Classifying Colon Cancer Tissues Using Probabilistic Principle Components Analysis with the Consistent Information Complexity (CICOMP) in Logistic Regression

Abdulaziz Saud Alkabaa
Department of Industrial Engineering
King Abdulaziz University
Jeddah, Saudi Arabia
aalkabaa@kau.edu.sa

Abstract

Genetic science has become one of the most important areas of research study for classification and clustering and building predictive models in the light of very highly correlated and noisy gene expression cancer data. Thus, the difficulties and complexity arise in the problem of visualizing data that have many variables. One of the goals of genetic research is to better understand and discover the mechanisms of a disease so that new treatment approaches and preventative measures can be proposed.

This paper proposes a novel heuristic procedure for modeling and classifying high dimension correlated variables by applying a Kernelized Probabilistic Principle Component Analysis (KPPCA) as a dimension reduction method using the estimated underlying kernel probability density of the gene expressions. The Gaussian kernel function along with the Asymptotic Mean Integrated Squared Error (AMISE) is used as optimal criterion to select the bandwidth for smoothing purposes, for 2000 genes for colon cancer tissues collected on 62 patients. This paper uses the Consistent Information Complexity (CICOMP) of Bozdogan, for the purpose of selecting the best numbers of probabilistic principle components (PPCs), which are linear combinations of the original variables. Later, a logistic regression model, as a classifier, is applied to describe the relationship between the patient types and the responsible cancer genes on the first 200 genes after applying PPCA. This paper shows a noticed improvement of sensitivity and specificity for the proposed logistic regression. The proposed heuristic steps in this paper is suggested for investigating such a problem of interests or can be extended in any industry problems. Moreover, the proposed heuristic procedures can be applied for multinomial logistic regression and liner regression model as well.

Keywords
Colon cancer classification, Logistic regression, Variables reduction, Probabilistic principle component analysis with Information complexity, and gene expression statistical research.

1. Introduction

Pathology is one of the most and critical fields in medical sciences. Many researchers have done research in this area to understand and discover causes of different cancer disease types and to discover patterns to improve the
causes associated with these diseases. As stated by the Stanford Cancer Institute (online 2019) on how genes cause cancer, the formation of tumors basically results from cell growth that gets out of control. In the human genome, there are many different types of genes that control cell growth in a very systematic, precise way. When these genes have an error in their DNA code, they may not work properly, and are said to be "altered" or mutated. An accumulation of many mutations in different genes occurring in a specific group of cells over time is required to cause malignancy. The different types of genes, that when mutated, can lead to the development of cancer are described below. It takes mutations in several of these genes for a person to develop cancer. What specifically causes mutations to occur in these genes is largely unknown. However, mutations can be caused by carcinogens (environmental factors known to increase the risk of cancer). The development of mutations is also a natural part of the aging process.

Vieira (2014) discusses one of the goals of genetic research to better understand the mechanisms of disease so that new treatment approaches and preventative measures can be proposed.

Technology has come a long way in this regard, and it is currently possible to simultaneously investigate almost one million sites in any individual's genomic DNA with the goal of finding associations between a given disease and genetic variation. However, technological advances have also created new problems and challenges for the scientists, such as how best to analyze the millions of data points involved in genetic studies of disease, how to deal with undersized high dimensional gene expression data sets.

Since the expression of genes is measured by the DNA microarrays, it can be inferred that there is an important necessity to create several analytical approaches to utilize and to analyze the information stored in gene expression data Raychaudhuri et al. (2000) and Yeung and Ruzzo (2001). With the enormous growing of various studies in the literature of gene expression data from microarrays being produced, new methodologies for prediction, classification, and clustering techniques are being studied and presented for the analysis of such data sets Yeung and Ruzzo (2001). For more details regarding methodologies and techniques that deal with gene expression data from the prospective of data mining, the readers are suggested to refer to Pamukçu et al. (2015) Raychaudhuri et al. (2000), Yeung and Ruzzo (2001), Chen et al. (2008), Yang et al. (2009) Ma and Kosorok (2009), and Nyamundanda et al. (2010).

2. Problem Statement

Classical mathematical and statistical models must be improved to accommodate the growing amount of data generated by today's genetic research. Because researchers now understand that the genetic contribution to many disease types is complex and that the same disease does not manifest in the same way in all people, descriptions that involve gradients of sickness and health are typically more effective than those that classify individuals as either "sick" or "healthy". Subsequently, this paper considers the well-known benchmark colon cancer dataset from Alon et al. (1999) as our illustrative example in our analysis. The dataset consist of 2000 genes with 22 normal and 40 cancerous tumors of colon tissues, totaling of 62 colon tissues. This is truly an undersized data set as most other benchmark gene expression data sets are. Our goal here is to introduce kernel density estimation using the Gaussian kernel along with choosing the optimal bandwidth. Motivation of kernelization of the data stems from the fact that such data sets are highly noisy, hence our goal is to smooth the data as our pre-processing tool. The proposed heuristic solution is coded by using Matlab function developed by Bozdogan (2015) to carry out the density estimation and then use the Consistent Information Criteria (CICOMP) of Bozdogan (2014) in Probabilistic Principal Component Analysis (PPCA) to reduce the dimension of the data set. In order to validate and practically represent the proposed methodology, a logistic regression model is applied and fitted on the first 200 genes after applying PPCA to choose the best subset of PCs as our new set predictors.

3. Scientific Background

Kernel density estimation (KDE) is a non-parametric approach to estimate the probability density function of a random variable. Kernel density estimation is a fundamental data smoothing tool which has many applications. A kernel is a non-negative real-valued integrable function K satisfying the following two conditions:

\[ \int_{-\infty}^{\infty} K(u) \, \, \, (1) \]
The first condition ensures that the method of kernel density estimation results in a probability density function. The second condition ensures that the average of the corresponding distribution is equal to that of the sample used. If \( K \) is a kernel, then so is the function \( K^* \) defined by \( K^*(u) = \lambda K(\lambda u) \), where \( \lambda > 0 \) a scalar. This can be used to select a scale that is appropriate for the data.

In our model the Gaussian kernel (function) as in equation (3) below is used with the optimal bandwidth estimator given in equation (4). For an illustrative purpose, Figure 1 (wand 1995) shows the shapes of different kernel functions.

\[
K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \quad \quad (3)
\]

\[
h_{AMISE} = \frac{R(K)^{1/5}}{m_2(K)^{5/8}n^{5/8}R(f''/2)^{1/5}n^{1/5}} \quad \quad (4)
\]

where \( m_2(K) = \int x^2K(x)dx \) and \( f \) is a density function.

After smoothing the data set as our pre-processing approach using the KDE on the 2000 genes, we use the new smoothed transformed data set and carry out the dimension reduction using the PPCA. The consistent information complexity (CICOMP) in the PPCA model is used to reduce the dimension. CICOMP for the PPCA model is given by

\[
CICOMP(PPCA) = n \sum_{j=1}^{m^*} \log(\lambda_j) + n(\pi - m^*)\log(\hat{\sigma}^2) + k[\log(n) + 1] + 2C_{1F}(\Sigma) \quad \quad (5)
\]

where

\[
m^* \leq m \text{ is the number of nonzero eigenvalues of the estimated PPCA covariance matrix},
\]

\[
\hat{\sigma}^2 = \frac{1}{p-m^*} \sum_{j=m^*+1}^{p} \lambda_j \quad \text{and}
\]

\[
k = m^*p + 1 - m^*(m^*-1)/2 \text{ is the number of parameters estimated in the model}. \]

For more details on CICOMP, for more information please see Pamukcu et al. (2015).

As the last phase of our analysis, we fit a logistic regression model with the forward stepwise subset procedure on the PPCs for the first 200 genes to determine the best subsets to search for the most significant genes that are causing the colon cancer. The cumulative distribution function (cdf) of the logistic function is given in by

\[
F(t) = \frac{e^t}{1+e^{t+1}} = \frac{1}{1+e^{-t}}. \quad \quad (7)
\]

Assuming \( t \) as a linear function of an explanatory variable (or of a linear combination of explanatory variables), the simple logistic function can be written as
\[ F(x) = \frac{1}{1+e^{-(\beta_0 + \beta_1 x)}} \]  

(8)

This will be interpreted as the probability of the dependent variable equaling a "success" or "case" rather than a failure or non-case. The inverse of the logistic function is given by

\[ g(x) = \log \left( \frac{F(x)}{1-F(x)} \right) = \beta_0 + \beta_1 x. \]  

(9)

The exponential function is the antilog for the logarithm using the natural log scale. Equivalently, equation (9) becomes:

\[ \frac{F(x)}{1-F(x)} = e^{\beta_0 + \beta_1 x} \]  

(10)

### 3.1 Proposed Approach

As mentioned above, this article proposes to use the non-parametric density estimation via the kernel density estimation to smooth our data. Figure 2 illustrates the comparison between parametric density approach with that of the kernelized density.

![Figure 2. Normal distribution and kernel density plots](image)

Figure 2. Normal distribution and kernel density plots

Figure 2-a displays that a normal underlying distribution on the data which shows non-normality. In contrast, figure 2-b shows smoothed data after applying the Gaussian kernel with optimal bandwidth value \( h=77.5741 \). In addition, the QQ-plot for the data shows that the majority of the data points fall near or on the upper 95% boundary when fitting the normal distribution on the original data. In addition, the histogram plot is skewed to the right indicating that the data set is not normally distributed. Comparing this with the plot in figure 2-b, we see that kernel estimation gives more confidence for the proposed kernel density estimation.

Consequently, based on the smoothed kernelized data set, the PPCA is carried out to reduce the dimension in the data set to approximate the best number of PPCs that can be extracted.

### 4. Analysis and Results

The quality of a kernel estimate depends less on the shape of the kernel function \( K \) than on the value of its bandwidth \( h \). It is important to choose the most appropriate bandwidth as a value that is too small or too large is not useful. It has been noticed that small values of \( h \) lead to very spiky estimates (not much smoothing) while
larger \( h \) values lead to over smoothing. Figure 4 shows different bandwidths for the first 9 genes. When the bandwidth is too small (very narrow) then the kernel density estimate is said to be under smoothed as the bandwidth is too small. When the band width is too large then it is said to be over smoothed and this has obscured most of the data’s structure. This brings to the problem of how to choose the bandwidth or the smoothing parameter \( h \). The Kernel density with Gaussian kernel is applied for the whole 2000 genes. For space considerations, the result for one gene with optimal bandwidth value \( h \) and its corresponding histogram plot is shown in figure 3.

![Kernel Density Estimate with \( h = 581.3818 \)](image)

Figure 3. Plot of kernel density for different band width \( h \)

As it is shown in figure 3 we can visually observe that the shape has an optimum value for the bandwidth \( h = 583.3818 \)

### 4.1 PPCA with CICOMP using Kernelized Data

The PPCA is used on the smoothed kernelized data set using CICOMP as our yardstick to determine the number of PPCs that should be retained. Table 1 shows the minimum of the information criteria at 31 PPCs.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Minimum Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CICOMP</td>
<td>-6.8073e+6</td>
</tr>
<tr>
<td>AIC</td>
<td>-7.0001e+6</td>
</tr>
<tr>
<td>CAIC</td>
<td>-6.8077e+6</td>
</tr>
</tbody>
</table>

Table 1 represents the minimum of the calculated criterion. Consequently, the number of PCs that should retain in the reduction scope is evidently shown by the plot in figure 4 which displays the minimum score for CICOMP in y-axis corresponds to the number of PPC in x-axis.
As it can be seen from figure 4 that the minimum score occurs on the 31 PCs. Thus, the first 31 PCs have been chosen as a dimension reduction. In contrast, the traditional way of computing the PCA has selected the first 10 PCs as dimension reduction with cumulative variances or eigenvalues that explain 95.5% of the total variance. The minimum score from CICOMP, has different number of PPCs that should be retained (31 PPCs) to be considered. Consequently, the information criterion here recommends an optimal way to describe the number of PPCs in the analysis and considering the amount of the information that remain in other PPCs. On the other hand, if the classical PCA had been relied with only the first 10 PCs would have been ignored all other PPCs that might have a lot important information in our case, specially the type of such a problem we have in this research. The loading factor has been studied the correlation between the 2000 genes and each 31 PCs. As a result, it has been founded that the most PPCs that have many loadings are PPCs number 2 and 14, which incorporate the most of genes. Therefore, the CICOMP has significantly incorporated the most significant PPCs in its consideration.

4.2 Logistic Regression Based on PPCA

This section describes the logistic regression as a classification method with Logit link function for mean response on the first 200 genes such as example, as follows:

- Applying PPCA on the first 200 genes for the seeking of dimension reduction
- Getting the PCs that represent 97% of total variance explained by PCs; which are numbers 1-30
- Applying stepwise forward selection method with tissue type (tumor=1, normal=0) as a response in this model and get the best subsets based on P-value for each predictor that has error rate less than 5%
- Carry out the best subset and run the logistic regression again with the new predictor set
- Test the significance of the estimated parameters βj for each predictor (PCs) based on Likelihood Ratio test and Wald test
- Study the loading factors of these significance predictors to explore the most effective genes that cause the tissue cancer.

The best subset is PCs number 2, 14, and 17 as appears in Table 2 below.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
<th>Df</th>
<th>ChiSq</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.76642631</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>PC 2</td>
<td>-0.1599852</td>
<td>1</td>
<td>4.201202</td>
<td>0.0404</td>
</tr>
<tr>
<td>PC 14</td>
<td>0.67234815</td>
<td>1</td>
<td>5.444746</td>
<td>0.01963</td>
</tr>
<tr>
<td>PC 17</td>
<td>-0.559081</td>
<td>1</td>
<td>4.043451</td>
<td>0.04434</td>
</tr>
</tbody>
</table>
The using of the logistic regression is to study the relationship between the PCs and the likelihood of tissue type. As it mentions in table 3 below that it is indeed analogues to the liner model in exploring the significance of a model, however, this test follows chi-sq X2 distribution. To judge of the model significance the error rate of 5% is compared with the P-value which must be < 0.05 to be a significance; which is achieved in the suggested model with misclassification error equals 0.2903.

Table 3. Logistic regression whole model testing

<table>
<thead>
<tr>
<th>Model</th>
<th>-LogLikeood</th>
<th>DF</th>
<th>ChiSqaure</th>
<th>Prob&gt;ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Difference</td>
<td>6.923805</td>
<td>3</td>
<td>13.84761</td>
<td>0.0031</td>
</tr>
<tr>
<td>Full</td>
<td>33.400415</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reduced</td>
<td>40.324220</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In addition, all the PCs in the model have been tested under chi-square Wald test and found significant based on each P-value.

The unit odds ratio quantifies the unit change in each predictor that accomplishes the positive tissue type in the response comparing with the negative tissue type. In our case, the unit odds ratio for PC 2, PC 4, and PC 17 are 0.85215 , 1.95883 , and 0.571734 respectively, and it indicates that; the odds of getting a cancer tissue rather than normal tissue improves by factors of 0.85215, 1.95883, and 0.571734 for each increase of one unit of PC 2, PC 4, and PC 17 respectively.

The accuracy of the proposed model has sensitivity when the cancer type is positive and the probability of getting the cancer type also positive by the model equals 73%. In contrast, the model specificity when the cancer type is negative and the probability of getting the cancer type also negative by the model equals 67.3 %.

The important step is to identify and distinguish what genes have significantly associated for cancer tissues, thus it is needed to discover the PCs that is incorporated in this model. Consequently, the loading factor is required explore the correlation between the 200 genes and PCs 2, 14, and 17. That can be done by the strong correlation between all of these PCs and the genes in our data set, thus the estimated genes that have high association are as follows: 5,8,15,31,46-49,67,91,97,139,186, and 199.

5. Conclusion

In this research the kernel density estimators are applied to the whole 2000 genes for the seeking of smoothing the data, then carried all of density estimated with optimal bandwidth on the PPCA to reduce the dimension and decide about the number of the PPCs that should be retained based on the CICOMP, after that this approach is compared with traditional approach of computing PCs and showed that how this approach do not depend on the first some PPCs but it substantially cares about some other PPCs that may carry many effects on the loading factors. Secondly, the logistic regression, as classifier, is studied with Logit link function for the first 200 genes based on PCs as dimension reduction and carry out the PCs scores that represent 97% of total variance explained, on the subset selection through stepwise forward approach with the response of the tissue types. The model is run and investigated the significance of the estimated model based on different test methods. Finally, the most significant and effective genes are discovered and proposed for cancer tissue type based on the proposed logistic regression model. As an extended future research, the Bozdogan's ICOMP can be used as fitness function on Genetic Algorithm GA to get the best subset for building the logistic regression model based on PPCs as predictors. This research is heuristic steps that is suggested for considering such a problem of interest or related. Moreover, it can be applied for multinomial logistic regression and liner regression model as well. As an extension
of this research, it is suggested to consider the kernelized PPCA with CICOMP through discriminate analysis with different discriminate methods as new classifier scope and comparing that with logistic regression.

References

Biography
Abdulaziz Saud Alkabaa is an assistant professor of industrial engineering in the Department of Industrial Engineering at King Abdulaziz University, Jeddah. He received his PhD in ISE, and his second Master’s degree in Statistics from the University of Tennessee, and achieved several academic recognitions during his graduate studies. Dr. Alkabaa research of interests are stochastic process, healthcare engineering, industrial statistics, optimal design of industrial experiments, forecasting models, high dimensional data mining, and optimal classification.