLS algorithm. Logistic regression models were constructed using predictor variables extracted from these data sets. The response was threefold and indicated which tumor classifications each patient belonged. Misclassification rates were determined for the IP algorithm and the PLS algorithm. The

rates correct classification associated with the IP algorithm were equal or better than those rates associated with the PLS algorithm.

Acknowledgments

I have to acknowledge first my advisor Dennis Tolley. Despite several serious roadblocks, he never lost faith that I would finish this project. This paper is an extension of his research and I could not have done it without him. I also want to thank Gilbert Fellingham who always had good advice and helped me pull it together in the end. Finally I have to thank McKay Curtis who cheerfully deflected my pessimism by laughing at me every time I complained too much about problems I was having with the project.

Table of Contents

List of Tables

List of Figures

Chapter 1 - Introduction

The information age affects every aspect of our civilization. Advances in computing and laboratory techniques make it possible to collect more data than we can handle. Computing power and data availability have progressed faster than the statistical methods required to analyze the data. This has become both a blessing and a curse, especially in proteomics, the study of proteins and their effect on biological processes.

Mass spectrometry is a valuable tool in proteomic studies. Data from mass spectrometry has the potential to unlock the causes of disease, the aging process and the secrets of human performance at the cellular and sub-cellular level. Mass spectrometry has the capability to produce tens of thousands of observations per subject, which allows thousands of proteins to be studied simultaneously. However, there is difficulty associated with so much data. Traditional statistical techniques rely on the number of observations per subject being less than the number of subjects.

Specialized techniques exist to deal with this problem. However, these specialized techniques require a researcher to make many assumptions about the data. Tolley et al. (2004) used the term, “Conventional Wisdom” to describe these assumptions. The researcher reduces the dimensionality of the data by making judgment calls (often arbitrary judgment calls) regarding which part of the data cannot or should not be used. The analysis is only accurate to the degree of the researcher’s assumptions being correct. It would be useful if a technique could be applied to proteomics data from mass spectrometry (heretofore referred to as mass protein data) that does not throw out data and makes a minimum number of assumptions.

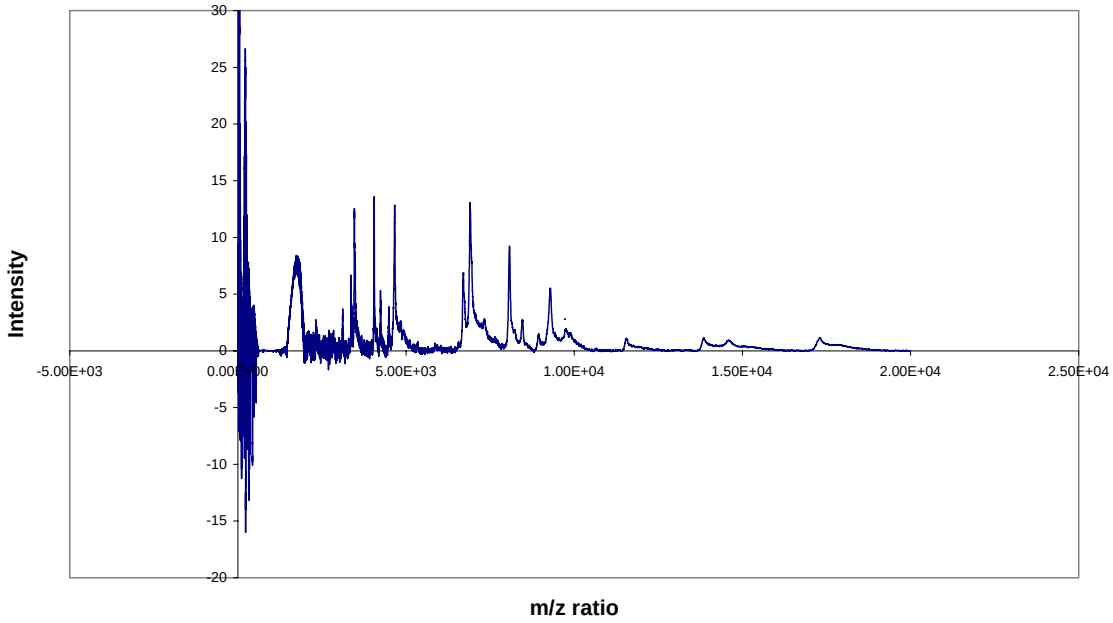
Chapter 2 – Literature Review

Dimension Reduction:

Analysis of mass protein data typically involves two problems. One problem is reduction of the dimensionality of the data, and the other problem is the classification or clustering of the data in order to draw conclusions. Dimension reduction is most often accomplished by principle components analysis (PCA) or partial least squares (PLS). Lee et al. (2003) used both principle components analysis and partial least squares to reduce the number of variables produced from mass protein data. Partial least squares is similar to PCA because it creates a new set of orthogonal variables that are linear combinations of the predictor variables. However, unlike PCA, the new variables do more than maximize the variation of the predictor variables. They are constrained by an iterative process so that they maximize the correlation between the new variables and the response variables. The new variables produced by PLS and PCA are assigned weights. In almost all PCA and PLS analysis, variables with lower weights are usually thrown out while variables with higher weights are kept. The weight threshold, after the variables are discarded, is partially an arbitrary decision by the researcher.

The values produced by mass spectrometry are intensities and are a function of mass divided by charge (m/z ratios). Each distinct m/z ratio is a separate variable. If intensity is plotted as a function of m/z ratio, the pattern is one of multiple peaks and valleys. Figure 1 shows this pattern for mass protein data of a single subject from a 2002 study by Petricoin et al. There are 15,154 separate m/z ratios in figure 1.

Figure 1 – Raw Mass Protein Data



A common dimension reduction technique is to use only variables associated with peak heights; consequently only a fraction of peak heights are selected, which are those higher than a defined threshold determined by a certain (or specific) researcher. Neville et al. (2003) did a study of 24 diseased patients and 17 healthy patients using mass protein data. 55 peaks were selected for analysis based on height and local noise; which is somewhat arbitrary and problematic because variance associated with the peaks increase as the peak height increases. Thus, choosing the highest peaks is not always the best strategy.

Other techniques are also employed to reduce the number of variables. Hilario et al. (2003) published a study on mass protein data from lung cancer patients that used two dimension reduction techniques. One technique (InfoGain) ranks predictors according to their mutual information with a class variable. The other technique (Relief-F) assigns

weights to variables using a method similar to K-nearest neighbors. The K-nearest neighbors for a given subject of the same experimental condition are near hits. The K-nearest neighbors for a given subject of a different experimental condition are the near misses. The difference between each variable (for a given subject) and the corresponding variables in its near miss and near hit subjects is calculated. The weight assigned to each variable decreases as the average distance to its near hits increases and decreases as the average distance to its near misses decreases. In both techniques, the final subset of variables is based on a user defined weight threshold.

Trends to Quantization

A decision that must be made prior to selection of a classification algorithm is whether to discretize mass protein data. Throughout the literature, the m/z density values are treated as continuous variables. However, mass protein data often contains enough noise that specialized techniques designed for noisy data are useful (Dancik 1999). Furthermore, the variance of the density values is highly correlated with the magnitude of the density values. Thus, very small m/z density values can be just as informative as very large ones. Therefore, some researchers consider the density values as discrete variables.

Boros et al. (2000) implemented a technique called Logical Analysis of Data (LAD), first introduced by Crama in 1988, wherein continuous random variables are discretized or binned into categories before a special classification algorithm is used. LAD was implemented for six pre-selected data sets ranging from heart disease data to predicting housing prices in Boston data. Boros showed that for all six data sets, the LAD classification system had comparable, if not better, misclassification rates than common classification algorithms that treated the variables as continuous. In 2004,

Alexe et al. used LAD to classify mass protein data from patients with ovarian cancer. By discretizing m/z density values, Alexe implemented a classification system with a sensitivity of 93.8 percent and a specificity of 100 percent in leave one out testing.

Other studies have also found that discretization of mass protein data is appropriate. For example, Hilario et al. (2003) applied several classification algorithms to m/z density peaks, treating them as continuous variables, then discretizing them. He concluded that:

> "…discretization helps generalization for all learning methods, the mean prediction error is significantly lower after discretization of peak heights".

Thus, research shows, discretization of mass protein data is starting to be accepted as a necessary preprocessing step to accurately reflect the noisy nature of the data and the non-homogeneity of the variance.

Classification

A myriad of classification algorithms have been used to discriminate subjects based on mass protein data, but only some of them are described in this section.

Two studies, both published in 2003 used tree based classification algorithms. Markey et al. (2003) used a classification and regression tree (CART) to predict the probability of each subject belonging to the healthy or diseased group. The second (Zhu 2003) used a wavelet transformation where peak height can be expressed as a function of the m/z ratio. A subset of the coefficients corresponding to the wavelet transformation created a classification tree. A classification tree uses a series of if-then rules to classify subjects. It also uses a recursive partitioning of the sample into increasingly more

homogeneous sub-samples in order to refine the if-then rules. Classification trees are impractical when they have too many variables. In both cases aggressive data reduction (less then 21 variables) was applied before the classification tree was be implemented.

Neville et al. (2003) used classification trees, logistic regression and linear discrimination in order to distinguish between healthy and diseased subjects. Each of these techniques required significant reduction in the dimension of the data before classification.

The LAD system of classification used by Alexe et al. (2004) relies on a method similar to a classification tree. Bounding conditions or 'cuttoff points' are defined for a small number of m/z ratios (the other m/z ratios are ignored). Based on the bounding conditions, each subject has a negative or a positive value for a given m/z ratio. In order to be classified into one group or another, a subject has to simultaneously meet a sufficient number of positive and negative conditions. This method works as long as there are a small number of m/z ratios that can accurately classify all subjects into their correct groups.

The Ugly Duckling Theorem

The Ugly Duckling theorem (Watanabe 1969) states that for any number of characteristics that are enumerated that are similar between a swan and a rose, an equal number of characteristics can be found in common between a duck and a swan. To state this theorem more generally, given all possible classifiers, a population can be grouped into any conceivable configuration. Only when certain classifiers are weighted more heavily than others do consistent groupings emerge. For example, if we choose to only weight feather color and presence of wings, a duck and a swan are identical.

However, once we choose to weight wingspan, a duck and a swan are different. Furthermore, if we weight retinal patterns, then each individual swan would be different and would be placed in its own group.

Conventional Wisdom

When the dimension of data is reduced, weights are being assigned to classifiers (variables) in such a way that the number of resultant variables with non-zero weight is reduced. The problem arises of how to weight the old variables to obtain the best set of new variables. Another potential problem of dimension reduction schemes is they weight variables based upon a researcher's assumptions of how the data should behave in addition to observed characteristics of the data.

When analyzing mass protein data, this 'Conventional Wisdom' can be very dangerous. In proteomics data it is rare to have more than 100 subjects. The dimension of the data is usually over 15,000. Most classification schemes require the number of variables to be well under the number of subjects. Almost all the classification schemes cited earlier in this paper used less than 25 variables in their analysis of mass protein data. This results in discarding or giving zero weight to thousands of observations that are potentially good classifiers.

Discarding potentially good observations is the case even with PCA where each new variable contains information from several old variables. As stated previously, the variance of m/z densities decreases with decreasing intensity. Thus, variables with very weak expression values can be good discriminators between groups. PCA will often give these variables with weak expression values little weight. Consequently, they are

discarded from the analysis by the researcher. This illustrates how assumptions inherent in PCA can lead a researcher to throw out good classifiers.

The argument made in this section is that conventional dimension reduction techniques make assumptions (conventional wisdom) that are often harmful to analysis of mass protein data. A good dimension reduction technique would be one which makes a minimum number of assumptions about the data.

Maximum Entropy

In 1948 Shannon defined Entropy (in an information theory sense) as:

$$S = -\sum p_i \log p_i \,, \qquad (1)$$

where p is the probability of a given state and the sum is over all possible states. If there are more states then S is bigger. If the states have more equal probabilities then S is bigger. Therefore, S is essentially a measure of the uncertainty in a system in which certain discrete states take on specific probabilities. In his 1957 paper, Jaynes showed that:

> "…in making inferences on the basis of partial information, we must use that probability distribution which has maximum entropy subject to whatever is known. This is the only unbiased assignment we can make; to use any other would amount to arbitrary assumption of information which by hypothesis we do not have."

A distribution which maximizes entropy relies only on the data and not on assumptions made by the researcher. Thus, a dimension reduction scheme that utilizes a maximum entropy distribution would be appropriate for mass protein data.

Information Partition Function

The Information Partition (IP) algorithm produces a distribution based on the principle of maximum entropy. The IP algorithm is a modification of the Grade of Membership (GoM) model. I will first describe the GoM model, then I will describe the Information Partition Function.

GoM Model

The Grade of Membership model was first proposed in 1974 by Woodbury and Clive. It was designed to model high dimensional categorical (discrete) data. Unlike most models which are used as tools to discover properties of a population, the GoM analysis tries to discover the properties of individuals, and thus predict the probability they will respond in certain way with respect to the discrete random variables.

The GoM model assumes there exist K pure types, or fundamental groups. These pure types are similar to the latent variables in latent variable analysis. Each of these pure types exhibits a characteristic behavior. For example, if the pure types were different diseases, the behavior for each pure type would be the presence or absence of the symptoms of that particular disease. The GoM model also assumes that individuals can belong to more than one pure type. Their grade of membership (hence the name) to each pure type is a value between 0 and 1. The total membership to all pure types for a given individual must sum to 1. The goal of the GoM model is to assign grades of membership from all pure types to each individual, thus specifying the relevant properties of all individuals.

GoM Likelihood

This paper deals with mass protein data, thus, it is convenient to refer to the categorical variables in the GoM model as m/z ratios and to their outcomes as intensities, assuming the intensities have been discretized. If the grades of membership for an individual are known, then the probability that a given individual has a certain intensity for a specific m/z ratio is known. The relationship that gives this probability according the GoM model is below (Woodbury et al. 1974):

$$P_{ijl} = \sum_k g_{ik} \lambda_{kjl} \ , \qquad (2)$$

where g_{ik} and λ_{kjl} are constrained as according the following relations:

$$g_{ik} \geq 0, \sum_k g_{ik} = 1 \qquad (3)$$

$$\lambda_{kjl} \geq 0, \sum_l \lambda_{kjl} = 1 \qquad (4)$$

Table 1 describes the terms in the model.

Table 1

i	-	index on individual observations
j	-	index on m/z ratios
k	-	index on pure types
l	-	category of descretized m/z ratio
g_{ik}	-	grade of membership of individual for each pure type
λ_{kjl}	-	probability of intensity l for the jth m/z ratio by an individual only of type k

The likelihood is easily derivable from this:

$$L = \prod_i \prod_j \prod_l \left(g_{ik} \lambda_{kjl} \right)^{y_{ijl}} \ . \qquad (5)$$

Y_{ijl} refers to the observed values. It is one for only one l in each j and zero for the other l's in each j because each individual had only one response to each m/z ratio.

The unknowns in the GoM model are the g_{ik}'s and the λ_{kjl}'s and K, the number of pure types. Kovtun et al. in 2004 determined that a solution for these unknowns can be approximated by maximizing the likelihood in (5).

Conditional Independence

The pure types cannot be measured directly. However, they can be estimated from the measurable variables associated with each individual. The grades of membership of each individual for each pure type can be represented in a space called gik space. The GoM model becomes a mapping function that transforms an individual from his measurable variable space to his Gik space. Manton et al. (1994) hypothesized that the coordinates of an observation in Gik space make that observation's coordinates in its measurable variable space conditionally independent. In other words, P(A,B|C) = P(A|C)P(B|C) where C is (A,B)'s position in gik space (Oliphant 2003).

The Information Partition Algorithm

In 2003 Oliphant derived the Information Partition algorithm. The IP algorithm is used to model the same discrete categories as the GoM model; however, unlike the GoM model, the IP incorporates maximum entropy into the development of the likelihood. Using the same notation and indexing as the GoM model, Oliphant first rewrote (1) as:

$$H = -\sum_{i} \sum_{j} \sum_{l} p_{ijl} \log(p_{ijl}) \quad . \qquad (6)$$

He then maximized (6) subject to the following constraints:

$$\sum_{l} p_{ijl} = 1 \tag{7}$$

$$\sum_{i} \sum_{j} \sum_{l} p_{ijl} g_{ik} w_{kjl} = E_k \,. \tag{8}$$

According to Oliphant, w_{kjl} represents a prior assumed distribution of a fixed amount of energy E of type K throughout cells jl. P_{ijl} and g_{ik} represent the same thing they do in equation (2). Referring equation (8) he said that:

> "[equation 8] can be thought of as the energy constraint and finds its thermodynamic analog in the equipartition theorem."

Maximizing (6) subject to (7) and (8) result in the following expression:

$$p_{ijl} = \exp\left(-\sum_{k} g_{ik} \lambda_k w_{kjl} \right). \tag{9}$$

Oliphant substituted this expression into the GoM likelihood which resulted in the Information Partition Function given below. In this likelihood, values of w_{kjl} were assumed unknown and combined with λ_k as λ_{kjl}.

$$L = \prod_{i} \prod_{j} \prod_{l} \exp\left(-\sum_{k} g_{ik} \lambda_{kjl} \right)^{y_{ijl}}, \tag{10}$$

where the estimates must meet the constraints:

$$\sum_{l} \exp\left(-\sum_{k} g_{ik} \lambda_{kjl} \right) = 1 \,, \sum g_{ik} = 1 \text{ and } \; g_{ik} \geq 0. \tag{11}$$

Oliphant showed that by maximizing (10) according to the constraints given in (11), estimates for g_{ik} and λ_{kjl} can be obtained. The IP algorithm is advantageous

because it produces a maximum entropy distribution subject to the assumption that some set of K groups exist with each individual having some level of membership in these groups. It also has the advantage that it generates a natural reduction in the dimension of the data. If there were only 50 subjects and 150 variables, the IP algorithm would use all the information from all 150 variables to assign grades of membership to each subject. While computational time provides a practical restriction to the number of variables, there is no theoretical limit to the number of variables the IP algorithm can incorporate.

The Moment Matrix

One of the problems of GoM analysis (and the IP algorithm by extension) is how many pure types are truly represented in the data. A partial solution to this problem was proposed by Kovtun et al. in 2004.

Grade of Membership analysis considers J m/z ratios, each ratio having L_j possible descretized intensities. A set of random variables $X_1..., X_j$, describe the probability of observing a specific intensity for each of the J ratios. However, $X_1..., X_j$ are different for each individual, depending on each individual's grades of membership to the K pure types. These differences can be described by a random variable β_{jl}. Let the probability that the ith individual has intensity l for the jth ratio be:

$$\Pr(X_j^i = l) = \beta_{jl}^i \quad . \tag{12}$$

Realizations of β_{jl} are distribution laws $(X^i{}_1..., X^i{}_j)$ for individuals. Let μ_β be a probabilistic measure giving the distribution of β_{jl}. In Kovtun's 2004 paper, he showed that the moments of μ_β can be used to calculate a lower bound for 'K', the

number of pure types in the data. These moments are:

$$M_l(\mu_\beta) = \int \prod_{j:l_j \neq 0} \beta_{jl_j} \mu_b(d\beta) \,. \qquad (13)$$

The $l_j \neq 0$ notation accounts for marginal distributions. If $l_j = 0$, we are not concerned with the jth ratio and its probability is not included in the product.

These moments are the probabilities of obtaining a given pattern of intensities. According to Kovtun, frequencies of the various intensity patterns in the data are consistent and efficient estimators for these moments. These estimators are organized into vectors and these vectors are formed into a matrix. Since not all intensity patterns are observable, missing values are present in this matrix. In order to get a lower bound on K, a part of the moment matrix which is complete is considered. Singular value decomposition is performed on this sub-matrix. From the singular values, a lower bound on the dimensionality of the entire moment matrix is obtained. This is a lower bound for K (the number of pure types).

The moments are the probabilities of obtaining a given pattern of intensities. The moments can be used to calculate a lower bound for K because in GoM analysis,

$$P_{ijl} = \sum_k g_{ik} \lambda_{kjl} \qquad \text{(equation 2)}$$

which is also the probability of obtaining a given pattern of intensities. However, in the IP algorithm, $P_{ijl} = \sum_k g_{ik} \lambda_{kjl}$ is not the probability of obtaining a given pattern of intensities. The probability of obtaining a given pattern of intensities for the IP algorithm

is given in equation (9). This suggests that for the IP algorithm, a lower bound for K can be determined by constructing the moment matrix by using the logs of the frequencies observed in the data instead of the frequencies themselves.

Chapter 3 - Methods

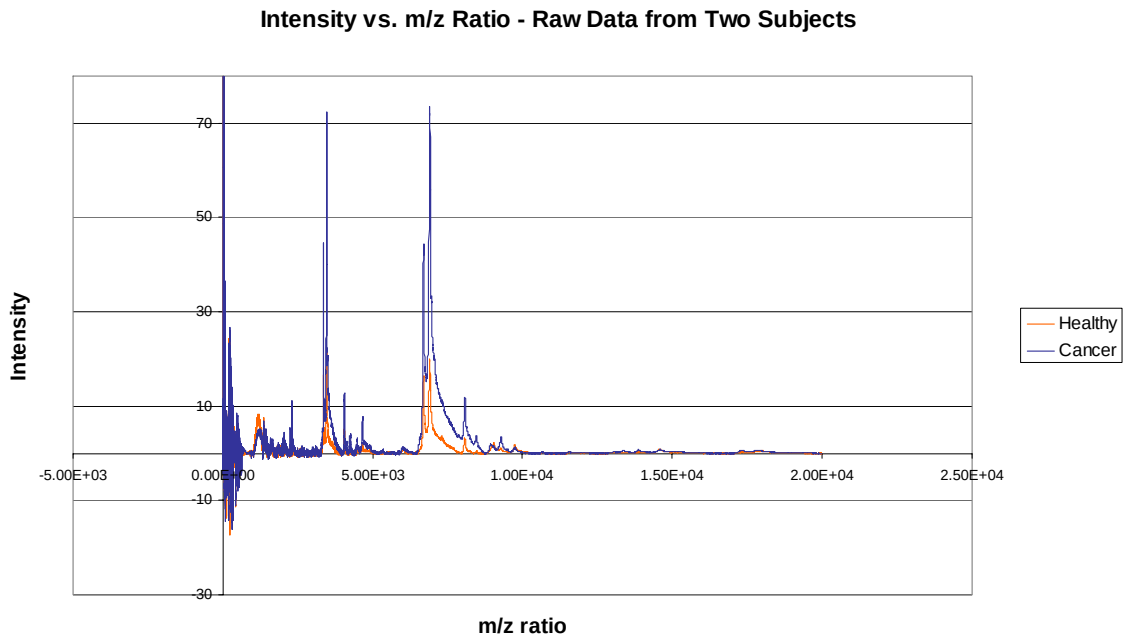
I have argued that it is appropriate to discretize mass protein data. Additionally I have argued that a classification scheme that incorporates the maximum entropy principle should be applied to mass protein data. A clustering algorithm that accepts discrete data and relies on the principle of maximum entropy is the Information Partition algorithm. Furthermore, the IP algorithm has no theoretical limit on the number of variables (information) that it can incorporate, which recommends it for use with mass protein data.

I took the mass protein data from 216 subjects. 100 of the subjects have malignant ovarian cancer. 16 of the subjects have benign ovarian cancer. 100 of the subjects have no cancer. Each subject has intensities for 15,154 distinct m/z ratios. Thus, there are 15,154 variables associated with each subject. Figure 2 displays the raw data from two subjects. One subject is from the cancer group and the other is from the healthy group.

I applied the IP algorithm to all 216 of the subjects. However, before I applied the IP algorithm on any of these subjects I conducted several preprocessing steps. The first step was to discard m/z ratios below 1000. At least some of the m/z ratios below 1000 were due to molecules in the solution in which the samples were placed before ionization. Furthermore, a high percentage of intensities below 1000 m/z were large negative numbers, which was true in all subjects. A large amount of noise is expected in mass protein data, however, large negative values are not. The m/z spectrums have almost no negative values above 1000 m/z; therefore I was not able to justify a baseline

correction that would have made the negative values positive. Consequently, I discarded all values below 1000 m/z in the data from the analysis.

Figure 2



Normalizing and Binning

The m/z spectrums for all subjects were normalized so the subjects could be directly compared to each other. This was done by dividing all intensities by a different constant for each subject. By dividing by this constant, the area under the curve of each spectrum was the same from subject to subject.

Next, I reduced the number of variables from 15,154 distinct m/z ratios to 190 variables. I did this by dividing the m/z spectrum for each subject into 190 bins of equal length. The length of each bin was 100 m/z ratio units. Each bin contained many intensities associated with many separate m/z ratios. The value assigned to each bin was

the average of the intensity values contained in the bin. There were 190 bins, resulting in 190 new variables. The goal was to choose bins small enough so the relevant features of the data (the peaks) were retained while filtering out noise and reducing the correlation among the variables. I also chose the number of bins greater or approximately equal to 216 (the number of subjects) to illustrate the usefulness of the IP algorithm in handling high dimensional data. Furthermore, reducing the data to 190 variables alleviated computational difficulties I was having reading such a large data set into memory. Figure 3 shows raw data from two subjects in the 8000 m/z to 9000 m/z range. Figure 4 shows the same subjects as figure 3 after the normalizing and binning operations have been performed.

Quantization of Data

After normalizing and binning the data, I quantized the intensities, by using two discretization schemes. In one scheme I used five discrete levels. In the other I used seven discrete levels. Because the variance of peaks changes with peak height, the

Table 2

Range of Intensity	Discrete level assigned
$0 \le y < .5$	1
$.5 \le y < 1$	2
$1 \le y < 2$	3
$2 \le y < 4$	4
$4 \le y < 8$	5
$8 \le y < 16$	6
$16 \le y$	7

Table 3

Range of Intensity	Discrete level assigned
$0 \le y < .75$	1
$.75 \le y < 1.5$	2
$1.5 \le y < 5$	3
$5 \le y < 15$	4
$15 \le y$	5

spacing of the levels was exponential rather than linear. The discretization schemes are given in tables 2 and 3.

Figure 3

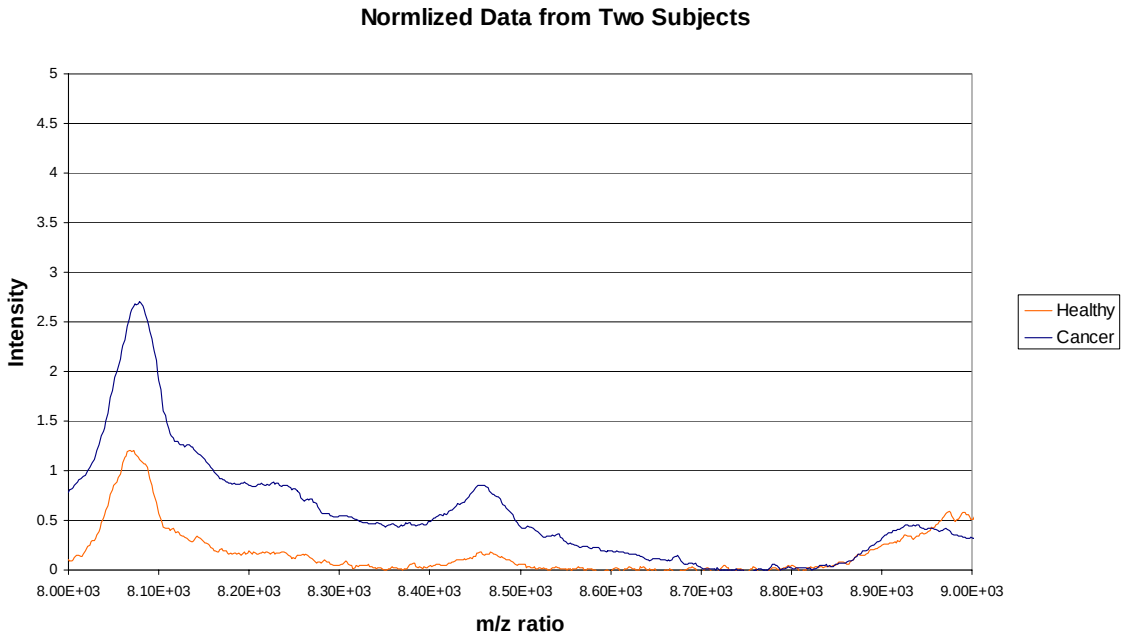
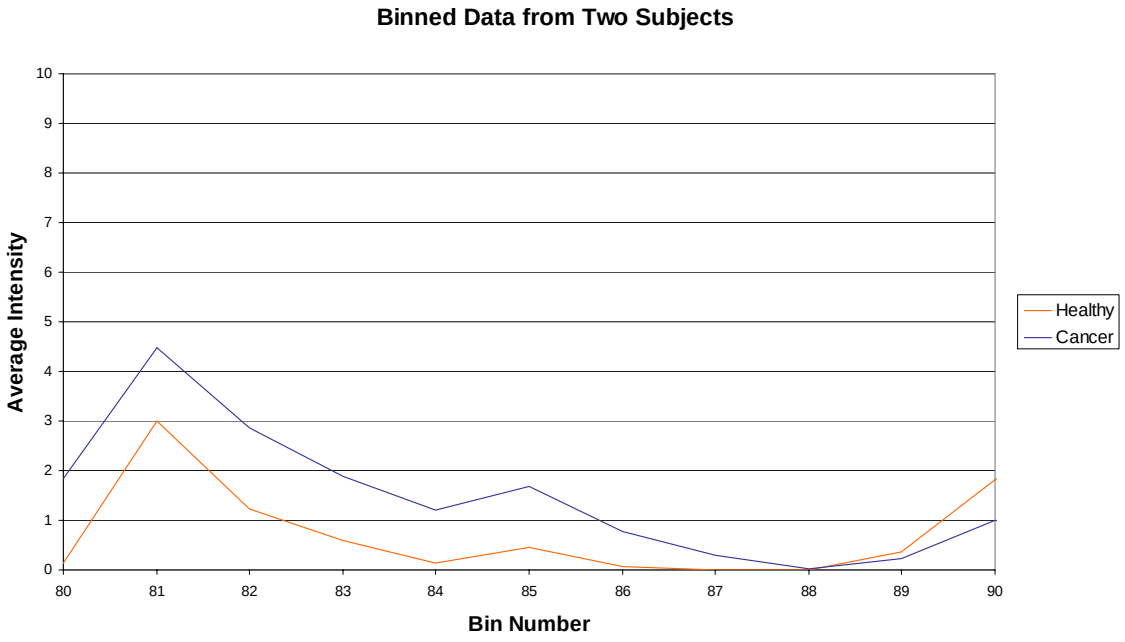


Figure 4



The Moment Matrix

After quantization of the data, K was approximated by dividing the moment matrix into three distinct sub-matrices. Each of these sub-matrices contained moments relating to 14 of the 190 bin variables. These 14 variables were chosen randomly for each sub-matrix with each sub-matrix having 49 rows and 50 columns.

Singular value decomposition was applied to each sub-matrix. Plots of the singular values are given in appendix A. Analyzing the singular values revealed the dimensionality of the moment matrix was at least three, but as large as five. Thus, there was strong evidence for at least three pure types, but possibly as many as five.

IP Algorithm

The IP algorithm was applied to all 216 subjects. In applying the IP algorithm, two things had to be considered. First there were at least 3 pure types, but as many as five; consequently the IP algorithm was to be applied to the subjects using K=3 and K=5. Secondly, the IP algorithm had to be applied to the subjects using both discretization schemes (with five and seven levels). In order to cover all combinations of these two considerations, the IP algorithm was applied in four separate cases. Table 4 gives each combination of the two considerations for each case.

Table 4

Case 1	K=5	7 discrete levels of intensity
Case 2	K=3	7 discrete levels of intensity
Case 3	K=5	5 discrete levels of intensity
Case 4	K=3	5 discrete levels of intensity

The algorithm produced a set of scores for each subject detailing the grade of membership that each subject had to each pure type, which are the gik scores. Once the

gik scores were obtained, a logistic model was fit to 170 of the subjects. 80 of these 170 subjects came from the healthy group, 80 from the malignant tumor group and 10 from the benign tumor group.

The purpose of the logistic model was to predict, based on the gik scores, whether a subject belonged in the healthy group, malignant tumor group or benign tumor group, coded as 1, 2 and 3 respectively. Once the logistic model was fitted to 170 subjects, it was used to classify the remaining 46 subjects based on their gik scores. The misclassification rate of the 46 holdout subjects was used to gage the effectiveness of the IP algorithm in reducing the dimension of the data without discarding relevant information.

Partial Least Squares

A partial least squares (PLS) algorithm was applied to the all the subjects. The purpose of the PLS application was to have a dimension reduction algorithm commonly used with mass protein data that could be compared to the IP algorithm. The PLS algorithm extracted nine variables from the original 190 bin variables once the m/z ratios of the 190 bin variables had been quantized. Four logistic models were fit to these variables using only 170 of the subjects. These were the same 170 subjects fit to the logistic model using the gik scores. The first model used only three of the nine variables. The second, third and fourth models used five, seven and nine of the variables respectively. The response was coded as 1, 2 or 3 depending on whether the subject was classified in the healthy, malignant tumor or benign tumor group.

By fitting the logistic models to 170 subjects, they classified the remaining 46 subjects based upon the variables produced by the PLS algorithm. The misclassification

rate of the 46 holdout subjects was used to gage the effectiveness of the PLS algorithm in reducing the dimension of the data without discarding relevant information. The effectiveness of the IP algorithm was then compared to PLS algorithm based on their respective misclassification rates.

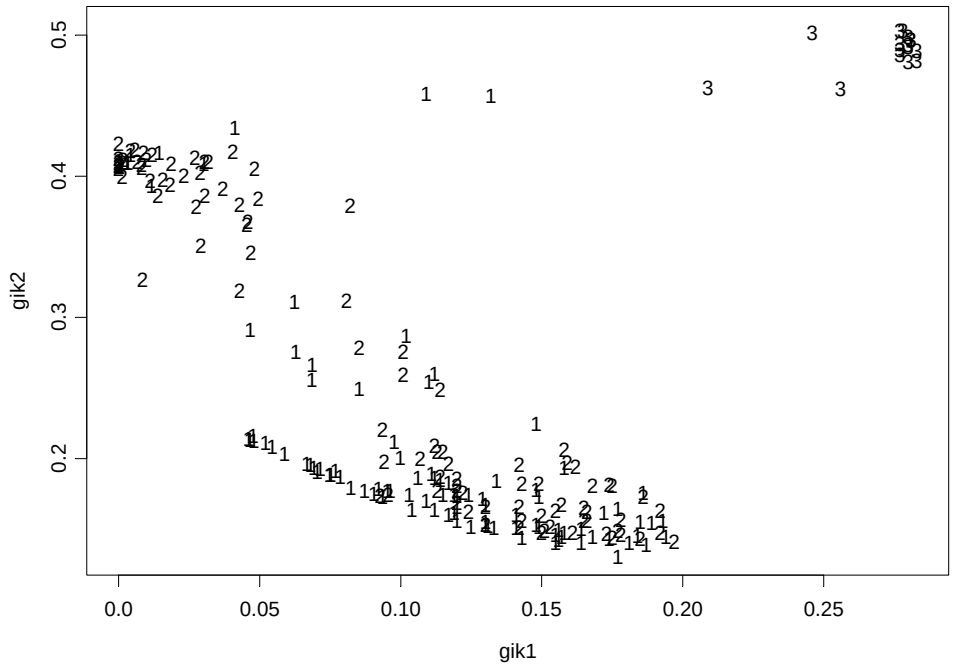
All logistic models using both giks and PLS components were implemented in SAS using PROC LOGISTIC. In each case the specific logistic model used was a cumulative logit model.

Chapter 4 – Results

IP Algorithm

The IP algorithm was applied to the data in four cases. Figure 5 shows gik1 versus gik2 for case 1. Gik1 and gik2 are the grades of membership for each subject to the first and second pure types respectively. Gik1 and gik2 were chosen for the plot because they distinguished between tumor and non tumor groups better than other pure types for case 1.

Figure 5 – Gik1 versus Gik2 for Case1
1 = Healthy, 2 = Malignant Tumor, 3 = Benign Tumor



In figure 5, gik1 and gik2 separate the benign tumor group completely from the other two groups, as well as separating the majority of healthy and malignant tumor subjects from one another. Figures showing graphical separation of groups for the other three cases are given in Appendix B.

The IP algorithm produces gik scores by maximizing the likelihood associated with the IP algorithm (equation 10). Table 5 displays the -2 log-likelihood for all four cases as well as giving the rms scores for each case. The rms score is the root mean squared error associated with the constraints.

Table 5

Case	-2loglikelihood	rms
1	25065	.007016
2	23388	.00567
3	14255	.00612
4	19702	.00841

Table 6 gives the misclassification rates from the logistic model applied to data from all four cases. Misclassification rates are for the holdout group of 46 subjects whose responses were not used in the construction of any of the logistic models.

Table 6

Case	Total Misclassification Rate	Malignant Tumor Misclassified	Healthy Misclassified	Benign Tumor Misclassified
1	11%	15%	10%	0%
2	17%	15%	20%	15%
3	26%	45%	15%	0%
4	13%	15%	10%	13%

Case 1 (K=5 and 7 discrete groups) yielded the best classification performance. However, Case 4 (K=3 and 5 discrete groups) yielded very similar results. The worst scenario was case 3 (K=5 and 5 discrete groups).

Table 7 gives average misclassification rates for pure types and the number of discrete levels.

Table 7

Category	Average Misclassification Rate	Average Misclassification Rate – Malignant Tumor
K=5	18.5%	30%
K=3	15%	15%
7 Discrete levels	14%	15%
5 Discrete levels	19.5%	30%

PLS Algorithm

The PLS extracted nine components from the 190 variables. Four logistic models were fit using PLS components. Model 1 used three components, model 2 used five components, model 3 used seven components and model 4 used nine components.

Table 8 gives the misclassification rates from the logistic model applied to data from all four models. Misclassification rates are for the holdout group of 46 subjects whose responses were not used in the construction of any of the logistic models.

Table 8

Model	Total Misclassification Rate	Malignant Tumor Misclassified	Healthy Misclassified	Benign Tumor Misclassified
1	24%	45%	10%	0%
2	17%	30%	10%	0%
3	15%	25%	10%	0%
4	24%	20%	15%	67%

Model 3 (7 components) had the best performance with an overall misclassification rate of 15%.

Discussion

Cases with three pure types exhibited a lower average misclassification rate than cases with five pure types. This is reasonable because three main clinical differences

within the subjects are obvious: healthy, malignant tumor and benign tumor. The IP algorithm recognized the differences between these three groups and accounted for them.

Cases with 7 discrete levels had lower average misclassification rate than cases with 5 discrete levels. This suggests that seven or more discrete levels should be used when applying the IP algorithm to mass protein data.

Table 9 compares misclassification rates of the IP algorithm and the PLS algorithm.

Table 9

Method	Total Misclassification Rate (average)	Total Misclassification Rate (best scenario)	Malignant Tumor Misclassified (average)
IP Algorithm	16.75%	11%	22.5%
PLS	20%	15%	30%

In all three categories the IP algorithm compares favorably with the PLS algorithm, which shows that the IP algorithm performs better than a commonly used dimension reduction algorithm applied to mass protein data.

Regarding the IP algorithm, case 1 which had the best performance had the lowest log-likelihood. However, case 3 had the poorest performance but had the highest log-likelihood.

Chapter 5 – Conclusion

The IP algorithm was compared to the PLS algorithm with regard to mass protein data. Both algorithms were used to reduce the dimension of the data. The IP algorithm was applied in four settings by extracting three and five variables from the data and using both five and seven pure types. The IP algorithm was applied four times to account for all four combinations of the different settings. For each setting, logistic regression was applied to the extracted variables to classify each subject in the healthy, malignant tumor or benign tumor group.

The PLS algorithm extracted nine variables. In four separate settings three variables were used, followed by five, seven and nine variables. For each setting, logistic regression was applied to the extracted variables in order to classify each subject in the healthy, malignant tumor or benign tumor group.

Using misclassification rates from the logistic regression, the IP algorithm performed better than the PLS algorithm, which demonstrates the potential usefulness of using an algorithm based on maximum entropy (minimum assumptions) with mass protein data. Using seven discrete levels instead of five resulted in lower misclassification rates with regard to the IP algorithm, suggesting that at least seven discrete levels of intensity should be considered when applying the IP algorithm to mass protein data.

Bibliography

Alexe G., Alexe S., Liotta L., Petricoin E., Reiss M., Hammer L. P., (2004), "Ovarian Cancer Detection by Logical Analysis of Proteomic Data," *Proteomics*, 4, 766-783.

Boros E., Hammer P., Ibaraki T., Kogan A., Mayoraz E., Muchnik I., (2000) "An Implementation of Logical Analysis of Data," *IEEE Transactions on Knowledge and Data Engineering*, 12, 2, 292-306.

Crama Y., Hammer P. L., Ibaraki T., (1988) "Cause-Effect Relationships and Partially Defined Boolean Functions," *Annals of Operations Research*, 16, 299-326.

Dancik V., Addona T., Clauser K., Vath J., Pevzner P., (1999) "De Novo Peptide Sequencing Via Tandem Mass Spectrometry," *Journal of Computational Biology*, 6, 327-342.

Hilario M., Kalousis A., Muller M., Pellegrini C., (2003), "Machine Learning Approaches to Lung Cancer Prediction from Mass Spectra," *Proteomics*, 3, 1716-1719.

Jaynes, E. T., "Information Theory and Statistical Mechanics", (1957) *Department of Physics, Standford University.*

Kovtun M., Akushevich I., Manton, G. K.,Tolley D., (2004), "Grade of Membership Analysis: Newest Development with Application to National Long Term Care Survey Data".

Lee K., Lin X., Park D., Eslava S., (2003) "Megavariate Data Analysis of Mass Spectrometric Proteomics Data Using Latent Variable Projection Method," *Proteomics*, 3, 1680-1686.

Markey M. K., Tourassi D. G., Floyd E. C., (2003), "Decision Tree Classification of Proteins Identified by Mass Spectrometry of Blood Serum Samples From People With and Without Lung Cancer," *Proteomics*, 3, 1678-1703.

Neville P., Tan P., Mann G., Wolfinger R., (2003), "Generalizable Mass Spectrometry Mining Used to Identify Disease State Biomarkers from Blood Serum," *Proteomics*, 3, 1710-1715.

Oliphant R. J., (2003) "The Information Partition Function," *Department of Statistics, Brigham Young University.*

Petricoin E., Ardekani A., Hitt B., Levine P., Fusaro V., Steinberg S., Mills G., Simone C., Fishman D., Kohn E., Liotta L., (2002), "Use of Proteomic Patterns in Serum to Indentify Ovarian Cancer," *Mechanisms of Disease*, 359, 572-577.

Shannon C. E., Weaver W., (1949), *The Mathematical Theory of Communication*, Urbana, Illinois: University of Illinois Press.

Tolley D., Oliphant J., Fellingham W. G., (2004) "High Dimensional Categorical Models using Conventional Wisdom." Department of Statistics, Brigham Young University.
Watanabe S., (1969) *Knowing and Guessing*, New York: John Wiley & Sons, Inc.

Woodbury M., Clive J., (1974), "Clinical Pure Types as a Fuzzy Partition," *Journal of Cybernetics*, 4, 3, 111-121.

Zhu H., Yu, Chang-Yung X., Heping Z., (2003), "Tree-Based Disease Classification Using Protein Data," *Proteomics*, 3, 1673-1677.

Appendix A

The three following graphs contain the singular values for the 3 sub-matrices of the moment matrix. In each graph, the first singular value is over 400 and is not plotted so that the lower singular values can be seen more easily.

Figure A.1

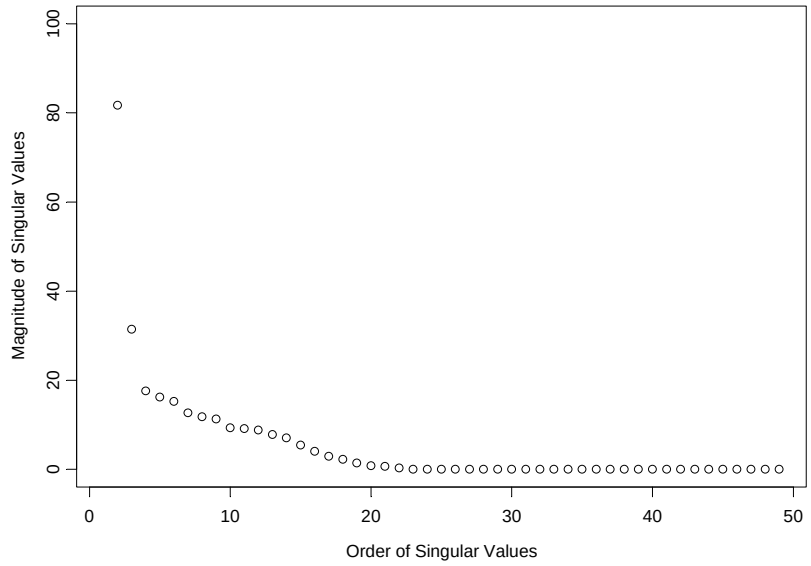


Figure A.2

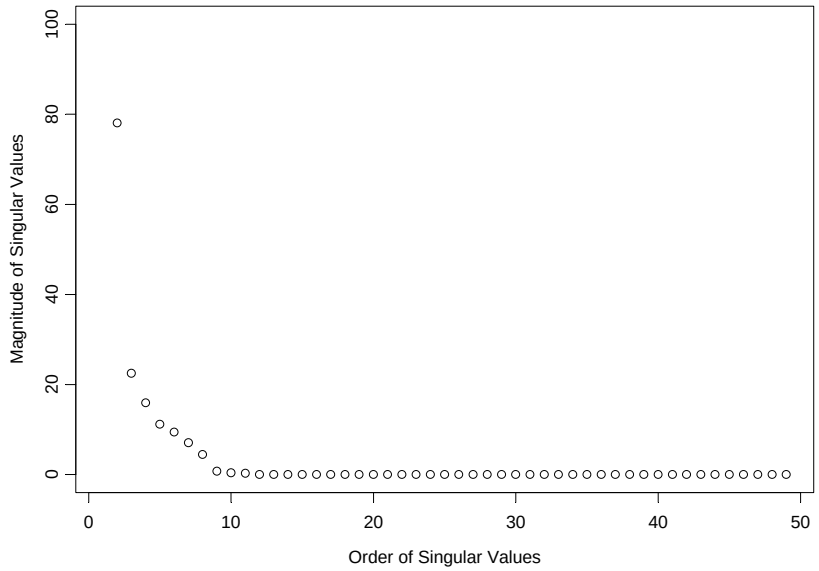
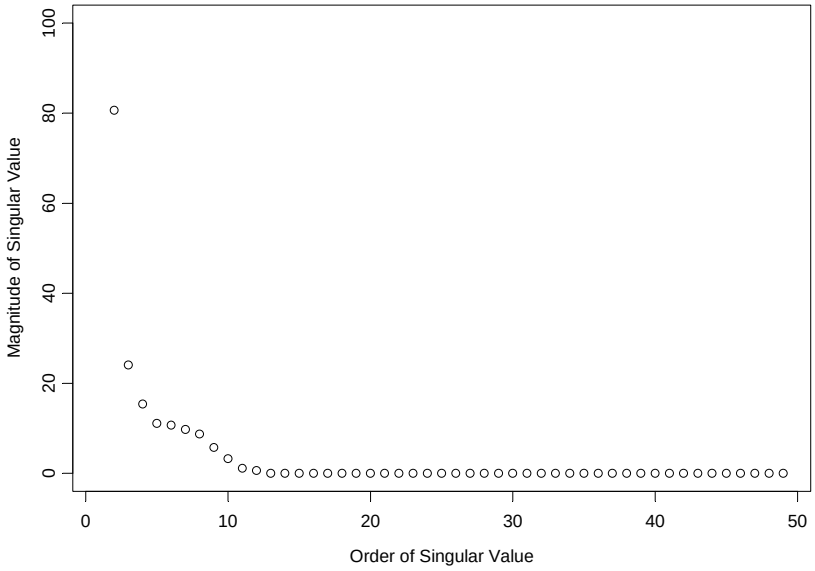


Figure A.3


Magnitude of Singular Value
0
20
40
60
80
100
0
10
20
30
40
50
Order of Singular Value

Appendix B

Gik1 versus gik3 for Case 2. For case 2, the gik scores were clustered in two groups. The distance between the two groups was to large to graph showing both groups because it excluded details within the groups. Therefore, the two separate graphs below, each show gik1 versus gik3 for one of the two groups.

For all graphs: 1 = normal, 2 = malignant tumor, 3 = benign tumor

Figure B.1 - Zoom in on one of two groups for case 2.

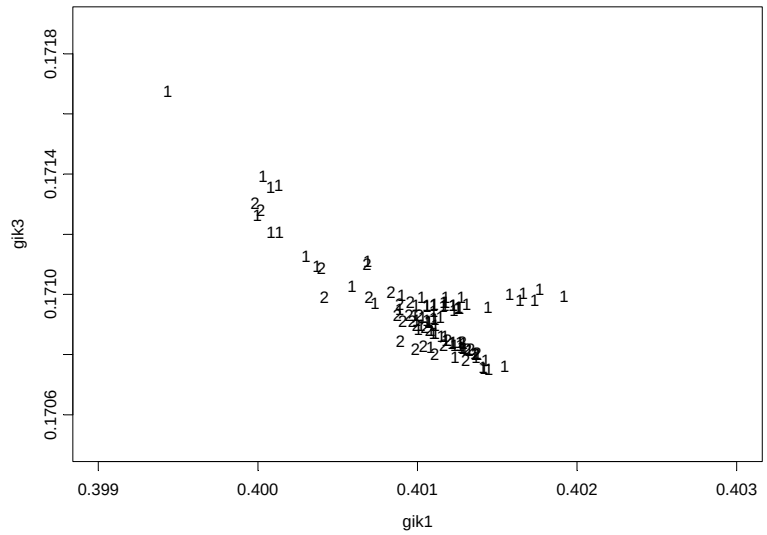


Figure B.2 - Zoom in on second of 2 of two groups for case 2.

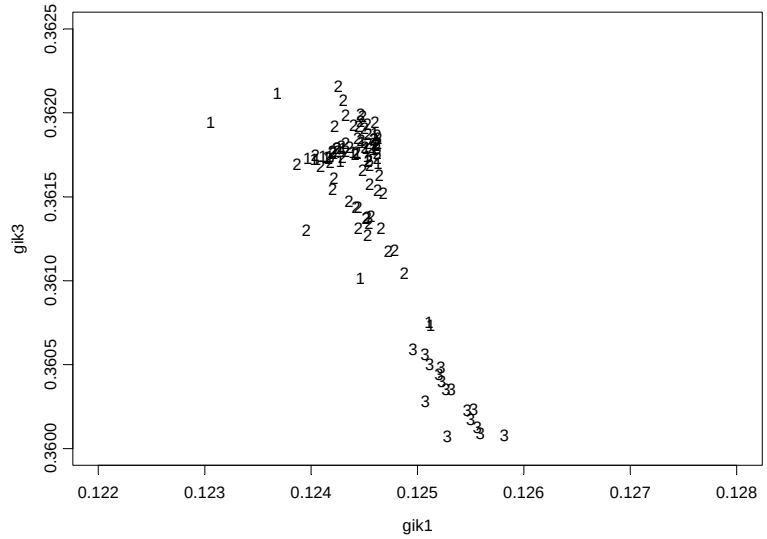


Figure B.3 - Gik4 versus gik4 for case 3.

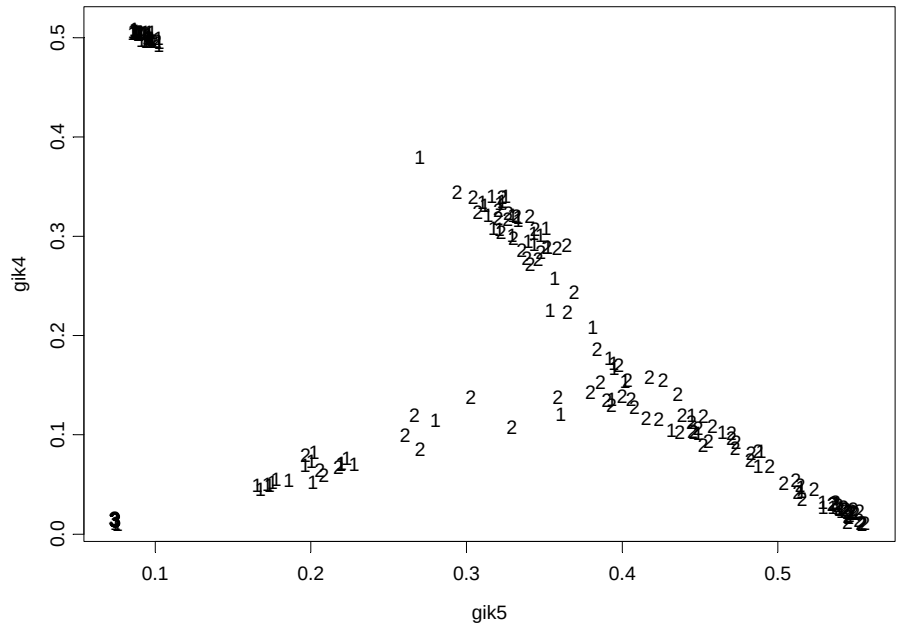
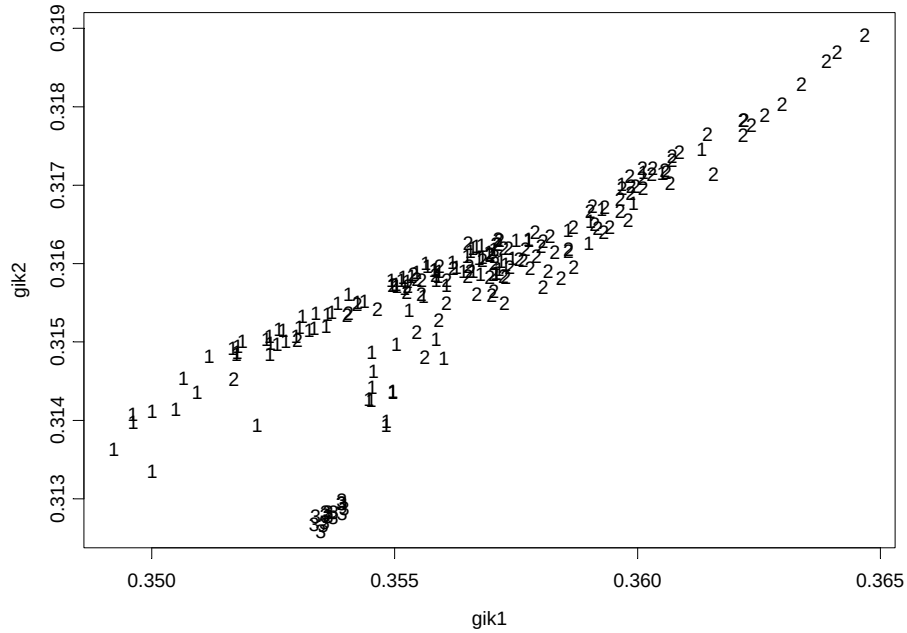


Figure B.4 - Gik1 versus gik2 for case 4.



For each case, the giks plotted in the graphs were the ones the tended to graphically separate the tumor and healthy groups more effectively.

Appendix C

C.1 - Commands and parameters used to execute IP algorithm in R.

```
library('Ripf')
answers<-ipf(state,y,trace=1,debug=1,w=1,maxiter=100,tol=.01)
```

C.2 – Code for PLS analysis in SAS

```
proc PLS data=PLS.test nfac=9 details;
model response2 = col11-col200/solution;
output out=load xscore=x yscore=y predicted=p_response;
run;

data temp;
set load;
keep _name_ response2 x1 x2 x3 x4 x5 x6 x7 x8 x9;run;

data temp1;
set temp;
if (_name_ > 80 & _name_ < 101) then response2 = .;
if (_name_ > 180 & _name_ < 201) then response2 = .;
if (_name_ > 210 & _name_ <= 216) then response2 = .;
if (_name_ > 80 & _name_ < 101) then _name_ = 1000;
if (_name_ > 180 & _name_ < 201) then _name_ = 1000;
if (_name_ > 210 & _name_ <= 216) then _name_ = 1000;
run;

proc sort data=temp1;by _name_;run;

proc logistic data=temp1;
model response2 = x1 x2 x3 x4 x5 x6 x7 x8 x9;
output out=log predprobs=I;
run;

data temp2;set log;keep IP_1 IP_2 IP_3;run;
proc print data=temp2;run;
```