2008

Peridocity, Change Detection and Prediction in Microarrays

Mohammad Shahidul Islam
Western University, aim@stats.uwo.ca

PERIDOCITY, CHANGE DETECTION AND PREDICTION IN MICROARRAYS

(Spine title: Peridocity, Change Detection and Prediction in Microarrays)

(Thesis format: Integrated-Article)

by

Mohammad Shahidul Islam

Graduate Program
in
Statistics

A thesis submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

Faculty of Graduate Studies
The University of Western Ontario
London, Ontario, Canada

ABSTRACT

Three topics in the analysis of microarray genomic data are discussed and improved statistical methods are developed in each case.

A statistical test with higher power is developed for detecting periodicity in microarray time series data. Periodicity in short series, with non-Fourier frequencies, is detected through a Pearson curve calibrated to the null distribution obtained by computer simulation. Unlike other traditional methods, this approach is applicable even in the presence of missing values or unequal time intervals. The usefulness of the new method is demonstrated on simulated series as well as actual microarray time series.

The second topic develops a new method for detection of changes in DNA or gene copy number. Regions for DNA copy number aberrations in chromosomal material are detected using maximum overlapping discrete wavelet transform (MODWT). It is shown how repeated application of MODWT to a series can be used to confirm the presence of change points. Application to simulated as well as array CGH (*Comparative Genomic Hybridization*) data confirms the excellent performance of this method.

In the third topic, it is shown that an improved class predictor for tissue samples in microarray experiments is developed by incorporating *nearest neighbour covariates* (NNC). It is demonstrated that this method reduces the mis-classification errors in both simulated and actual microarray data.

KEY WORDS: Beowulf cluster computing with R, change points in array CGH DNA copy number, class prediction in microarray datasets, detection of periodicity in microarray time series experiments, nearest neighbour covariates, wavelet change-point detection.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 0

INTRODUCTION

Microarray experiment is a promising technology to monitor the expression levels for thousands of genes simultaneously. This technology is relevant to almost all fields of life sciences. Microarrays provide a more complete understanding of the molecular variations among tumors, and hence direct to better diagnosis and treatment strategies for many diseases. DNA microarray experiments, followed in defined time period, are highly suitable to gene expression levels during a biological process. Apart from monitoring transcript or messenger ribonuceic acid (mRNA) levels, DNA microarrays are used to detect single nucleotide changes, unbalanced chromosome aberrations by *Comparative Genomic Hybridization* (CGH) experiment (Nuber, 2005).

The analysis of microarrays demands solving a number of statistical problems ranging from normalization to different supervised and unsupervised studies. Growth and development of any organism requires appropriate regulation of cell division cycle (Whitefield *et al.*, 2002). In cancer cell, the molecular processes for duplication of cell are erratic. So, advent of treatment for cancer or some other diseases might get possible through proper understanding of cell division cycle. There are well established theory and application to test for periodicity in short time series but with Fourier frequencies. However, most of the microarray time series are short and there is no guarantee that the series will only have Fourier frequencies. Wichert *et al.* (2004) discussed the issue of investigating periodicity in the microarray cell cycle data using Fisher's g statistic. Our proposed method can lead to substantial improvement

in power of the test when non-Fourier frequencies are present in the series. Moreover, other traditional methods fail to operate in presence of missing values in the series. But the proposed method is not affected by the missing values or unequal time interval.

Due to the presence of large number of genes for each single array, the issue of multiple testing in a genome-wide data analysis plays a great rule in reaching the final conclusion. A significant p-value obtained from a given setting for a specific gene would very unlikely refer to randomness rather than true features of this gene. But the presence of large number of genes makes it possible to get false positive and false negatives for a defined hypothesis. Wichert *et al.* (2004) used a method of False Discovery Rate (FDR), first proposed by Benjamini and Hochberg (1995), as multiple testing procedure. False positive rate, which leads to p-value, differ conceptually from FDR. Storey and Tibshirani (2003) suggested working with positive FDR (pFDR) for multiple testing. Pounds *et al.* (2004) proposed a method, called the spacing LOESS histogram (SPLOSH) for estimating the conditional FDR (cFDR) and claimed that this approach is more stable than the $q$value. Simulation results and implementation to real data show the variation of selecting the number of periodically expressed genes through different multiple testing methods. SPLOSH revealed to be most conservative while $q$value approach seems to be liberal in detecting correct number of periodic genes.

Copy number changes, also called chromosome gains or losses in the DNA content, often cause to tumorigenesis. Array CGH is a molecular-cytogenetic method that provides a way to do genomewide screening for such loss and gain regions referring to genetic alterations. To study and solve the challenge of efficiently identifying the regions with DNA copy number alterations, a number of methods have already been proposed. Pollack *et al.* (2002) applied a moving average to the process of ratios, and

use normal versus normal hybridization to compute the threshold. Maximum likelihood approach to fit mixture models corresponding to gain, loss and normal regions was used by Hodgson *et al.* (2001). An algorithm, proposed by Wang *et al.* (2005), builds hierarchical clustering-style trees along each chromosome, and then selects the clusters by controlling the FDR at a specific level. Wang (1995) develops a method for identifying the jumps in a time series by comparing wavelet coefficients of the data with a proposed threshold. In chapter 2, we propose a method using maximum overlapping discrete wavelet transform (MODWT) to detect the amplification or deletion points of DNA copy number. The region is defined to be gain or loss region using parametric bootstrap procedure.

A successful diagnosis and treatment of cancer depend on the classification of tumors through high-throughput microarray data analysis and this is one of the mostly studied issues in microarray experiment. Golub *et al.* (1999) worked with qualitative disease phenotypes. Comparison of different classification methods was provided by Dudoit *et al.* (2002) and Simon *et al.* (2004). Traditional k-nearest neighbour method selects single nearest neighbour for the purpose of predicting future observations. In Chapter 3, we consider plausibility of taking first nearest neighbour covariates in the set of original inputs. The performance of four methods in four miroarray and one simulated data set was investigated. We found that this type of augmented covariate set can result in better prediction.

REFERENCES

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, **57**, 289-300.

Dudoit, S., Fridlyand, J. and Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Stat. Assoc.*, **97**, 77-87.

Fort, G. and Lambert-Lacroix, S. (2005). Classification using partial least squares with penalized logistic regression. *Bioinformatics*, **21** , 1104-1111.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. and Lander, E. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286 (5439)**, 531-537.

Hodgson, G., Hager, J., Volik, S., Hariono, S., Wernick, M., Moore, D., Nowak, N., Albertson, D., Pinkel, D., Collins, C., Hanahan, D. and Gray, J.W. (2001). Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nature Genetics* **29**, 491.

Holmes, C. C. and Adams N. M. (2003). Likelihood inference in nearest-neighbour classification models. *Biometrika*, **90**, 1, 99-112.

Nuber U.A.(2005) *DNA Microarrays*. Taylor & Francis, New York.

Pollack J.R., Sorlie, T., Perou, C.M., Rees, C.A., Jeffry, S.S., Lonning, P.E., Tibshirani, R., Botstein D., Borrsen-Dale, A. and Brown, P.O. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences*, USA, **99**, 1296312968

Pounds,S. and Cheng C. (2004) Improving false discovery rate estimation, *Bioinformatics*, **20**, 1737-1745.

Simon, R. M., Korn, E. L., McShane, L. M., Radmacher, M. D., Wright, G. W., Zhao, Y. (2004). *Design and Analysis of DNA Microarray Investigations*. New York: Springer.

Storey,J.D. and Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci., USA*, **100**, 9440-9445.

Wang, Y. (1995). Jump and Sharp Cusp Detection by Wavelets. *Biometrika*, **82**, 385-397.

Wang, P., Kim, Y., Pollack, J., Narasimhan, B., Tibshirani, R. (2005). Studies in crop variation. I. A method for calling gains and losses in array CGH data. *Biostatistics*, **6**, 1, 4558

Whitefield,M.L., Sherloc,G., Saldanha,A.J., Murrary,J.I., Ball,C.A., Alexander,K.E., Matese,J.C., Perou,C.M., Hurt,M.M., Brown,P.O. and Botstein,D. (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977-2000.

Wichert,S., Fokianos K. and Strimmer K. (2004) Identifying periodically expressed transcrips in microarray time series data. *Bioinformatics*, **18**, 5-20.

Chapter 1

TESTING PERIODICITY AND APPLICATION TO GENE EXPRESSION DATA

1.1 INTRODUCTION

Searching and studying the behavior of periodically expressed genes in time series gene expression data have drawn recent interest in microarray technology. There are well established theory and application to test for periodicity in time series with small sequence size and Fourier frequencies. However, most of the microarray time series are short and there is no guarantee that the series will only have Fourier frequencies. Wichert *et al.* (2004) discussed the issue of investigating periodicity in the microarray cell cycle data. They introduced a graphical approach, called *average periodogram* (AP), for a quick view for searching periodicity in the data set. Our investigations suggest that this method does not perform well if the data set contains unequal frequencies for different series. For some of the cell cycle data sets, it was found that the AP gives misleading results. Application of Fisher's g statistic for detecting the periodicity in gene expression data fails to perform correctly even in some explicit periodic series. This is, in fact, mostly due to the non-Fourier frequencies in the series. Specifically, Fisher's g test has very low power for testing periodicity in short series with non-Fourier frequencies. In Section 1.4 we show that $n \geq 50$ is needed.

In our procedure we use log-likelihood ratio (LLR) for any kind of frequencies. We simulate data from completely white noise process and find the LLR for the series.

The simulation is done very large number of times and a Pearson VI curve is fitted to the calculated -2LLR. Thereafter the p-value for the required test can be obtained from the CDF of the curve.

In testing periodicity, missing time points are either imputed by interpolation or the genes with missing values are removed from the analysis. Our proposed method is completely simulation based and so there is no need to omit any gene with missing observations. We just simulate the observations according to the time points of our working data set.

Permutation test is an exact test, introduced by Fisher and Pitman in the 1930's. Initially it represented a theoretical standard rather than a practical approach. But with the improvement of computer speed, the permutation test was applied to a wider and wider variety of problems (Good, 2000). We can obtain as small as 0.02% percentage of variance for a desired p-value of 0.001 when 5000 random permutations are taken. The permutation test cannot generate very low p-values. In the multiple testing situation we face in time series microarray experiments, this implies that it will have lower power than the method we have developed.

Large number of statistical hypothesis tests conducted in the microarray data analysis can potentially lead to a large number of false discoveries, which is significant findings that arise solely by chance mechanisms. Therefore, reaching a correct decision in such case requires the use of an efficient multiple test method. A short simulation result shows the variation of selecting the number of periodically expressed genes using different multiple testing methods. Pearson Curve fitting method was applied to various microarray data sets for periodicity detection; then Benjamini and Hochberg's FDR, Storey's q-value approach and Pounds & Cheng's SPLOSH were used for detecting the number of genes of interest. The results depend highly on the choice of multiple test method. Other issues might affect the analysis e.g. non-

randomness or autoregressive-behavior of the error process in the sinusoidal model. After all, the result of the whole analysis should be verified by biological interpretation for final conclusion.

The whole procedure was implemented in software `R`. Moreover, `GeneTS` and `qvalue` packages from CRAN were used for Fisher's g test and Storey's q-value approach. The method for SPLOSH is available online (Pounds *et al.*, 2004).

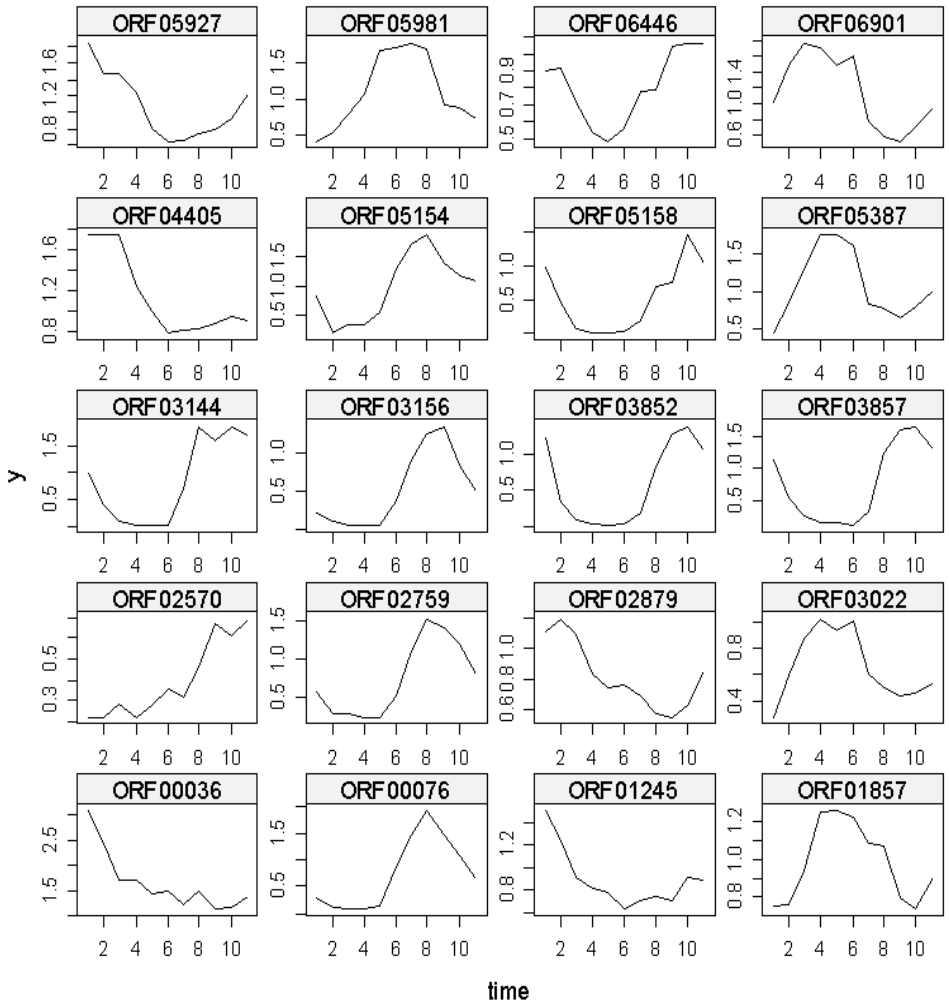

Figure 1.1: First 24 most significantly periodic ORFs in *Caulobacter crescentus* cell cycle data. The periodicity was detected using Pearson curve fitting method. Here the periodic pattern in all series is not same. The genes are indicated in the top level.

1.2 METHODS

In its simplest form we assume that the time series, $z_t, \quad t = 1, \ldots, n$ consists simply of a sinusoid plus random error.

$$z_t = \mu + \alpha \cos(2\ \pi\lambda t + \beta) + e_t \tag{1.1}$$

where e_t is the error or disturbance and α, β and λ are the parameters. It is frequently assumed that e_t is IID with mean zero and variance σ^2. This lack of autocorrelation is probably not suitable. It is worthwhile to explore a generalization of this assumption so that e_t is assumed to be generated by a stationary time series model such as an AR(1).

The model in equation 1.1 can also be written in the form,

$$z_t = \mu + A \cos(2\ \pi\lambda\ t) + B \sin(2\ \pi\ \lambda\ t) + e_t \tag{1.2}$$

where A and B jointly determine the amplitude and phase of the sinusoid and λ is the frequency. The periodogram is usually computed at the Fourier frequencies, $\lambda_j = j/n$ where $j = 1, ..., [n/2]$.

$$I(\lambda_j) = \frac{1}{n} | \sum_{t=1}^{n} z_t e^{-2\pi i \lambda_j t} |^2 \tag{1.3}$$

Fisher's g statistic can be written as,

$$g = \frac{\max_j I(\lambda_j)}{\sum_{j=1}^{m} I(\lambda_j)} \tag{1.4}$$

where m is (n-1)/2 or (n-2)/2 according to n is odd or even.

Then under the assumption that e_t is NID $(0, \sigma^2)$ and the null hypothesis H_0 : $\lambda = 0$ the CDF of g is given by,

$$F(x) = \Pr(g \leq x) = 1 - \sum_{i=1}^{\rho} \binom{n}{i} (-1)^{(i-1)} (1 - ix)^{(n-1)} \tag{1.5}$$

where $\rho = \lfloor 1/x \rfloor$. An upper tail test is used so the observed p-value is given by $1 - F(g)$.

Note that although the null hypothesis is $H_0 : \lambda = 0$, Fisher's g test will also detect any departure from an uncorrelated sequence provided the sample is large enough. The test will be optimal against an alternative hypothesis $H_1 : \lambda = \lambda_j$ where $\lambda_j \in 1/n,\ 2/n, ...,\ [n/2]/n$.

It should be noted here that, when the data are independent random numbers drawn from a Gaussian distribution, the periodogram ordinates at the Fourier frequencies are independently exponentially distributed. Walker (1965) searched for the maximum likelihood of the frequency over the range 0 to π; and thus it was not restricted to only the Fourier frequency. Turkman and Walker (1984) gave the asymptotic distribution of

$$G_T^{'} = \max_{f \in (0,\pi)} I^x(f)/\mu \tag{1.6}$$

where $f = 2\pi\lambda$ and μ is the mean of the distribution of the periodogram ordinates. It was shown that

$$P[G_T^{'} \leq z + \log n\ + \log(\log n)/2 - \log(3\pi)/2] = \exp[-\exp(-z)] + o(1) \tag{1.7}$$

Chiu (1989) proposed a modified statistic which is proportional to the ratio of the maximum periodogram to a trimmed mean of the periodogram. It was shown that the method has same asymptotic power as that of Fisher's test. Let $\hat{\mu} = n^{-1} \sum_{j=1}^{m} I(\lambda_j)$ and $\tilde{I}(\lambda_j) = I(\lambda_j)/\hat{\mu}$ be the normalized periodogram. It was mentioned that Fisher's test statistic is proportional to the maximum of the periodogram ordinates normalized

by the sample mean of the periodogram ordinates. When the series contains periodic components, the periodogram ordinates at frequencies close to the frequencies of these components have large magnitude and these periodogram ordinates can be viewed as outliers in an exponential sample. The sample mean is affected by outliers, but Fisher's test uses the sample mean to normalize the periodogram. The sample mean of the periodogram ordinates tends to be larger than μ when the series contains periodic components; therefore, $\tilde{I}(\lambda_j)$ tends to be smaller than $I(\lambda_j)/\mu$. So the power of Fisher's test will be smaller than the power of the test based on the maximum of $I(\lambda_j)/\mu$. This issue leads to another test that uses the maximum of periodogram ordinates normalized by a robust estimate of μ. Chiu (1989) proposed the test statistic to be

$$R_T(\beta) = I_n / \sum_{j=1}^{n\beta} I_j \tag{1.8}$$

where $I_1 < I_2 < \ldots < I_n$ are the order statistic of the periodogram ordinates $I(\lambda_j)$ and β is a constant between zero and unity which determines the proportion of periodogram ordinates trimmed. Usually $\beta = 0.9$ or $\beta = 0.95$ will give good result. The performance of the test is not very sensitive to the choice of β. A value of $\beta = 1$ refers to Fisher's test as an extreme case.

In the case of non-Fourier frequency, Chiu (1989) extended the method in Equation (1.6) to give a new form of statistic. Instead of using all ordinates of the periodogram, the statistic takes trimmed mean by considering only lower $100\beta\%$ $I(f')$'s in account. The statistic is defined as

$$G_T = \max_{f \in (0,\pi)} cI^x(f)/\mu_\beta \tag{1.9}$$

where $c = 1 + (1-\beta)\log(1-\beta)/\beta$.

1.2.1 PEARSON SYSTEM OF EQUATION

In statistical literature some families of distributions have been constructed to provide approximations to a variety of observed distributions. Such families are often called system of distributions and elaborate discussion was given by Johnson and Kotz (1970). Karl Pearson proposed a family of systems where every member has a probability density function $p(x)$ which satisfies a differential equation of the form

$$\frac{1}{p}\frac{dp}{\mathrm{dx}} = -\frac{a+x}{c_0 + c_1 x + c_2 x^2} \tag{1.10}$$

Type IV occurs when $c_0 + c_1 x + c_2 x^2$ does not have real roots. Type VII is the special symmetrical case of Type IV, and it occurs when $c_1 = a = 0$. This nests Student's t distribution. Type III (Gamma distribution) occurs when $c_2 = 0$. Type V arises when the quadratic $c_0 + c_1 x + c_2 x^2 = 0$ has one real root. In this case $c_1^2 - 4c_0 c_2 = 0$. The Normal distribution is obtained when $c_1 = c_2 = 0$.

That leaves Type I, Type II and Type VI. These cases occur if $c_0 + c_1 x + c_2 x^2 = 0$ has two real roots, r_1 and r_2. In particular, Type I occurs if $r_1 < 0 < r_2$; that is, roots are of opposite sign with domain $r_1 < x < r_2$. This nests the Beta distribution. Type II is identical to Type I, except that here we further assume that $r_1 = -r_2$. This yields a symmetrical curve with $\beta_1 = 0$. Type VI occurs if r_1 and r_2 are the same sign; the domain is $x > r_2$ if $0 < r_1 < r_2$, or $x < r_2$ if $r_2 < r_1 < 0$. In the case of Type VI, with two real roots of the same sign, one can express $c_0 + c_1 x + c_2 x^2$ as $c2(x - r_1)(x - r_2)$. The family of solutions is then:

$$P(x) = K(x - r_1)^{\frac{a+r_1}{-c_2 r_1 + c_2 r_2}} (x - r_2)^{\frac{a+r_2}{c_2 r_1 + c_2 r_2}} \tag{1.11}$$

where K is a constant of integration which can now be solved for the relevant domain. The shape of the resulting distribution will clearly depend on the Pearson parameters

(a, c_0, c_1, c_2). These parameters can be expressed in terms of the first four moments of the distribution. Thus, if we know the first four moments, we can construct a density function that is consistent with those moments. This provides a nice way of constructing density functions that approximate a given set of data. Karl Pearson grouped the family into a number of types. These types can be classified in terms of (β_1,β_2) space, where

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} \text{ and } \beta_2 = \frac{\mu_4}{\mu_2^2}$$

We can consider some higher order polynomial. For example, taking $c_0 + c_1x + c_2x^2 + c_3x^3$ in the denominator of the differential equation will provide flexible fit given that the population moments are known. However, if the sample size is large enough then using sample moments will still provide reliable fit. In terms of moments, the solutions of the coefficients will be

$$\begin{aligned}
c_0 &= \frac{\acute{\mu}_1\acute{\mu}_3(\acute{\mu}_2{}^2 + \acute{\mu}_4) + \acute{\mu}_2(3\acute{\mu}_3{}^2 - 4\acute{\mu}_2\acute{\mu}_4) + \acute{\mu}_1{}^2(-4\acute{\mu}_3{}^2 + 3\acute{\mu}_2\acute{\mu}_4)}{2(9\acute{\mu}_2{}^3 + 4\acute{\mu}_1\acute{\mu}_3 - 16\acute{\mu}_1\acute{\mu}_2\acute{\mu}_3) + 6\acute{\mu}_3{}^2 - 5\acute{\mu}_2\acute{\mu}_4 + \acute{\mu}_1{}^2(-3\acute{\mu}_2{}^2 + 5\acute{\mu}_4)} \\
a &= \frac{20\acute{\mu}_1{}^2\acute{\mu}_2\acute{\mu}_3 - 12\acute{\mu}_1{}^3\acute{\mu}_4 - \acute{\mu}_3(3\acute{\mu}_2{}^2 + \acute{\mu}_4) + \acute{\mu}_1(-9\acute{\mu}_2{}^3 - 8\acute{\mu}_3{}^2 + 13\acute{\mu}_2\acute{\mu}_4)}{2(9\acute{\mu}_2{}^3 + 4\acute{\mu}_1\acute{\mu}_3 - 16\acute{\mu}_1\acute{\mu}_2\acute{\mu}_3) + 6\acute{\mu}_3{}^2 - 5\acute{\mu}_2\acute{\mu}_4 + \acute{\mu}_1{}^2(-3\acute{\mu}_2{}^2 + 5\acute{\mu}_4)} \\
c_1 &= \frac{8\acute{\mu}_1{}^2\acute{\mu}_2\acute{\mu}_3 - 6\acute{\mu}_1{}^3\acute{\mu}_4 - \acute{\mu}_3(3\acute{\mu}_2{}^2 + \acute{\mu}_4) + \acute{\mu}_1(-3\acute{\mu}_2{}^3 - 2\acute{\mu}_3{}^2 + 7\acute{\mu}_2\acute{\mu}_4)}{2(9\acute{\mu}_2{}^3 + 4\acute{\mu}_1\acute{\mu}_3 - 16\acute{\mu}_1\acute{\mu}_2\acute{\mu}_3) + 6\acute{\mu}_3{}^2 - 5\acute{\mu}_2\acute{\mu}_4 + \acute{\mu}_1{}^2(-3\acute{\mu}_2{}^2 + 5\acute{\mu}_4)} \\
c_2 &= \frac{6\acute{\mu}^3 + 4\acute{\mu}_1\acute{\mu}_3 - 10\acute{\mu}_1\acute{\mu}_2\acute{\mu}_3 + 3\acute{\mu}_3{}^2 - 2\acute{\mu}_2\acute{\mu}_4 + \acute{\mu}_1{}^2(-3\acute{\mu}_2{}^2 + 2\acute{\mu}_4)}{2(9\acute{\mu}_2{}^3 + 4\acute{\mu}_1\acute{\mu}_3 - 16\acute{\mu}_1\acute{\mu}_2\acute{\mu}_3) + 6\acute{\mu}_3{}^2 - 5\acute{\mu}_2\acute{\mu}_4 + \acute{\mu}_1{}^2(-3\acute{\mu}_2{}^2 + 5\acute{\mu}_4)}
\end{aligned}$$

We can use the relative likelihood function to compare two statistical models, say M_1 and M_2. Let Ł(M_1) and Ł(M_2) denote the likelihood for models M_1 and M_2. Then the relative plausibility of model M_1 vs. M_2 is defined as $R = \text{Ł}(M_1)/\text{Ł}(M_2)$. Our aim is to test

$$H_0 : \alpha = 0 \qquad \text{against} \qquad H_0 : \alpha > 0$$

The likelihood ratio after dropping the constant is then given as $R = (S_2/S_1)^{n/2}$, where S_1 and S_2 are given as

$$S_1 = \sum_{t=1}^{n}(z_t - \bar{z})^2 \text{ and } S_2 = \sum_{t=1}^{n}(z_t - \hat{A}\cos(2\pi\hat{\lambda}t) - \hat{B}\sin(2\pi\hat{\lambda}t))^2$$

We partition the whole range of frequency in discrete parts; that is, for every series of length 101 or less we take 50 frequency values ranging from 1/101 to 50/101. Thus we can find the sum squares of residuals (S) in the regression model for each value of λ and thereafter pick the value which minimizes the value of S. In the setting, log-likelihood ratio can depict the presence of sinusoid in a given series. However, non-identifiability of the distribution in the case when $\lambda = 0$ makes the maximum likelihood estimates non-normal, and so -2LLR does not follow χ^2 distribution. $100,000$ simulated series of required size are obtained from white noise process and for each series -2LLR is calculated. The resultant values follow Pearson Type VI distribution. Therefore, we can find CDF and hence the p-value from this fitted curve. As mentioned before we can fit cubic Pearson-style distribution to the values and henceforth find the p-values. It was explored that the difference between these two curve fitting is negligible. Another curve fitting approach, with Pearson type VI distribution up to 99% quantile and Pareto distribution thereafter at the tail, can also be used.

None of the traditional method is able cope with testing periodicity if there exist unequal time interval or missing observations in the series. In this simulation method, we simulate the observations exactly according to the time points where the original series appear. For example, if the original series takes the values at time points 1, 3, 4, 5, 7, 8, 9, 10, 11, 13, then we fit the Pearson type VI curve given as

$$f(x) = \frac{1.06080*10^{105}(-4.68199+1x)^{1.24837}}{(111.31353+1x)^{51.30084}}, \quad 4.68199 \leq x \leq \infty$$

after finding LLR from such simulated series. This allows us not to omit any series with missing observations from our analysis.

1.3 MULTIPLE TESTING

In microarray experiments we have thousands of short time series and these are tested against some null hypothesis. Multiple testing procedures attempt to adjust p-values derived from multiple statistical tests to correct for occurrence of false positives. The aim is to estimate a measure of significance that is easily interpreted in terms of the simultaneous testing of thousands of genes. Therefore, the final conclusion in gene selection is based on multiple testing.

Suppose we have expression levels for G genes measured in time sequence. Using some specific method of test of periodicity, let the p-values be $p_1, p_2, \ldots, p_G$. Wichert *et al.* (2004) used False Discovery Rate (FDR), first proposed by Benjamini and Hochberg (1995), as multiple testing procedure. There is big conceptual difference between false positive rate and FDR. When features are said to be significant, the false positive rate is the rate that truly null features are called significant. However, the FDR is the rate that significant features are truly null. There are different other multiple testing procedures present in the literature. However, we discuss two most recent but appealing methods.

1.3.1 False Discovery Rate and q-value

p-value is a measure of significance in terms of false positive rate. Although the idea of q-value is similar to that of p-value, the former is an extension of a quantity called false discovery rate (Storey, 2002). Storey and Tibshirani (2003) claimed that the method offers a sensible balance between the number of true and false positives that is automatically calibrated. If the features with q-value less than α are called significant, this means that among the significant features, a value of FDR will be $\alpha\%$. The features with p-value less than α provides false positive rate of $\alpha\%$ among

all null features. Suppose we have total number of M features in the genomewide study and consider table 1.1 for the development of q-value approach.

Possible outcomes in tests of hypothesis

	Significant	Not significant	Total
Null True	F	M_0 -F	M_0
Alternative True	T	M_1 -T	M_1
Total	S	M-S	M

Table 1.1: Different possible outcomes in hypothesis testing when we consider total of M features in the genome-wide study. S is the number of outcomes we call significant. Here M_0 and M_1 represent the number of features under null and alternative hypotheses respectively.

Under notations of table 1.1, the expected value of false positive $E(F) \leq 0.05M$ is guaranteed by a p-value threshold of 0.05. In a genomewide study, the value of M will be large and hence the value of $E(F)$; this leads the false positive rate to be very liberal. Again, controlling the family wise error rate $\Pr(F \geq 1)$ would be too conservative as we generally have a number of significant genes in a microarray data analysis. A balance between these two situation can be done by considering FDR which is the expected value of the ratio between number of false positives and total number of significant genes.

$$\text{FDR} = E(\frac{F}{S}) \tag{1.12}$$

Now the FDR is to be estimated such that a feature is called significant when the p-value is less than or equal to a threshold t ($0 < t < 1$). Let $p_1, p_2, ..., p_M$ be the p-values and

$$\begin{aligned} F(t) &= \#\{\text{null } p_i \leq t; i = 1, 2, ..., M\} \\ S(t) &= \#\{p_i \leq t; i = 1, 2, ..., M\} \end{aligned}$$

Since the number of features is very large, we can write,

$$\text{FDR}(t) = E(\frac{F(t)}{S(t)}) \approx \frac{E(F(t))}{E(S(t))} \tag{1.13}$$

A simple estimate of $E(S(t))$ is the observed $S(t)$. The null p-values are uniformly distributed over the range of 0 and 1 and so $\Pr(\text{null } p \leq t) = t$. We can write $E(F(t)) = M_0 t$; but M_0 is unknown and has to be estimated. This is equivalent to estimating $\pi_0 = M_0/M$. As the null p-values are uniformly distributed, an estimate of π_0 with respect to tuning parameter λ, can be expressed in the following form:

$$\hat{\pi}_0(\lambda) = \frac{\#\{p_i > \lambda; i = 1, 2, ..., M\}}{M(1-\lambda)} \tag{1.14}$$

There is a tradeoff between bias and variance in choosing the value of λ. If λ increases, the bias of $\hat{\pi}_0(\lambda)$ decreases and the bias is the minimum when $\lambda \to 1$. Thus we need to estimate the quantity $\lim_{\lambda \to 1} \hat{\pi}_0(\lambda) \equiv \hat{\pi}_0(\lambda = 1)$. In order to do so, the value of $\hat{\pi_0(\lambda)}$ is plotted for a sequence of values of λ ranging from 0.001 to 0.95. A cubic spline is fitted to these data points, and then evaluated at $\lambda = 1$. This fitted value is the final estimate for π_0.

The estimated value of FDR can be quantified as

$$\widehat{\text{FDR}}(t) = \frac{\hat{\pi}_0 mt}{S(t)} = \frac{\hat{\pi}_0 mt}{\#\{p_i \leq t\}} \tag{1.15}$$

Now, the q value of feature i can be estimated by using the estimated FDR from above equation

$$\hat{q}(p_i) = \min_{t \geq p_i} \widehat{\text{FDR}}(t) \tag{1.16}$$

Storey and Tibshirani (2003) used positive FDR (pFDR) as an alternative quantity to FDR. This is defined as, $\text{pFDR} = E(F/S|S > 0)$. Most technically, a q value

is defined as the minimum pFDR at which the feature is called significant. However, $\Pr(S > 0)$ approximates 1 due to large number of features in a study, and so clearly pFDR $\approx E(F)/E(S)$. This quantity is approximately equal to FDR; thus the distinction between these two was not considered to be crucial.

Providing an example, Storey and Tibshirani (2003) mentioned that Benjamini's approach of using FDR is too conservative and thus has smaller power as it assumes $\pi_0 = 1$. Another restrictive and impractical behavior of Benjamini's approach is that a single acceptable level of FDR has to be chosen beforehand.

Storey and Tibshirani (2003) suggested an automated algorithm to calculate q-value:

1. Let $p_{(1)} \leq P_{(2)} \leq ... \leq p_{(m)}$ be the ordered p-values.

2. For a large range of λ, say $\lambda = 0.001, 0.002, ..., 0.95$, calculate

$$\hat{\pi_0(\lambda)} = \frac{\#\{p_i > \lambda; i=1,2,...,m\}}{m(1-\lambda)}$$

3. Let $\hat{f}$ be the natural cubic spline with 3 df of $\hat{\pi}_0(\lambda)$ on λ.

4. Set the estimate of π_0 to be $\hat{\pi}_0 = \hat{f}(1)$

5. Calculate

$$\hat{q}(p_{(m)}) = \min_{t \geq p_{(m)}} \frac{\hat{\pi}_0 mt}{\#\{p_j \leq t\}} = \hat{\pi}_0 p_{(m)}$$

6. For $i = m-1, m-2, \ldots, 1$ calculate

$$\hat{q}(p_{(m)}) = \min_{t \geq p_{(m)}} \frac{\hat{\pi}_0 mt}{\#\{p_j \leq t\}} = \hat{\pi}_0 p_{(m)} = \min(\frac{\hat{\pi}_0 mt}{i}, \hat{q}(p_{(i+1)})$$

7. The estimated q-value for the ith most significant feature is $\hat{q}(p_{(i)})$

1.3.2 SPLOSH

This method, called spacings LOESS, was recently proposed by Pounds *et al.* (2004). It works through estimating conditional FDR (cFDR), the expected proportion of false positives given we have r significant genes. Pounds *et al.* (2004) claimed that their method is more stable than Storey's q-value approach. cFDR is defined as

$$\text{cFDR} = E(\frac{V}{R}|R = r) = \frac{E(v|R = r)}{r} \tag{1.17}$$

Using the definition of FDR, we can write

$$\text{cFDR} = \text{FDR} = \frac{\pi_0 t}{\Pr(p \leq t)} = \frac{\pi_0 t}{F(t)} \tag{1.18}$$

and so an estimate of this would be

$$\widehat{\text{cFDR}} = \frac{\hat{\pi}_0 t}{\hat{F}(t)} \tag{1.19}$$

We see in q-value approach that $\hat{F}(t)$ is considered to be the observed proportion of p-values less than t and π_0 is estimated using cubic spline. SPLOSH works with getting a smooth estimate for $F(t)$ in estimating cFDR. Now, $\hat{F}(t)$ can be considered as an estimate of p-value CDF, $F(t)$. A smooth estimate of $F(t)$ could be obtained by integrating an estimate of p-value PDF, $f(t)$. Pounds *et al.* (2004) described the whole procedure for getting smooth estimate of cFDR after giving the notations for various calculations. Let $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(g)}$ be the ordered p-values and $a_{(i)} = (i - 1/2)/g$ be their adjusted ranks. Assume that there are $\tilde{g}$ unique p-values written as $\tilde{p}_{(1)} \leq \tilde{p}_{(2)} \leq ... \leq \tilde{p}_{(g)}$. For $j = 1, 2, ..., \tilde{g}$, let $\tilde{a}_j$ be the average of $a_{(i)}$ for all i such that $p_{(i)} = \tilde{p}_{(j)}$. Define $\tilde{p}_{(0)} = 0$ and $\tilde{a}_{(0)} = 0$ if $\tilde{p}_{(1)} > 0$. Also define $\tilde{p}_{(\tilde{g}+1)} = 1$ and $\tilde{a}_{(\tilde{g}+1)} = 1$ if $\tilde{p}_{(\tilde{g})} < 1$. l and u respectively represent the lower and

upper indices j of the set $\tilde{p}_{(j)}$:

$$l = \begin{cases} 0, & \text{if } p_{(1)} > 0 \\ 1, & \text{otherwise} \end{cases}$$

and

$$u = \begin{cases} \tilde{g}, & \text{if } p_{(\tilde{g})} < 1 \\ \tilde{g} + 1, & \text{otherwise} \end{cases}$$

For $j = l, ..., u - 1$, define

$$m_j = \frac{\tilde{p}_{(j+1)} + \tilde{p}_{(j)}}{2} \tag{1.20}$$

$$\triangle_j = \tilde{p}_{(j+1)} - \tilde{p}_{(j)} \tag{1.21}$$

$$\partial_j = \frac{\tilde{a}_{(j+1)} - \tilde{a}_{(j)}}{\tilde{p}_{(j+1)} - \tilde{p}_{(j)}} \tag{1.22}$$

$$\tilde{x}_j = \arcsin[2 \times (m_j - 1/2)] \tag{1.23}$$

$$\tilde{y}_j = \log(\partial_j) \tag{1.24}$$

$$\tag{1.25}$$

For $i = 1, 2, ..., G$, define

$$x_i = \arcsin[2 \times (p_i - 1/2)] \tag{1.26}$$

Then the procedure to get estimate of cFDR using SPLOSH algorithm is described in the following steps:

1. Compute the quantities m_j and x_i described above
2. Apply LOESS to $(\tilde{x}_j, \tilde{y}_j)$ for $j = l, ..., u - 1$ to obtain as estimated curve $\hat{y}(\cdot)$.
3. For $j = l, ..., u$, let $\hat{f}^*(\tilde{p}_{(j)}) = \exp[\hat{y}(\tilde{x}_{(j)})]$ be an estimate of $f(\tilde{p}_{(j)})$ up to a unitizing constant c.

4. Let $\hat{f}(p_i) = 1/c\hat{f}^*(\tilde{p}_{(j)})$ estimate the PDF at p_i for $i = 1, 2, ..., G$, where

$$c = \frac{1}{2}\sum_{j=l}^{u-1}[\hat{f}^*(\tilde{p}_{(j)}) + \hat{f}^*(\tilde{p}_{(j+1)})]\triangle_j \tag{1.27}$$

is determined by trapezoid rule integration.

5. Let $\hat{F}(\tilde{p}_{(i)}) = 0$ and for $k = l + 1, ..., u$ let

$$\hat{F}(\tilde{p}_{(k)}) = \frac{1}{2}\sum_{j=l}^{k-1}[\hat{f}(\tilde{p}_{(j)}) + \hat{f}(\tilde{p}_{(j+1)})]\triangle_j \tag{1.28}$$

be an estimate of $\hat{F}(\tilde{p}_{(k)})$ obtained by trapezoid rule integration.

6. Let $\hat{\pi}_0 = \min_{1\leq i\leq g}\hat{f}(p_i)$

7. For $i = 1, 2, ..., g$, obtain $r_{(i)} \equiv \hat{r}(p_{(i)})$ by substituting the value of $p_{(i)}$, $\hat{\pi}_0$ and $\hat{F}(p_{(i)})$ in the definition of $\widehat{\text{cFDR}}$. Use $\hat{\pi}_0/\hat{f}(0)$ as an estimate of cFDR for p-values equal to 0.

8. Define

$$h_{(i)} = \min_{k\geq i}(r_{(k)}) \tag{1.29}$$

as a monotone quantity based on the cFDR estimates $r_{(i)}$, for $i = 1, 2, ..., g$.

This algorithm is called SPLOSH because it applies LOESS to the p-value spacings to obtain a PDF estimate.

In the procedure described above, the log-transformation of $\partial_{(i)}$ ensures that the PDF estimate will be strictly positive after back-transformation. Moreover, this transformation makes the distribution of $y_{(j)}$ more symmetric, which is important to get reasonable estimate of $\hat{y}(\cdot)$ after applying LOESS algorithm. For p-values at extreme

ends, overborrowing information from the center of the p-value distribution in estimating $\hat{y}(\cdot)$ using LOESS is prevented by the arc-sine transformation of $m_{(j)}$. At p-values close to 0 and 1, this overborrowing of information tends to give respectively downward and upward bias estimate of $f(p)$.

Simulation as well as a real example was presented to show that SPLOSH exhibits greater stability than Storey's q-value method for small p-values. To estimate $\hat{F}(t)$, SPLOSH uses the information on both sides of t and thus maintains the level of stability. However, q-value approach uses the information only on the left side of t, and thus makes the method unstable.

1.4 RESULTS

We investigate the performance of our proposed method for different series sizes. The following tables show the power comparison, at a pre-specified significance level 0.05, using two methods - (a) Fisher's g test and (b) Pearson type VI curve fitting. The results are based on simulation studies each from three types of series - (a) white noise process, (b) series with Fourier frequency and (c) series with non-Fourier frequency. The simulation results for first kind of series, that is, series with $\lambda = 0$ refer to type I errors.

1.4.1 Sample Size-10

We simulate series of length 10 and calculate log-likelihood ratio. Figure 1.2 indicates that the estimated moments refer to the space of Pearson type VI density. This claim is confirmed by looking at figure 1.3, which indicates that the estimated and exact density function merge. From the simulated values, we can obtain a fitted density of

the form:

$$f(x) = \frac{4.9961^{128}(-4.55905 + x)^{0.77219}}{(147.49666 + 1x)^{59.30667}}, 4.55905 \leq x \leq \infty \tag{1.30}$$

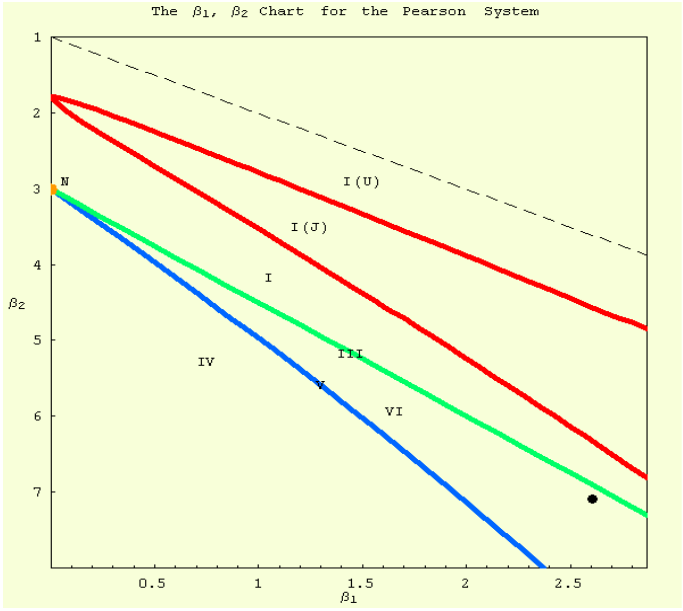

Figure 1.2: The distribution of simulated log-likelihood ratio for the test of periodicity in white noise series of length 10. The simulation was done for 10^5 times. It reveals that Pearson type VI is the most appropriate system of distribution in this case

We present the simulation result in Table 1.1. First and second elements in each column represent the values obtained from Fisher's g test and Pearson curve fitting method respectively. Series with Fourier frequency and small error variance (σ^2), both Fisher's g test and Pearson curve fitting give good power of the test. However, with the increase of error variance the former outperforms the latter.

A series with $\lambda = 0.15$ corresponds to the series with non-Fourier frequency. In such a case, Fisher's g test fails miserably to test periodicity in the series. As can be seen from Table 1.1, this test even fails 100% of times for the series with $\lambda = 0.15$ and very small value of σ. Pearson curve fitting still gives very large power of the test. There is almost no difference in the performance of this method for Fourier

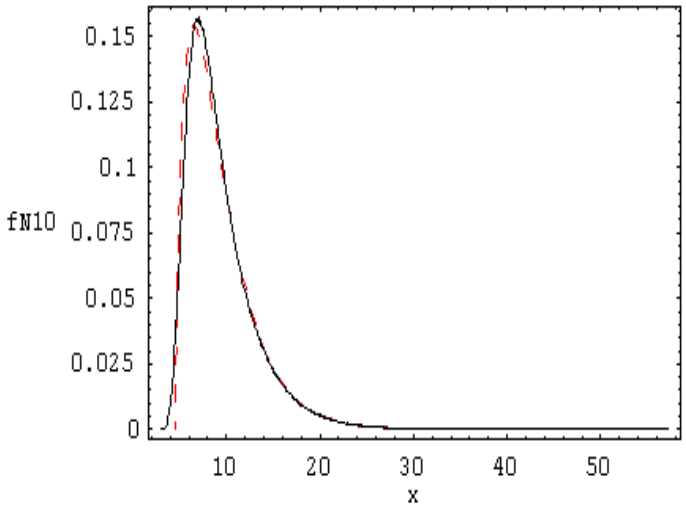

Figure 1.3: Comparison of fitted and theoretical Pearson type VI curve. Solid black curve represents the latter, while the dotted red curve represents the other. The fitted line was obtained from 10^5 simulated log-likelihood ratios when testing periodicity in white noise process of length $n = 10$.

	Power comparison; $n = 10$		
σ	$\lambda = 0.0$	$\lambda = 0.1$	$\lambda = 0.15$
0.1	0.0505|0.0506	1.0000|1.0000	0.0000|1.0000
0.2	0.0541|0.0517	1.0000|0.9998	0.0010|0.9997
0.3	0.0483|0.0523	0.9748|0.9455	0.0093|0.9281
0.4	0.0486|0.0501	0.8136|0.7300	0.0230|0.7043
0.5	0.0499|0.0530	0.5816|0.4950	0.0386|0.4822
0.6	0.0480|0.0503	0.4107|0.3332	0.0359|0.3205
0.7	0.0457|0.0501	0.2890|0.2348	0.0412|0.2248
0.8	0.0508|0.0508	0.2104|0.1712	0.0447|0.1693
0.9	0.0492|0.0501	0.1677|0.1355	0.0450|0.1326
1.0	0.0500|0.0489	0.1272|0.1117	0.0480|0.1089

Table 1.2: Power comparison of Fisher's test and Pearson curve fitting method for a series of length 10. The simulation was carried out 10^4 times. $\lambda = 0.1$ and $\lambda = 0.15$ correspond to Fourier and non-Fourier frequencies respectively. First element in each column represents the power for Fisher's g test and the second one represents that for our proposed method.

or non-Fourier frequencies. So the table reveals that the presence of even a perfect periodicity might not be detected by Fisher's g test when the sequence possesses non-Fourier frequencies. For example, Figure 1.4 is a plot for the series defined as

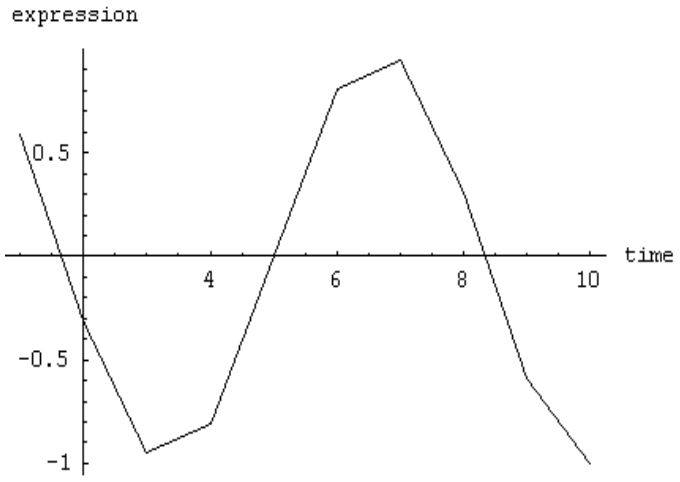

Figure 1.4: Representation of the series $\cos(2\pi\lambda t)$, where $\lambda = 0.15$ and series length is 10. Fisher's test is unable to detect the periodicity due to non-Fourier frequency in the series.

$\cos(2\pi(0.15)t)$ where series size is 10 and $\lambda = 0.15$, a non-Fourier frequency. Although this is a perfect sinusoidal series, Fisher's g test fails to detect the periodicity.

1.4.2 Sample Size-20

Estimated moments plotted in Pearson system diagram indicate that the distribution of simulated -2LLR follows Type VI density and an estimate can be given as:

$$f(x) = \frac{3.99819^{56}(-4.64171 + x)^{1.69362}}{(45.63029 + 1x)^{33.69305}}, 4.64171 \leq x \leq \infty \tag{1.31}$$

Table 1.2 indicates that the power of the test is almost same as that of the Fisher's g test when the series possesses Fourier frequency. Although the power of Fisher's g test in the series with non-Fourier frequency is better than that for $n = 10$, it is still not satisfactory.

Power comparison; $n = 20$			
σ	$\lambda = 0.0$	$\lambda = 0.1$	$\lambda = 0.125$
0.1	0.0503|0.0509	1.0000|1.0000	0.3848|1.0000
0.2	0.0493|0.0497	1.0000|1.0000	0.3310|1.0000
0.3	0.0507|0.0510	1.0000|1.0000	0.2883|1.0000
0.4	0.0517|0.0521	0.9995|0.9987	0.2400|0.9992
0.5	0.0522|0.0525	0.9858|0.9697	0.2042|0.9702
0.6	0.0485|0.0532	0.9078|0.8585	0.1653|0.8532
0.7	0.0504|0.0494	0.7610|0.6910	0.1389|0.6839
0.8	0.0567|0.0561	0.6048|0.5214	0.1143|0.5158
0.9	0.0501|0.0493	0.4730|0.3995	0.1011|0.3990
1.0	0.0497|0.0486	0.3583|0.3036	0.0856|0.2912

Table 1.3: Power comparison of Fisher's test and Pearson curve fitting method for a series of length 20. The simulation was carried out 10^4 times. $\lambda = 0.1$ and $\lambda = 0.125$ correspond to Fourier and non-Fourier frequencies respectively. First element in each column represents the power for Fisher's g test and the second one represents that for our proposed method.

1.4.3 Sample Size-50

Series size of 50 gives a simulated Pearson VI curve as:

$$f(x) = \frac{2.08^{53}(-5.1567 + x)^{2.49946}}{(32.93073 + 1x)^{34.22725}}, 5.1567. \leq x \leq \infty \tag{1.32}$$

Table 1.3 reveals that the performance of Pearson curve fitting and Fisher's g test is almost same for Fourier frequency. With non-Fourier frequency and large value of σ, the latter still does not have very good power of the test.

1.4.4 Multiple Tests

As shown in our previous simulation results, Pearson curve fitting has better power than Fisher's exact test. We make comparison of the multiple test methods to see how each one can pick the number of significant genes. Similar to what was described by Wichert *et al.* (2004), we simulate the data with 100 periodic and 1900 random genes for series length of $10, 20$ and 50. We consider $\lambda = 1/2\pi$ and $\sigma = 0.1$ in our original

Power comparison; $n = 50$

σ	$\lambda = 0.0$	$\lambda = \frac{1}{10}$	$\lambda = \frac{11}{100}$
0.1	0.0492|0.0493	1.0000|1.0000	1.0000|1.0000
0.2	0.0519|0.0487	1.0000|1.0000	1.0000|1.0000
0.3	0.0494|0.0482	1.0000|1.0000	1.0000|1.0000
0.4	0.0469|0.0458	1.0000|1.0000	0.9995|1.0000
0.5	0.0535|0.0508	1.0000|1.0000	0.9814|1.0000
0.6	0.0503|0.0524	1.0000|0.9999	0.8962|0.9999
0.7	0.0495|0.0492	0.9996|0.9987	0.7568|0.9981
0.8	0.0543|0.0531	0.9914|0.9859	0.5930|0.9877
0.9	0.0471|0.0437	0.9583|0.9380	0.4777|0.9416
1.0	0.0504|0.0515	0.9020|0.8617	0.3571|0.8659

Table 1.4: Power comparison of Fisher's test and Pearson curve fitting method for a series of length 50. The simulation was carried out 10^4 times. Here $\lambda = 1/10$ and $\lambda = 11/100$ correspond to Fourier and non-Fourier frequencies respectively. The format of the elements is same as that is in 1.1.

model. The multiple test levels are taken as $q = 0.001, 0.01, 0.05$, 0.1 and 0.15. The simulation results are presented in Table 1.5. When the error variance is considered to be very small, the detection of periodicity is obvious. In each case we see that Benjamini and Hochberg's FDR (BH) and q-value approaches give similar results. SPLOSH gives closest number of significant genes to the actual one. However, as can be expected the performance depends on the length of the series. For example, no multiple test method is able to detect any periodicity at a level of $q = 0.001$ when the series length is 10. But with increased level, e.g, $q = 0.05$, BH, q-value and SPLOSH detect 80, 83 and 22 periodic genes respectively. This simulation results also confirm that q-value approach is more liberal than BH approach.

1.5 APPLICATION TO GENE EXPRESSION DATA

Periodicity in different gene expression data was discussed by Wichert *et al.* (2004). We use all the data sets they analyzed for implementing our proposed method. These

Selecting periodic genes using different multiple test methods

q	$N = 10$	$N = 20$	$N = 50$	$N = 100$
0.001	0|0|0	100|100|98	101|101|100	100|100|100
0.01	16|24|3	103|103|100	103|103|101	101|102|100
0.05	80|83|22	110|110|100	108|109|101	109|110|100
0.1	101|105|33	116|117|100	115|118|101	117|119|100
0.15	113|105|33	117|123|100	121|125|101	126|129|100

Table 1.5: Simulation results for the performance of different multiple test methods to detect number of periodic genes. In each column first, second and third elements are the numbers obtained by BH, q-value and SPLOSH respectively. The simulation was carried out similar to that explained by Wichert *et al.* (2004). We simulate 100 periodic genes from a model $z_t = \cos(2\pi\lambda t) + e_t$, where $\lambda = 1/2\pi$ and $e_t \sim \text{NID}(0, 0.3)$; other 1900 genes are selected to have random Gaussian process.

cell cycle data sets contain various number of series lengths, and so the application of the new method and different multiple test methods would verify the performance that was discussed in simulation section. We shortly discuss the results obtained in all the experiments, but present in detail that for *Caulobacter crescentus* cell cycle data, which has one of the shortest series among all the experiments.

1.5.1 Yeast Cell Cycle

Spellman *et al.* (1998) analyzed the yeast *Saccharomyces cerevisiae* microarray experiments. There are four gene expression experiment data set with three different cell cycle synchronization techniques. The gene expression data sets are `cdc15`, `cdc28`, `alpha` and `elution`. The periodicity analysis of `cdc15` and `elution` indicates that the proposed method detects less periodic genes than that by Fisher's g test while different multiple test methods are taken into consideration. However, for the `cdc28` and `alpha` experiments, the number of periodic genes detected by Pearson curve fitting method is much higher than that by other approach.

Table 1.9 for `alpha` experiment indicates that the proposed method results in 421, 283 and 236 periodic genes using q-value, SPLOSH and BH approaches respectively.

Whereas, Fisher's g results in 347, 150 and 193 periodic genes using the three multiple tests respectively. However, it is difficult to distinguish cell-cycle specific variation from an artifact of the method used to synchronize the cells.

Number of Significant genes in `cdc15` data						
	$< 1e-04$	< 0.001	< 0.01	< 0.025	< 0.05	< 0.1
p-value	92|97	249|290	613|794	897|1121	1192|1406	1610|1807
q-value	11|0	50|57	293|493	540|893	879|1293	1422|1917
SPLOSH	29|26	64|80	171|256	251|738	721|1169	1318|1796
BH	9|0	33|7	171|216	324|473	490|767	797|1139

Table 1.6: Number of significant genes obtained in `cdc15` experiment of Yeast cell cycle data. The left hand side values are for Pearson curve fitting method and the other ones are for Fisher's g test.

Number of Significant genes in `cdc28` data						
	$< 1e-04$	< 0.001	< 0.01	< 0.025	< 0.05	< 0.1
p-value	21|8	76|27	172|112	247|185	322|265	430|361
q-value	0|0	10|0	76|8	119|25	196|56	294|141
SPLOSH	6|0	12|2	36|12	60|21	127|26	256|112
BH	0|0	8|0	34|6	89|13	123|27	205|95

Table 1.7: Number of significant genes obtained in `cdc28` experiment of Yeast cell cycle data. The left hand side values are for Pearson curve fitting method and the other ones are for Fisher's g test.

	Number of Significant genes in `alpha` data					
	$< 1e-04$	< 0.001	< 0.01	< 0.025	< 0.05	< 0.1
p-value	45|94	145|259	403|599	578|840	788|1109	1092|1493
q-value	1|0	8|45	80|266	179|449	307|682	494|1048
SPLOSH	3|12	16|56	52|156	86|221	124|501	295|940
BH	0|0	6|10	46|169	127|306	241|469	376|711

Table 1.8: Number of significant genes obtained in `alpha` experiment of Yeast cell cycle data. The left hand side values are for Pearson curve fitting method and the other ones are for Fisher's g test.

	Number of Significant genes in `elution` data					
	$< 1e-04$	< 0.001	< 0.01	< 0.025	< 0.05	< 0.1
p-value	49|44	166|151	534|519	841|836	1217|1191	1674|1678
q-value	0|0	11|1	83|61	214|151	421|347	924|767
SPLOSH	9|4	29|18	75|65	124|99	283|150	854|680
BH	0|0	7|1	43|34	123|91	236|193	475|458

Table 1.9: Number of significant genes obtained in `elution` experiment of Yeast cell cycle data. The left hand side values are for Pearson curve fitting method and the other ones are for Fisher's g test.

1.5.2 Bacterial Cell Cycle

This data set was analyzed by Laub *et al.* (2000). To identify cell cycle-dependent transcripts, the discrete cosine transform (DCT) was calculated for each of the 2966 expression profiles with valid data (Laub *et al.*, 2000). They identified 553 genes whose messenger RNA levels varied as a function of the cell cycle. The data set consists of very small sequence length (11) and so the test of periodicity using Fisher's g test would fail to detect the periodic genes if the true frequency is not Fourier frequency.

As our proposed method can handle the test of periodicity in a series with missing observation, we do not omit data with a single missing observation. Rather, in this data set we keep the series with three or less missing values. We do not analyze series whose ORF names are missing. *Caulobacter crescentus* cell cycle data has most of the missing observations at the end and so Fisher's g test can be applied to such series. We get 2460 ORFs having at most three missing values, out of which 1584,

463, 284, and 129 are the number of series with none, one, two and three missing values respectively. 2361 of the series have either no missing value or the missing values are at the end.

Number of Significant genes in bacterial cell cycle data

	$< 1e-04$	< 0.001	< 0.01	< 0.025	< 0.05	< 0.1	< 1
p-value	2|6	26|33	127|112	201|160	272|205	380|275	1425|1440
q-value	0|0	0|0	1|0	1|27	50|45	150|95	1440|1440
SPLOSH	0|0	1|0	4|3	13|10	26|20	131|29	1440|1440
BH	0|0	0|0	0|0	1|23	1|43	93|95	1396|1228

Table 1.10: Selection of periodic genes using different multiple test methods (BH, q-value and SPLOSH) after applying Fisher's g test and Pearson curve fitting method to the *Caulobacter crescentus* cell cycle data. The series with any missing value are excluded from the analysis. The left hand side values are for Pearson curve fitting method and the other ones are for Fisher's g test.

Fisher's g test taking missing at the end

	< 0.0001	< 0.001	< 0.01	< 0.025	< 0.05	< 0.1
p-value	7	40	150	227	298	413
FDR	0	0	0	0	37	92
q-value	0	0	0	0	37	92
SPLOSH	0	0	0	4	21	34

Table 1.11: Fisher's g test was applied to rows of *Caulobacter crescentus* cell cycle data where we have missing observations at the end. The results show the number of periodic genes selected using different multiple test methods (BH, q-value and SPLOSH).

Table 1.10 and 1.11 present the number of significant genes derived through different multiple test procedures when test of periodicity is applied to sequences with no missing values. Fisher's g test provides 205 genes having p-value less than 0.05. But multiple test level of $q = 0.05$ produces 43, 45 and 20 significant cell-cycle regulated genes using BH, q-value and SPLOSH respectively. If we apply our Pearson curve fitting approach, it gives 317 genes having p-values less than 0.05. By this method, 1, 50 and 26 significant genes are detected using BH, Storey's q-value and SPLOSH

Pearson Curve fitting taking up to three missing						
	< 0.0001	< 0.001	< 0.01	< 0.025	< 0.05	< 0.1
p-value	2	31	186	313	442	628
BH	1	1	1	1	1	76
q-value	0	1	1	1	1	152
SPLOSH	0	0	5	17	62	228

Table 1.12: Pearson curve fitting performance on finding periodic genes in *Caulobacter crescentus* cell cycle data when we take into account up to three missing values in the series. Results using three multiple test methods are included.

respectively when $q = 0.05$.

When the method is applied to the data set under the fact that ignoring series with missing cells might not be a good way to analyze it, we get different number of significant genes for different multiple testing methods and different periodicity testing methods. Table 1.12 shows the result when our proposed method is applied for the test of periodicity. At a significance level of $q = 0.05$, the number of significant ORFs are 1, 1 and 62 using BH, q-value and SPLOSH respectively. Fisher's g test is applied to the series having missing values at the ends. This gives number of significant genes to be 37, 37 and 21 respectively using BH, q-value and SPLOSH approach.

We see that all the ORFs, revealed to have periodic movement using Fisher's g test, are also significantly periodic using our method. Fisher's g test failed to detect some periodic gene; for example, ORF05387 is a highly periodic series, as can be seen from the Figure 1.5. It fails to remain in a list of 200 most significant genes when Fisher's g test is applied. However, Pearson curve fitting results a p-value of 0.0004108737 for this gene and thus indicating to be highly periodic.

One of the ORFs is repeated twice. SPLOSH produces 61 significant genes at a threshold level $q = 0.05$ when Pearson curve fitting is applied to the data. We present first 24 genes in Figure 1.1. If the error process in a periodic or random series has

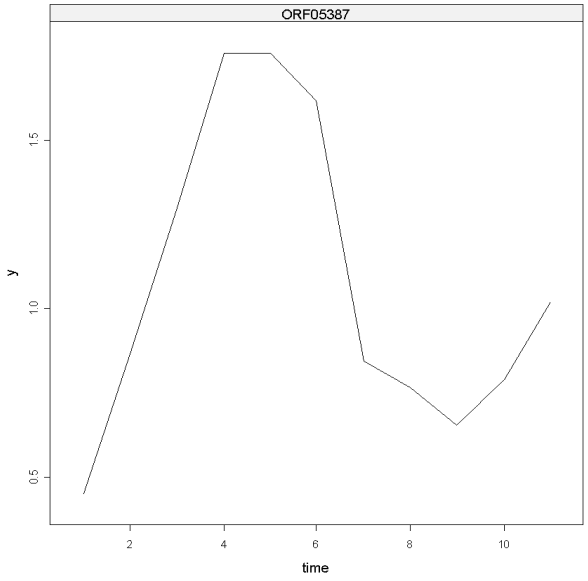

Figure 1.5: An ORF with highly periodic pattern. Fisher's g test fails to detect the periodicity, but our method provides a p-value of 0.0004108737 for the test.

autoregressive error process, this might yield more number of significant cell-cycle regulated genes. This is due to the entanglement of periodicity with autoregressive errors. We tested the non-whiteness of the error process in *Caulobacter crescentus* cell cycle data. None of the series was found to have non-white error process after applying multiple testing methods. 20 ORFs having the smallest p-values for the test of autoregressive error process are listed as:

$ORF07061$ $ORF04700$ $ORF02759$ $ORF04977$ $ORF03161$ $ORF03165$
$ORF04480$ $ORF03823$ $ORF03156$ $ORF01232$ $ORF00526$ $ORF00076$
$ORF00854$ $ORF08039$ $ORF07002$ $ORF02058$ $ORF00968$ $ORF05154$
$ORF05154$ $ORF00818$

All but four of these ORFs (ORF03161 ORF04480 ORF03165 ORF03823) are in the list of 61 most significantly periodic series. However, they are all periodic with p-values 0.002031544, 0.002522119, 0.003564576 and 0.003730593 respectively.

1.5.3 Human Fibroblasts

Cho *et al.* (2001) designed the microarray experiments for human fibroblasts cells. The data sets are two short time series with 12 observations. The implementation of Fisher's g test and Pearson curve fitting method along with multiple test methods detected almost no periodicity in the data sets and this is quite apparent in Tables 1.13 and 1.14. Contrary to what was described by Wichert *et al.* (2004), this non-periodicity cannot be derived by *average periodogram* (AP), which will be described in next section. Their work picked no periodicity through AP due to usage of unnecessary large scale for periodogram ordinates; but using appropriate scaling would refer periodicity through AP.

	Number of Significant genes in Human Fibroblasts N2 data					
	$< 1e-04$	< 0.001	< 0.01	< 0.025	< 0.05	< 0.1
p-value	1|2	8|6	50|79	136|205	296|372	602|756
q-value	0|0	0|0	0|0	0|0	0|1	0|2
SPLOSH	0|0	0|0	0|1	0|2	1|2	1|2
BH	0|0	0|0	0|0	0|0	0|1	0|2

Table 1.13: Performance of Pearson curve fitting method and Fisher's g test in selecting periodic genes in Human Fibroblasts N2 data. In each column, the results from former are at the left and those for other are at the right.

	Number of Significant genes in Human Fibroblasts N3 data					
	$< 1e-04$	< 0.001	< 0.01	< 0.025	< 0.05	< 0.1
p-value	1|1	7|14	80|105	187|253	379|497	765|891
q-value	0|0	0|0	0|0	0|0	0|0	0|0
SPLOSH	0|0	0|0	0|0	1|0	1|1	1|2
BH	0|0	0|0	0|0	0|0	0|0	0|0

Table 1.14: Performance of Pearson curve fitting method and Fisher's g test in selecting periodic genes in Human Fibroblasts N3 data. In each column, the results from former are at the left and those for other are at the right.

1.5.4 Human Cancer Cell Line

The design of the microarrary experiment was described by Whitefield *et al.* (2002). There are five experiments - `score1`, `score2`, `score3`, `score4` and `score5` with different series length. The measurements are taken for time points 12, 26, 48, 19 and 9 respectively. The expression levels were measured for varying number of genes. Three different cell cycle synchronization methods were used; namely, a double thymidine block (`score1`, `score2`, `score3`), thymidine followed by arrest in mitosis with nocodazole (`score4`) and mitotic shake-off using an automated cell shake (`score5`). Tables 1.15–1.19 indicate that there have been a huge difference in the outcome of Fisher's g test and Pearson curve fitting approach.

Number of Significant genes in `score1` of Human HeLa data

	$< 1e-04$	< 0.001	< 0.01	< 0.025	< 0.05	< 0.1
p-value	4|11	49|64	289|367	658|787	1175|1380	2196|2411
q-value	0|0	0|0	0|1	0|1	0|2	0|25
SPLOSH	0|1	0|1	2|6	3|9	6|16	8|29
BH	0|0	0|0	0|1	0|1	0|1	0|5

Table 1.15: Number of significant periodic genes obtained in `score1` of Human HeLa data. In each column, the left hand side values are for Pearson curve fitting method and the other ones are for Fisher's g test.

Number of Significant genes in `score2` of Human HeLa data

	$< 1e-04$	< 0.001	< 0.01	< 0.025	< 0.05	< 0.1
p-value	56|59	219|251	829|922	1352|1552	2032|2295	3083|3508
q-value	0|0	5|3	22|40	59|96	171|307	439|781
SPLOSH	6|4	11|16	37|57	64|96	99|148	158|236
BH	0|0	5|2	12|17	40|57	122|148	340|403

Table 1.16: Result of Pearson curve fitting method and Fisher's g test in selecting periodic genes. Both methods were applied to `score2` of Human HeLa data. The left hand side values are for Pearson curve fitting method and the other ones are for Fisher's g test.

Number of Significant genes in `score3` of Human HeLa data

	$< 1e-04$	< 0.001	< 0.01	< 0.025	< 0.05	< 0.1
p-value	2563|2491	3762|3966	5864|6604	7362|8425	9014|10440	11511|13255
q-value	1571|1444	2452|2437	3891|4382	4774|5690	5748|7063	7145|9166
SPLOSH	1181|1132	1687|1728	2467|2677	2872|3203	3219|3677	3607|5987
BH	1503|1304	2342|2213	3676|3914	4500|5053	5371|6128	6530|7808

Table 1.17: In each column, the left and right entries represent the number of periodic genes obtained using Pearson curve fitting method and Fisher's g test respectively. Both methods were applied to `score3` of Human HeLa data.

Number of Significant genes in `score4` of Human HeLa data

	$< 1e-04$	< 0.001	< 0.01	< 0.025	< 0.05	< 0.1
p-value	159|65	525|240	1912|954	3214|1708	4790|2907	7440|5070
q-value	0|0	12|1	103|2	229|21	507|57	242|128
SPLOSH	12|2	40|5	130|23	201|43	279|60	405|94
BH	0|0	1|1	64|2	157|21	314|57	767|126

Table 1.18: In each column, the left and right entries represent the number of periodic genes obtained using Pearson curve fitting method and Fisher's g test respectively. Both methods were applied to `score4` of Human HeLa data.

Number of Significant genes in `score5` of Human HeLa data

	$< 1e-04$	< 0.001	< 0.01	< 0.025	< 0.05	< 0.1
p-value	0|4	33|37	303|352	791|838	1566|1686	3212|3417
q-value	0|0	0|0	0|0	0|0	0|0	0|0
SPLOSH	0|0	0|0	0|0	0|0	0|1	0|4
BH	0|0	0|0	0|0	0|0	0|0	0|0

Table 1.19: Number of significant genes obtained in `score5` of Human HeLa data. The left hand side values are for Pearson curve fitting method and the other ones are for Fisher's g test.

1.6 AVERAGE PERIODOGRAM (AP)

Under certain conditions, the plot of periodogram against the Fourier frequencies can be a useful device to observe whether there is any sinusoidal pattern in the series. When thousands of series are analyzed together, Wichert *et al.* (2004) proposed using *average periodogram* (AP) for visual inspection for the presence of periodicity in the series. They defined the AP as

$$\mathrm{AI}(\lambda) = \frac{1}{G}\sum_{i=1}^{G} I_i(\lambda) \tag{1.33}$$

where G is the number of time series present in the analysis and $I_i(\lambda)$ is the periodogram for the i-th series. Wichert *et al.* (2004) justified the use of *average periodogram* as follows: if the data follows a pure random process then the periodogram of all time series is uniform and therefore the average estimate should reduce to a straight line; if there are a few time series exhibiting strong periodicity, then their corresponding periodogram ordinates dominate the AP. However, we believe that this graphical device would not be applicable if the series with strong periodicity do not have same frequency.

In practical use, all the series might not have equal frequencies, and so resorting to this kind of graphical device in signal detection can give misleading results. We see in figures 1.6, the frequencies are randomly selected from a set of Fourier frequencies as can happen in real data set. For each series with sizes $N = 10, 20$ and 40, AP is not capable of detecting periodicity.

In the summary, we can say that AP plots the distribution of estimated Fourier frequencies obtained from the working series. Figures 1.7 and 1.8 show the distribution of the estimated frequencies from different cell cycle data sets. We add AP with the corresponding distribution and this confirms the aforementioned statement.

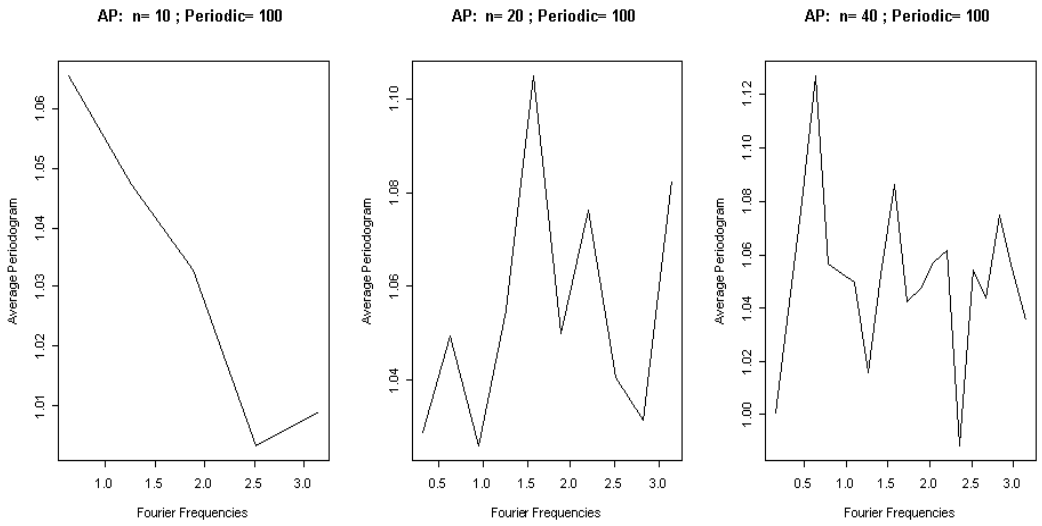

Figure 1.6: *Average periodogram* for $N = 10, 20$ and 40 when there are 100 periodic and 1900 random genes in the whole data set. The frequencies are not equal for all series but selected randomly from the set $2\pi/10$, $4\pi/10$, $6\pi/10$, $8\pi/10$, π for $N = 10$. For $N = 20$ and 40 the frequencies are selected from the sets $\{2\pi/20$, $4\pi/20$, $6\pi/20$, ... ,$\pi\}$ and $\{2\pi/40$, $4\pi/40$, $6\pi/40$, ..., $\pi\}$ respectively.

1.7 DISCUSSION

The analyses demonstrate that checking for cell cycle regulated genes in short microarray time series data requires consideration of both periodicity testing technique as well as multiple testing method. In fact, there is no guarantee that the series will possess only Fourier frequencies. There exists only asymptotic theory to test periodicity in a series with non-Fourier frequency. Details can be found in Turkman and Walker (1984). Simulation results show that Fisher's g test fail 100% of time even in an almost perfectly periodic series with non-Fourier frequency.

Average periodogram represents the distribution of estimated frequencies of the series. If the dominating frequencies are not close to each other, then this graphical device might not work. Various simulation procedures were done to evaluate the performance of this method proposed by Wichert *et al.* (2004).

If the series has unequal time intervals, all the traditional methods fail to perform

the test of periodicity. However, the proposed method is still useful in such situation. Permutation test is a straightforward method, but this imposes some lower bound to the p-values. Therefore, the result obtained from this method might not be very satisfactory in multiple testing. We plan to carry out further simulation experiments to compare the statistical power for our method with a permutation test. The permutation test is being implemented in R, so we can carry out these computations using the a Beowulf cluster computer. We believe this will demonstrate the superiority of the Pearson curve fitting method.

Although selecting the cell cycle regulated genes in a series with autoregressive error process was not a problem in the data we analyzed, this issue might need to be considered in other microarray time series. The lack of applicability of Fisher's g test and other test procedure in the series of missing observations or the series having unequal time sequence can be overcome by the proposed simulation method.

From the simulation result and the application of real data, it was seen that multiple test methods play a big role in the selection of cell cycle regulated genes. Our proposed method resulted far more significant genes in some of the data sets; namely `cdc28`, `elution` in yeast cell cycle and `score4` in Human cell cycle data set. Some data sets tested to have more sinusoidal gene expressions using Fisher's g test. Human fibroblasts with N2 and N3 experiments as well as two experiments `score1` and `score5` in human cell cycle data sets seem to have no noticeable genes of sinusoidal pattern after implementing Fisher's g or our proposed method. SPLOSH seems to be more conservative in our study. However, the final decision in detecting the genes should be made in accordance with biological interpretation.

For efficiency of the computation, R is interfaced to C routines for generating the null distributions. An R package, `GenePeriodicity`, has been developed for all the implementations of the method in this chapter.

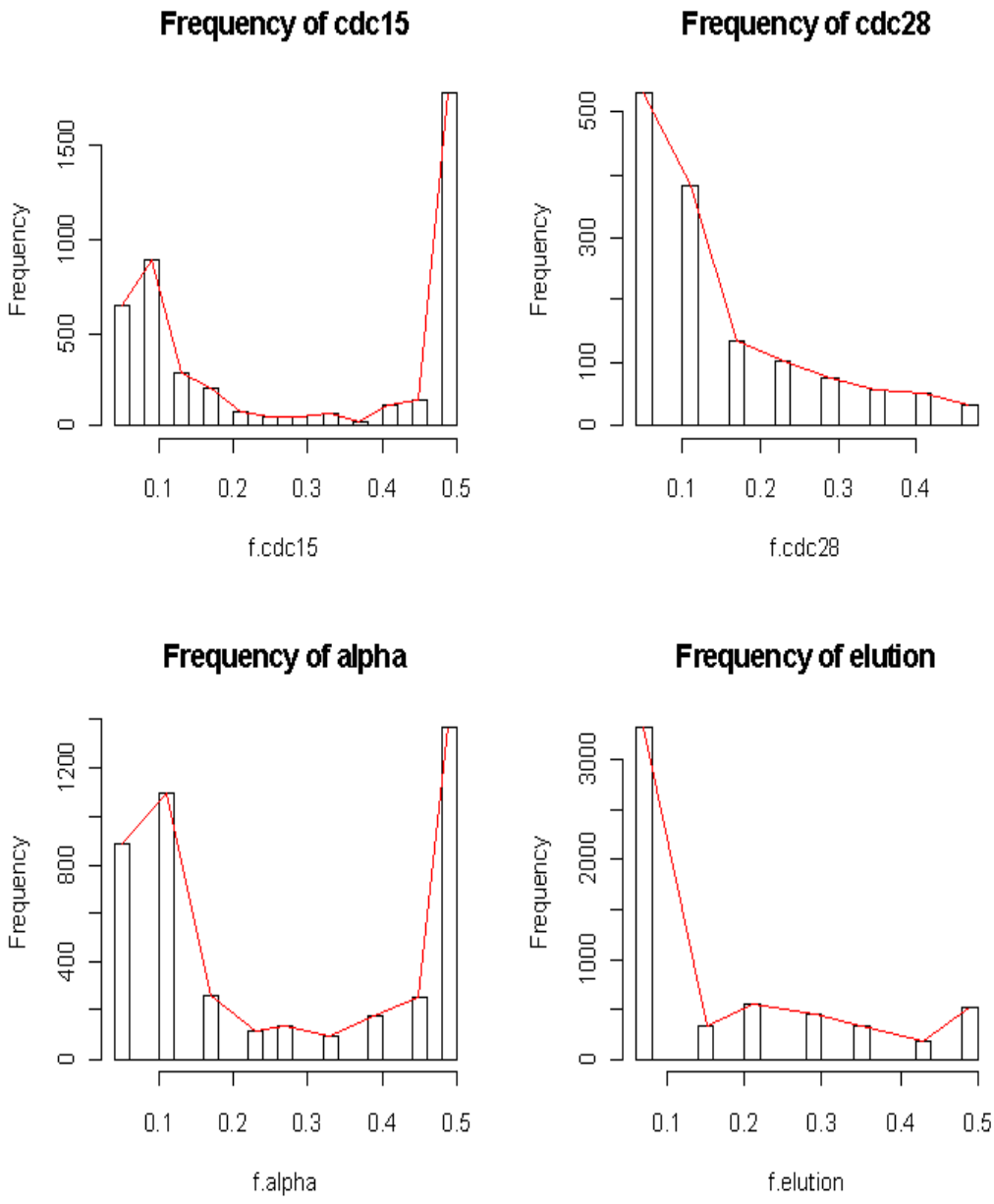

Figure 1.7: Distribution of estimated Fourier frequencies for yeast *Saccharomyces cerevisiae* microarray experiments. The line represents AP which is added to each of the histograms. These indicate that distribution of periodogram and AP represent the same feature.

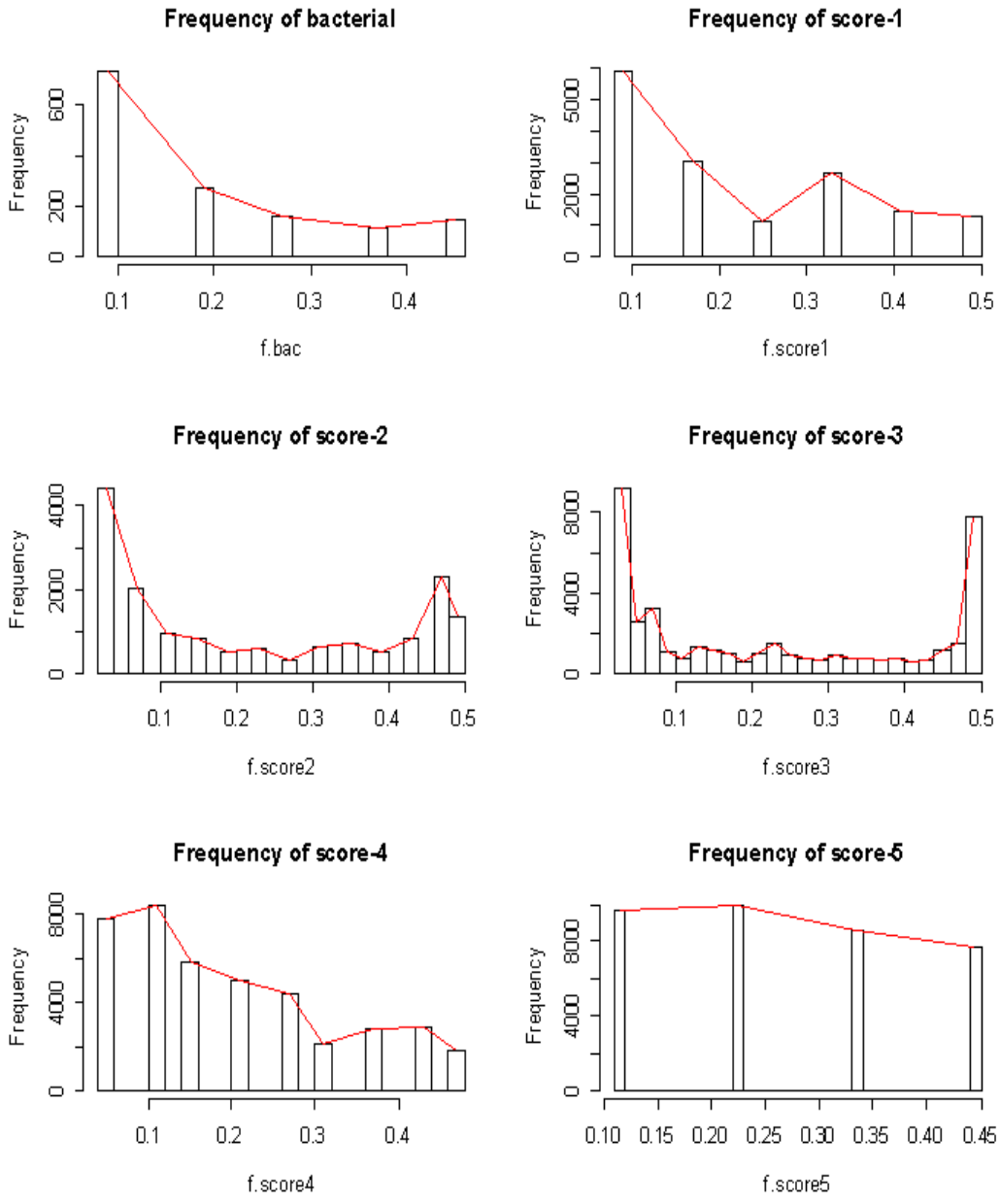

Figure 1.8: Distribution of estimated Fourier frequencies for bacterial cell cycle. The line represents AP which is added to each of the histograms. These indicate that distribution of periodogram and AP represent the same feature.

REFERENCES

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, **57**, 289-300.

Bloomfield, P. (2000) *Fourier Analysis of Time Series: An Introduction*. John Wiley & sons, Inc.

Chiu, S. (1989) Detecting Periodic Components in a White Gaussian Time Series. *J. R. Statist. Soc. B*, **51(2)**, 249-259.

Cho, R. J., Huang, M., Dong, H., Steinmetz, L., Saponoso, L., Hampotn, G., Elledge, S. J., DAvis, R. W., Lockhadt, D. J. and Campbell, M. J. (2001) Transcriptional regulation and function during the human cell cycle. *Nat. genet.*, **27**, 48-54.

Good, P. (2000) *Permutation Tests*. Springer-Verlag, New York.

Johanson, N. I. and Kotz, S. (1970) *Continuous univariate distributions -1*. Houghton Mifflin Company, Boston.

Laub, M. T., McAdams, H. H., Feldblyum, T., Fraser, C. M. and Shapiro, L. (2000) Global analysis of the genetic network controlling a bacterial cell cycle *Science*, **290**, 2144-2148.

Pounds, S. and Cheng, C. (2004) Improving false discovery rate estimation, *Bioinformatics*, **20**, 1737-1745.

Quin, B. G. and Hannan, E. J. (2001) *The Estimation and Tracking of Frequency*. Cambridge University Press.

Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Anders, K., Eisen, M. B., Brown, P. O., Botstein, D. and Futcher, B. (1998) Comperative identification of cell cycle-regulated genes of the yeast *Saccharomyces cervisiae* by microarray hybridization. *Mol. Biol. Cell*, **9**, 3273-3297.

Storey, J. D.(2002) A direct approach to false discovery rates. *J. R. Stat. Soc. B*,**64**, 479-498.

Storey, J. D. and Tibshirani, R. (2003) Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci., USA*, **100**, 9440-9445.

Turkman, K. F. and Walker, A. M. (1984) On the asymptotic distributions of maxima of trigonometric polynomials with random coefficients. *Adv. Appl. Prob.*, **16**, 819-842.

Walker, A. M. (1965) Some asymptotic results for the periodogram of a stationary time series. *J. Aust. Math. Soc.*, **5**, 107-128.

Wichert, S., Fokianos, K. and Strimmer, K. (2004) Identifying periodically expressed transcrips in microarray time series data. *Bioinformatics*, **18**, 5-20.

Whitefield, M. L., Sherloc, G., Saldanha, A. J., Murrary, J. I., Ball, C. A., Alexander, K. E., Matese, J. C., Perou, C. M., Hurt, M. M., Brown, P. O. and Botstein, D. (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977-2000.

Chapter 2

A METHOD FOR ANALYSIS OF CGH MICROARRAY DATA

2.1 INTRODUCTION

Genetic DNA copy number alterations are important features for the development of many diseases. A normal human cell contains two copies of each of the 22 non-sex chromosomes. DNA copy numbers change from two in case of genetic alterations. Deletions of copy numbers contribute to the alterations in the expression of tumor-suppressor genes, whereas amplifications contribute to the alterations in oncogenes. The changes in gene expression modify the normal growth control and survival pathways. Thus, for understanding disease phenotype and for localizing important genes, it is important to characterize the DNA copy number changes. *Comparative Genomic Hybridization* (CGH) microarray is a technique for measuring such changes (Pinkel and Albertson, 2005). As a high throughput technique, it offers many advantages over other cytogenetic techniques such as *Fluorescence In Situ Hybridization* (FISH). More recently, cDNA and oligonucleotide arrays have become popular for CGH. The shorter probes on these arrays provide design flexibility and greater coverage, and the resultant high-throughput CGH data have prompted the development of various methods for data analysis. See Lai *et al.* (2005) and Willenbrock and Fridlyand (2005) for comparative reviews of the analysis methods.

In a CGH experiment, a test sample labelled red (Cy5) is hybridized to a reference normal sample labelled green (Cy3), and the resulting data consists of the ratio of

the fluorescence intensities from test versus reference sample, indexed by the physical location of the clones on the genome. The arrays in CGH experiment are constructed with the assumption that the ratio of binding of test and control DNA is proportional to the ratio of the copy numbers of the corresponding DNA sequences. Alterations in DNA copy number typically occur through the gain or loss of chromosomal segments. In a homogenous cell population the actual DNA copy number profile of the genome consists of a series of plateaus of constant copy number, bounded by sharp transitions. Thus the alterations correspond to the regions of concentrated high or low log-ratios on the genome.

Various methods have already been proposed to study and solve the challenge of efficiently identifying the regions with DNA copy number alterations. For example, Pollack *et al.* (2002) applied a moving average to the process of ratios and used normal versus normal hybridization to compute the threshold; Hodgson *et al.* (2001) used a maximum likelihood to fit mixture models corresponding to gain, loss and normal regions; Lingjaerde *et al.* (2001) employed a simple smoothing to signs of neighbours and significance is described by comparing both the height and weight of the observed segments with their joint null distribution. Wang *et al.* (2005) proposed an algorithm *Cluster Along Chromosomes* (CLAC), which builds hierarchical clustering-style trees along each chromosome arm (or chromosome), and then selects the clusters by controlling the *False Discovery Rate* (FDR) at a certain level. CLAC is available as an R package, `clac`, from CRAN.

The log-ratio sequence is viewed as a time series sequence along the genome by considering the possible correlation between clones at closer physical locations on the genome. The problem of change point detection in such series is closely related to the problem of detecting discontinuities in signal processing and edge-detection in image analysis. Wavelet methods are widely used for these problems. For example,

for detecting discontinuity, one method recommends using the Haar Wavelet and looking at the lowest two levels of detail (Matlab, 2007). The MatLab approach is purely exploratory. Wang (1995) proposed a method for identifying the jumps in a time series by checking if wavelet transformation of the data has large absolute values across fine scale levels.

We propose a new method for determining the change point of log-ratio. Maximum overlapping discrete wavelet transform (MODWT) is employed for this purpose. This technique provides higher resolution for the location on the chromosome where the break occurs. The method can automatically and efficiently detect the change points and hence the gain and loss regions along the whole genome. This method utilizes Wang's threshold value to define significant jumps from the previous region. Double application of MODWT at level one is used to confirm the presence of true abnormal regions in the sequence.

The organization of the chapter is as follows. In Section 2.2 we introduce the models and applications of wavelet methods to the CGH data. Some simulated examples are demonstrated in Section 2.3 to show the performance of the proposed method. Section 2.4 is devoted to the application of the method to real CGH data. A brief discussion is presented in Section 2.5.

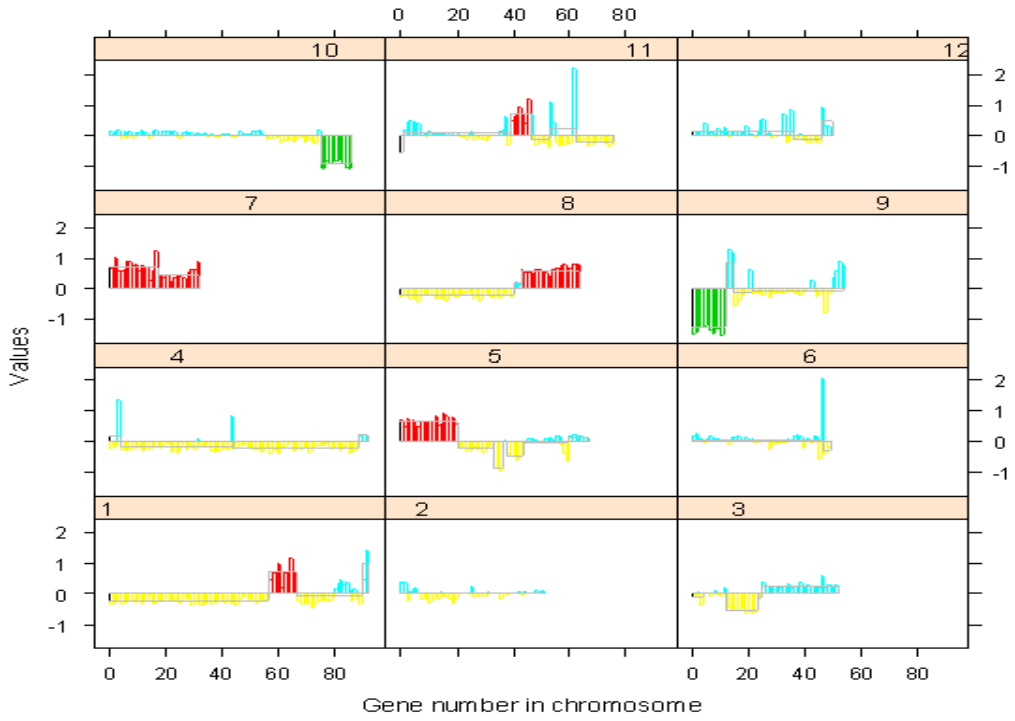

Figure 2.1: Representation of regions of copy number changes detected by Wavelet method to a CGH array. In this data set, 2400 BAC clones were measured each with three replicates (Snijders *et al.*, 2001). Measurements for log base 2 intensity ratio are provided. Average relative DNA copy number sequences of the three replicates in first 12 chromosomes are shown in this figure. Red color refers to detected copy number amplification region; whereas, green color refers to deletion region.

2.2 NOTATION AND MODELS

Microarray based CGH provides the relative copy number of the spotted DNA sequences by monitoring the differential hybridization of two samples to the sequences on the array. Let $z_t, t = 1, 2, ..., n$ be the measure of the relative DNA copy numbers of n clones along the genome. Usually z_t is the logarithm with base 2 of the intensity ratio of test sample versus the reference sample. There are systematic variations in microarray experiments and so normalization procedures are applied to remove those noises. We assume here that all the data are normalized. Identifying or screening

the genes that have DNA copy number gain or loss is equivalent to describing the genes locations on the genome where the DNA copy numbers increase or decrease. We assume that the DNA copy number follows a distribution F_0 in a region on the genome, and after the location k, the distribution is changed to F_1; so we can write,

$$\begin{aligned} z_1, z_2, \ldots, z_k &\sim F_0 \\ z_{k+1}, z_{k+2}, \ldots, z_n &\sim F_1 \end{aligned}$$

That is equivalent to finding the change point k, where the distribution of the relative copy numbers are different on both sides of k. Note that for the CGH data, there may be many change points along the genome and these points define the regions of gains or losses of the copy numbers. If the clones on the genome are close enough, they might affect each other on copy numbers. Thus we can assume that the copy number of a clone on the genome is associated with that of the previous clone. The copy numbers sequence along the genome can therefore be envisaged as a time series. Determination of change points is equivalent to the determination of abrupt change along the sequence. Wavelets are ideally suited for this purpose.

2.2.1 Wavelet Methods

Wavelets are well established in the mathematical sciences (Daubechies, 1992) and have been successfully applied in fields such as signal and image processing, numerical analysis and statistics. Wavelets literally means small waves. A function $\psi(.)$, defined over the entire real axis, is called a wavelet if $\psi(.) \to 0$ as $t \to \pm\infty$ and satisfying the

following conditions:

$$\int_{-\infty}^{\infty} \psi(u)\mathrm{d}u = 0 \tag{2.1}$$

$$\int_{-\infty}^{\infty} \psi^2(u)\mathrm{d}u = 1 \tag{2.2}$$

Wavelets are functions that can be used to describe a signal efficiently by breaking it down into its components at different scales and following their evolution in the time domain. Wavelets tell us the changes in averages in a time series. These changes in averages are computed in terms of weighted average differences of the series over different time scales, denoted by λ. The variation of λ can provide information about how averages of $x(.)$ over many different scales can change from one period of length λ to the next. The collection of variables $\{W(\lambda, t) : \lambda > 0, -\infty < t < \infty\}$, defined in Equation 2.3, is called continuous wavelet transform (CWT).

$$W(\lambda, t) = \int_{-\infty}^{\infty} \psi_{\lambda,t}(u)x(u)\mathrm{d}u \tag{2.3}$$

In Equation 2.3, $W(\lambda, t)$ is proportional to the difference between two adjacent averages of scale λ. Here the transformed series $x(.)$ ia a function of translation parameter t and scale parameter λ. The transforming function $\psi_{\lambda,t}(u)$ is called the mother wavelet.

Discrete wavelet transformations map data from the time domain to the wavelet domain (Percival and Walden, 2000); however, the difference from CWT is that the scale λ and translation parameter t are no longer continuous. These transformations result in a vector of the same size. If we have a series of size N, wavelet transformations can be defined by the matrices of dimension $N \times N$.

The partial DWT is a special orthonormal transformation:

$$(z_0, \ldots, z_{N-1}) \longleftrightarrow (W_1, ..., W_J, V_J),$$

where W_j is a vector of length $N_j = N/2^j$; V_J has the same length as that of W_J and N_J. For simplicity we have assumed that N is a multiple of 2^J. The vector W_j is called the vector of wavelet coefficients at level j and is associated with changes or differences on scale 2^{j-1}. Vector V_J, which is the scaling coefficients at level J, is associated with averages on scale 2^{J-1}.

We can write,

$$\begin{aligned} W &= \Gamma X, \text{ where} \\ \Gamma &= \begin{pmatrix} \Gamma_1 \\ \Gamma_2 \\ \vdots \\ V_J \end{pmatrix} \end{aligned}$$

In practice the DWT is computed using the pyramid algorithm which requires only $O(N)$ flops. There are two practical limitations of DWT; these are:

- Series should be of dyadic length.
- Selecting different starting point for the series changes the result of the analysis.

First problem can be dealt through polynomial extensions of the scaling coefficients, then the DWT can be practically implemented for any size of the series. However, it is not a trivial task to select an appropriate number of end points to fit or the order of fit (Constantine and Percival, 2003). The second problem refers that the DWT is not a shift-invariant transform and so shifting the time series circularly can totally change the DWT. Maximum overlap discrete wavelet transformation (MODWT) is used to overcome such limitations. Thus MODWT provides us the advantage of making the series length and shift-invariant.

Our goal here is to identify the change point in DNA sequence. We focus on change-point approaches to data dependent thresholding. The primary idea is to divide the wavelet coefficients into groups of small coefficients containing primary noise and one of large coefficients containing significant signal. Hypothesis testing techniques are employed to obtain an appropriate threshold and a test is performed to determine if the set of coefficients at that scale contains significant signal when coefficients exceed the threshold.

2.2.2 Wang's Threshold

The underlying model or null hypothesis may be written as $z_t = f(t) + a_t, t = 1, \ldots, n$ where $a_t \sim \text{NID}\,(0, \sigma^2)$ and $f(t)$ is a smooth function, ie. continuous and differentiable. This is a classic model in time series. Examples include the polynomial trend analysis (Fisher, 1921) and the lowess polynomial seasonal adjustment of Clevelend *et al.* (1990). For forecasting purposes, the ARMA family and its extensions are more useful models. Wang (1995) considers the problem of testing an abrupt change in the function $f(t)$. This model is more general and focuses on detecting the change points at which a jump or sharp cusp occurs. A sharp cusp occurs at point t_0 if there exists a constant $K > 0$ such that

$$|f(t_0 + h) - f(t_0)| \geq K|h|^{\alpha} \tag{2.4}$$

for all h as $h \to 0$ and $0 \leq \alpha < 1$. When $\alpha = 0$, the function has a jump. Wang (1995) shows that, asymptotically with probability 1, all wavelet coefficients will have absolute value less than the universal threshold value $\sigma\sqrt{2\log(n)}$ provided there are no jumps or cusps. The unknown value of σ may be estimated robustly by the median absolute value of the wavelet coefficients at level 1 divided by 0.6745.

2.2.3 Practical Implementation Details

We follow the notation in the book by Percival and Walden (2000) which is also used in the S-Plus, R and MatLab software. The principal differences are that level 1 refers highest time domain resolution and filter width rather than half-width is used in the naming of the Daubechies wavelets. So for example, Wang's $D(1)$ corresponds to $D(2)$ which is also equivalent to the Haar wavelet. Another difference is that Wang pads the real data so that n is a power of 2 because he uses Mallat's algorithm. For detection of changepoints, Wang (1995) recommends examining plots of the absolute empirical wavelets at various levels j and finding those values which exceed the threshold line and are larger than others. Daubechies wavelets are denoted as: $D(k)$, $k = 2, 4, \ldots, 20$. The level j should be chosen as small as possible in order to obtain the highest time domain resolution.

In practice, it is not very satisfactory to examine the wavelet coefficient plots at different levels and then select the jump points, since this is a subjective and tedious procedure. We need to use some automated procedure for selecting appropriate levels for different series. MODWT at level one serves as a good strategy for this purpose. By intuition, we can think that the gain or loss region cannot contain a single observation. We apply MODWT at level one and record the observation numbers where the wavelet coefficients are greater than the Wang's threshold value. In order to verify that the wavelet coefficients correspond to right jump detection, we delete the observations where the jumps were detected and rerun the procedure. If the new wavelet coefficient adjacent to the deleted observation is again greater than the Wang's threshold and has the same sign as that of the previous coefficients, then the deleted observation in previous step is considered as true signal of jump detection.

To reach a conclusion of the analysis, we have to define the loss or gain region.

We can define a threshold beyond which a region is called to be loss or gain region according to the sign of the wavelet coefficients. The selection of the threshold using some multiple test procedure is discussed in the following subsection. The region with multiple testing value, say q-value, greater than the threshold is colored as red and the region having multiple testing value less than the threshold is colored as green. Thus red corresponds to the gain region and green corresponds to the loss region. We put a line in each detected region to represent the mean.

2.2.4 Testing Region Means Using Bootstrap

We have $z_t, t = 1, 2, \ldots, n$ as the observations along a specific chromosome arm. The observations in ith region and tth position can be expressed as

$$z_{ti} = \mu_i + e_t,\ i = 1, 2, \ldots k \text{ and } t = 1, 2, \ldots, n$$

The error term e_t follows AR(p) process, the order of which can be estimated. That is,

$$e_t = \phi_1 e_{t-1} + \phi_2 e_{t-2} + \ldots + \phi_p e_{t-p} + a_t \tag{2.5}$$

where ϕ_1, ϕ_2, ..., ϕ_p are autoregressive parameters and $e_t \sim N(0, \sigma_a^2)$.

Suppose we have only one region and we would like to test whether the region mean is significantly different from zero. A t-test procedure that considers corrected variance of $\bar{z}$ in an AR(p) error process would seem to work for such case. A short simulation study with an AR(1) process was done to see the power of this test procedure. Table 2.1-2.2 reveal that the method does not perform very well even for small ϕ values. Hence with the increase of magnitude of ϕ, the method becomes incapable of handling such situation regardless of the series length. Moreover, series length refers to the length in a particular gain/loss region, which in real CGH data will not be very large.

Power comparison; $n = 50$, $\sigma_a = 0.2$

ϕ	$\mu = 0$	$\mu = 0.5$
0.0	0.056	0.942
0.1	0.084	0.9074
0.3	0.118	0.7472
0.5	0.108	0.5186
0.7	0.137	0.3100
0.9	0.236	0.2862

Table 2.1: Power of the test $\mu = 0$ in an AR(p) setting with series length 50. Here we consider standard deviation for error term to be 0.2. The test is done at 0.05 level of significance. The column for $\mu = 0$ represents type-I error.

Power comparison; $n = 100$, $\sigma_a = 0.2$

ϕ	$\mu = 0$	$\mu = 0.5$
0.0	0.0548	0.9988
0.1	0.0718	0.9950
0.3	0.0780	0.9406
0.5	0.0788	0.7216
0.7	0.0970	0.3962
0.9	0.1656	0.2010

Table 2.2: Power of the test $\mu = 0$ in an AR(p) setting. Here we consider sample size to be 100 and the standard deviation for error term to be 0.2. The test is done at 0.05 level of significance. First column represents type-I error and second column represents power of the test.

To overcome lack of power of the test in such phenomenon, we can resort to parametric bootstrapping procedure. This simple method can be outlined in the following few steps:

Step 1 Find the region means using MODWT procedure and then find $e_{ti} = y_{ti} - \hat{y}_i$.

Step 2 Select the AR order p.

Step 3 Estimate the parameters and innovation variance from the model selected in step 2.

Step 4 Simulate a mean-zero stationary Gaussian AR(p) time series, say e^*, with parameters $\hat{\phi}$ and innovation variance $\hat{\sigma}$ found in step 3. For null model $\mu = 0$, and so $y = e$. Do the simulation procedure large number of times, say $B = 10^4$ times.

Step 5 Find the means for each simulated series in all regions, $\bar{y}^*_{\gamma_1}, \bar{y}^*_{\gamma_2}, \ldots, \bar{y}^*_{\gamma_k}$, where the superscript *denotes the bootstrap sample. The p-value for region i is defined as, $p_i = \#\{\bar{y}^*_{\gamma_i} \geq \bar{y}_{\gamma_1}\}/B$

In the presence of large series, we can find the order of the AR(p) process from the series using BIC criterion.

The simulation study, presented in Tables 2.3-2.5, suggests that the bootstrapping method works well for testing mean in large series. The *False Positive Rate* (FPR) of the test is still high for large ϕ and short series. Nonetheless, this test procedure works better than the previously mentioned one.

Bootstrapping power comparison; $n = 50$, $\sigma_a = 0.2$

ϕ	$\mu = 0$	$\mu = 0.5$
0.0	0.064	1.00
0.1	0.064	1.00
0.3	0.066	1.00
0.5	0.08	1.00
0.7	0.118	0.998
0.9	0.244	0.752

Table 2.3: Power of the bootstrap method for testing $\mu = 0$ in AR(1) process for different values of ϕ. Here series length is 50 and $\sigma_a = 0.2$. For any value of σ, FPR is very high in this case.

Bootstrapping power comparison; $n = 100$, $\sigma_a = 0.2$

ϕ	$\mu = 0$	$\mu = 0.5$
0.0	0.054	1.00
0.1	0.056	1.00
0.3	0.064	1.00
0.5	0.062	1.00
0.7	0.072	1.00
0.9	0.134	0.83

Table 2.4: Power of the test $\mu = 0$ using bootstrap method for different values of ϕ. The AR(1) series has length 100 and $\sigma_a = 0.2$.

Bootstrapping power comparison; $n = 200$, $\sigma_a = 0.2$

ϕ	$\mu = 0$	$\mu = 0.5$
0.0	0.048	1.00
0.1	0.048	1.00
0.3	0.052	1.00
0.5	0.054	1.00
0.7	0.060	1.00
0.9	0.124	0.996

Table 2.5: Power of the bootstrap method for testing $\mu = 0$ in an AR(1) process for different value of ϕ. Here series length is 200 and the standard deviation for error term is $\sigma_a = 0.2$.

If there is only one region present in the study, the decision about the test can be done using this obtained p-value. However, in a GCH data analysis there will be several gain and loss regions and so the overall decision depends on multiple test method. Having obtained the p-values for all regions using the aforementioned bootstrap procedure, we need to calculate the multiple test values using some standard method. Benjamini and Hochberg (1995) proposed a method for multiple testing using *False Discovery Rate* (FDR). Another more recent approach, called q-value, was proposed by Storey (2002). To deal with multiple testing, Pounds *et al.* (2004) introduced spacings LOESS histogram, or SPLOSH. This aims at estimating conditional FDR

which is the expected proportion of false positives given we have r significant features. In the genome wide study of testing periodicity which was discussed in Chapter 1, SPLOSH revealed to be most conservative while q-value approach seems to be liberal in detecting the correct number of periodic genes. However, unlike the number of genes, the number of jump points or the number of regions will not be even hundreds. So it would be expected that all these methods would produce similar results in this simulation.

2.2.5 Determination of Gains and Losses

Assume that the relative copy number is a smooth function $f(k)$, where k denotes the position of the clone on the gene. To find the change points of $f(k)$, we can determine abrupt change of the function $f(k)$ through wavelet coefficients. The test threshold is calculated using the universal threshold $\sigma\sqrt{2\log(n)}$. Any wavelet coefficient that exceeds the point are specified as the position of abnormal change in DNA copy numbers. Once we specify distinct regions using the threshold, we need to define them as loss, gain or normal region through another preselected threshold T_2.

$$\mathcal{R}_i = \begin{cases} \text{Call gain}, & \text{if } \mathcal{M}_i > T_2 \\ \text{Call loss}, & \text{if } \mathcal{M}_i < -T_2 \\ \text{Call normal}, & \text{if } -T_2 \leq \mathcal{M}_i \leq T_2 \end{cases}$$

where $\mathcal{M}_i$ is the multiple test value of the i-th region. If we would like to call a region to be gain or loss region at a q-value of 0.05, then this is our selected T_2.

2.3 SIMULATED EXAMPLES

Let z_t, $t = \{1, 2, \ldots, n\}$ be the observations along a specific chromosome arm. In this section we present few simulated examples to demonstrate the performance of the proposed method. A comparison of the method with CLAC is provided. A preselected threshold of $q = 0.05$ is used to call a gain or loss region in all the simulated examples and real data.

2.3.1 Example-1: White Noise Series

Data of length 1040 are generated such that $z_t \sim N(0, 0.15^2)$. This means that no loss or gain region is present in the data shown in Figure 2.2. The proposed method, applied to raw data, worked well in providing the true feature of the series.

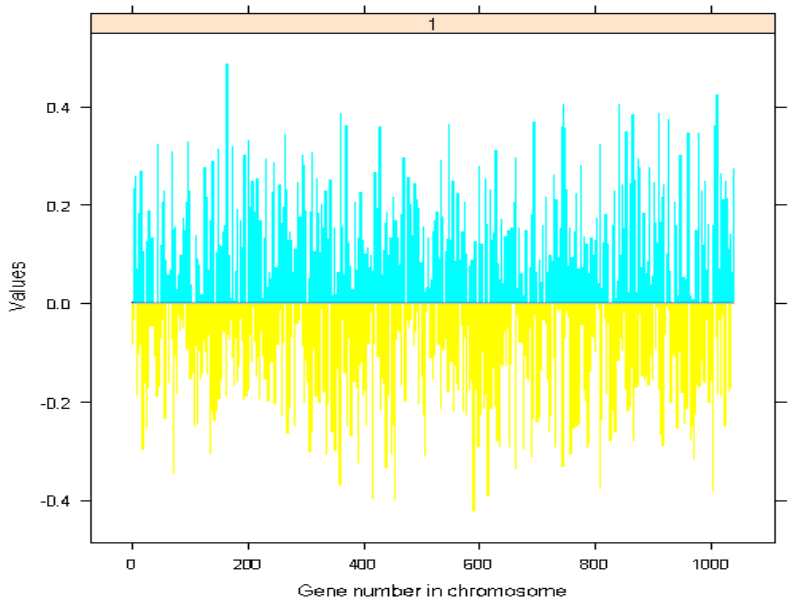

Figure 2.2: White noise series, where there is no jump. The data are simulated from $N(0, 0.15^2)$. The mean value of the region is almost in the zero line.

2.3.2 Example-2: Smooth Signal Plus White Noise

Some data, $n = 200$, was generated by adding random noise to a smooth curve presented in Figure 2.3. That is, the observations follow the relationship $z_t = g(x_t) + \epsilon_t$, where $g(x_t)$ is the smooth part and $\epsilon_t \sim (N(0, \sigma^2)$.

Wavelet method is applied to this data for plausible jump detection. We see from Figure 2.4 that the method is able to detect correctly the absence of any break points.

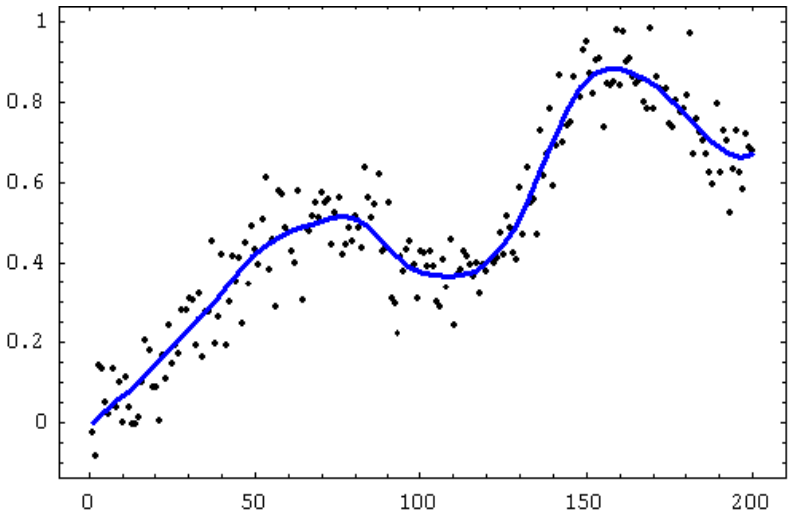

Figure 2.3: Scatter plot of simulated observations obtained by adding random noise to a smooth curve, which is also shown. Apparently there is no sharp jump point in the series.

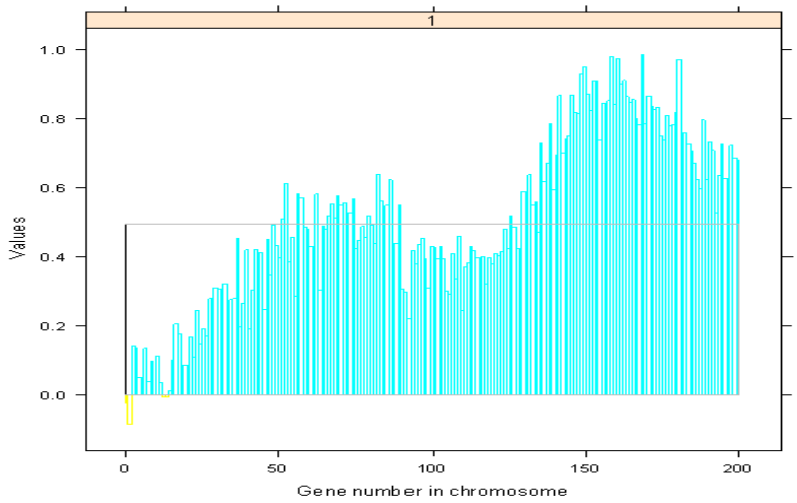

Figure 2.4: Application of wavelet method to the series shown in Figure 2.3. There is only one region, that is no jump was detected in this series. The mean value is given by a line.

2.3.3 Example-3: Two Loss/Gain Regions

Data set with $n = 270$ observations are simulated in two blocks representing two chromosomes. The model in both chromosomes is $z_t = \mu_t + e_t$, where μ_t takes on values 0, 0.7, and -0.7. That is,

$$\mu_{t1} = \begin{cases} 0, & 1 \leq t \leq 80 \\ -0.7, & 81 \leq t \leq 110 \\ 0, & 111 \leq t \leq 150 \end{cases} \quad \text{for chromosome 1}$$

$$\mu_{t2} = \begin{cases} 0, & 1 \leq t \leq 40 \\ -0.7, & 41 \leq t \leq 70 \\ 0, & 71 \leq t \leq 120 \end{cases} \quad \text{for chromosome 2}$$

For each chromosome,

$$e_t = \phi e_{t-1} + a_t, \quad a_t \sim \text{NID}\,(0, \sigma_a^2) \tag{2.6}$$

Since $\text{Var}(e_t) = \sigma_a^2/(1-\phi^2)$, we can write the innovation variance, $\sigma_a^2 = (1-\phi^2)\text{Var}(e_t)$. We consider three cases with ϕ values 0.4, 0.6 and 0.8. Here we do not provide the graphs for case $\phi = 0.6$ as it gives similar result as that for $\phi = 0.4$. Figures 2.5 and 2.7 show that the proposed method detects the jump points at right places in all three cases. CLAC method is applied in all data sets. For the implementation of CLAC method, normal array is generated from AR(1) process with corresponding value of ϕ used in the original data. This method seems to work well with low values of ϕ, as can be seen in Figure 2.6. However, Figure 2.8 indicates that the detection of loss and gain region is not perfect in the presence of high autocorrelation. It should be noted that the performance of the method relies on the selection of normal array.

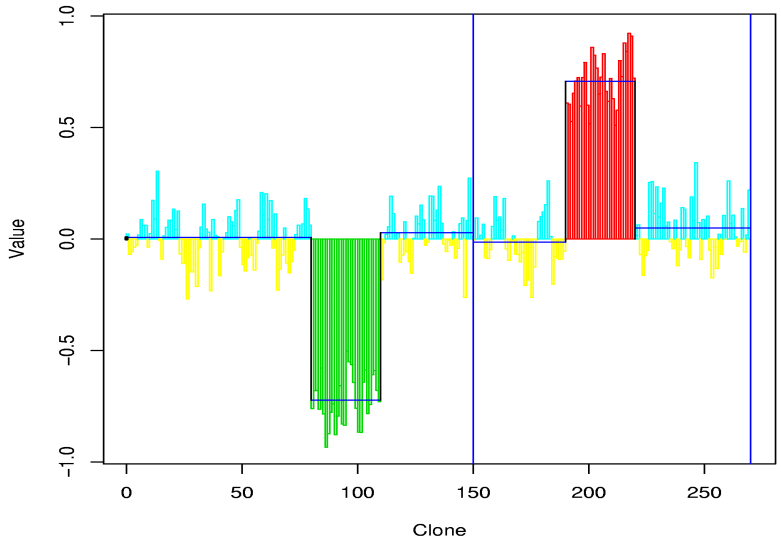

Figure 2.5: Application of wavelet method to the series with error term following AR(1) with $\phi = 0.4$. The method can detect the gain and loss region.

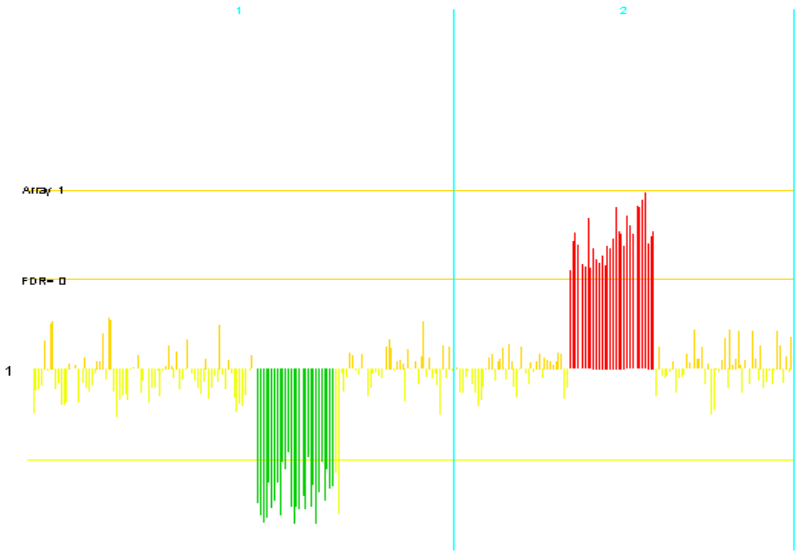

Figure 2.6: Application of CLAC method to the series with error term following AR(1) with $\phi = 0.4$. Gain and loss region is detected at the right places for this value of ϕ.

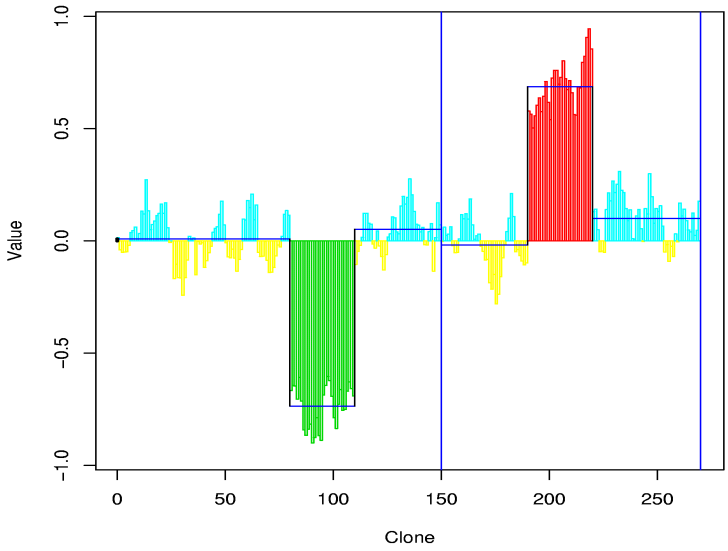

Figure 2.7: Application of wavelet method to the series with error term following AR(1) with $\phi = 0.8$. The method can detect correct gain and loss region.

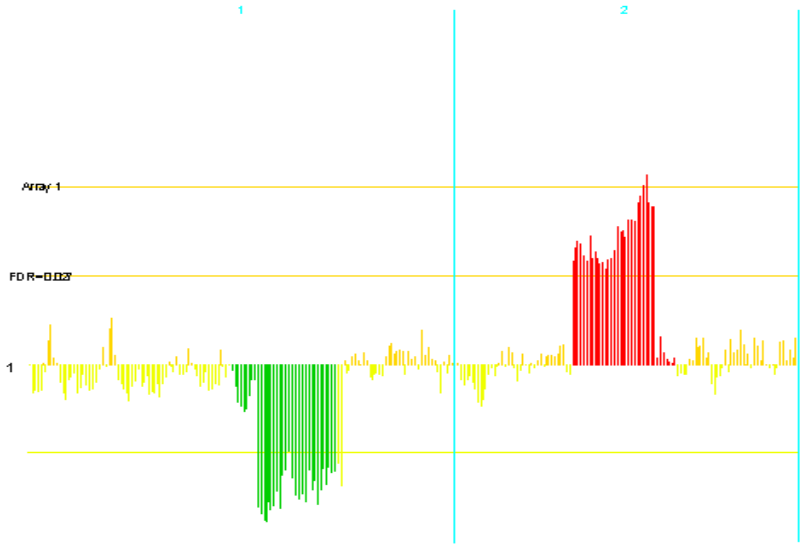

Figure 2.8: Application of CLAC method to the series with error term following $AR(1)$ with $\phi = 0.8$. We do not get exact detection of gain and loss region.

2.3.4 Example-4: Seven Jump Points

The data set consists of 200 observations having 7 jump points at 50, 60, 92, 106, 144, 169 and 181. Error terms are IID normal with mean 0 and standard deviation

0.5. We split the series into two chromosomes where 141 genes are assigned to first one and 59 genes assigned to second one. This is a typical example where there are two successive gain regions within second chromosome. We see from Figure 2.9 that the proposed method can detect the break points exactly and define the loss and gain regions according to the preselected threshold value.

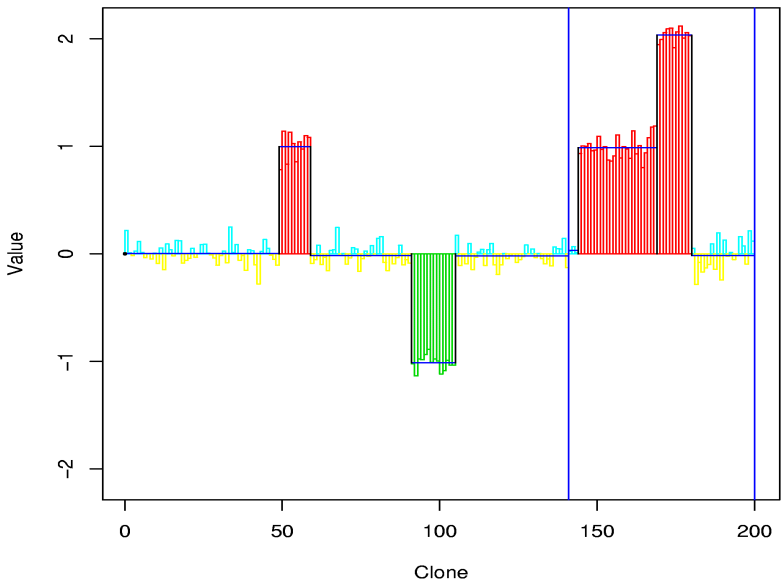

Figure 2.9: A series with seven jump points. Observations are divided into two chromosomes such that first 141 observations are in chromosome 1 and rest 59 observations are assigned to chromosome 2. The proposed method correctly detects the jump points.

2.3.5 Smoothing the Data

Wang (1995) suggested using simple moving average smoothing (MAS) with specific window size before applying the approach. If $\hat{z}$ be the running mean with neighbourhood size k, then the smoothed series would be:

$$\hat{z}_i = \frac{1}{2k+1}(z_{i-k} + z_{i-k+1} + \ldots + z_{i+k}) \tag{2.7}$$

for $i = k+1, k+2, \ldots, n-k$. For the other observations, say for $i = 1, 2, \ldots k$ and $i = n-k+1, n-k+2, \ldots, n$, define $u = \max(1, i-k)$ and $v = \min(n, i+k)$; then

$$\hat{z}_i = \frac{1}{\nu - u + 1}, (z_u + z_{u+1} + \ldots + z_{nu}) \tag{2.8}$$

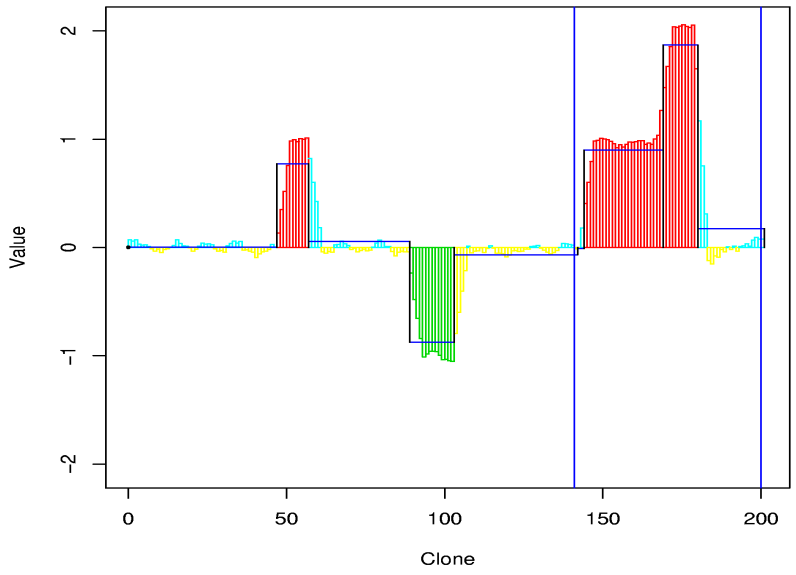

Figure 2.10: Detection of jump points when wavelet method is applied to the data demonstrated in Figure 2.9, but smoothing is done before the analysis. There are shifts in the jump point detection.

Investigations revealed that this smoothing results some shift in the break point detection when wavelet method is applied. For example, the series in Example 4 was smoothed and the wavelet method was applied thereafter. Figure 2.10 shows that the number of regions detected is correct; nevertheless, the detection points are not at the appropriate places.

2.4 APPLICATIONS TO CGH ARRAYS

We apply the proposed method in two real CGH arrays. The method detects several loss and gain regions. A comparison of the method with CLAC is illustrated through the second example.

2.4.1 Application-1

In CGH array, 2400 BAC clones were measured each with three replicates (Snijders *et al.*, 2001). Measurements for log base 2 intensity ratio are provided. Average relative DNA copy number sequences of the three replicates along the genome is shown in Figure 2.11. The figure also demonstrates the gain or loss regions that are detected using this method. As we can see, the measures are mostly along the zero line, which indicates that the test sample has the same DNA copy numbers as that of reference sample.

The log ratios along the genome are considered as a time series sequence. The proposed method is then applied to calculate the wavelet coefficients and to determine the abnormal positions. There are number of loss and gain regions detected by this method. Figure 2.1 in Section 2.1 demonstrates the gain and loss regions detected in first 12 chromosome. Figure 2.12 presents the chromosome-wise abnormal regions for other chromosomes. Red and green colors refer to the gain and loss regions respectively. There are presence of abnormal regions in several chromosomes, namely 1, 5, 7, 8, 9, 11, 14, 17, 20 and 23.

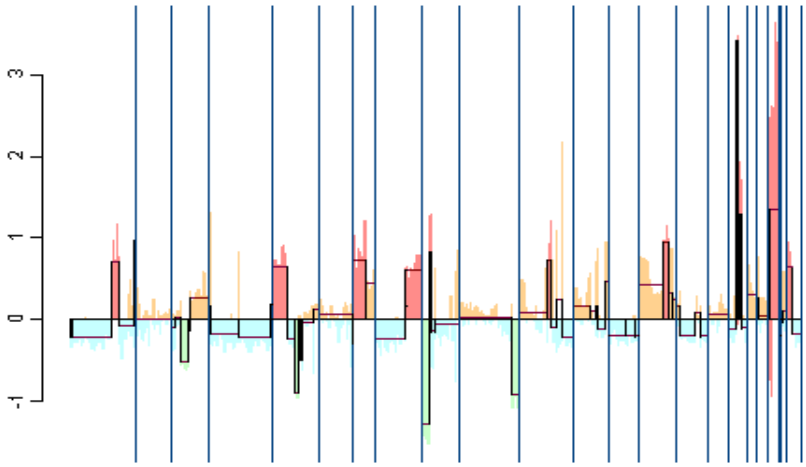

Figure 2.11: Application of wavelet method to CGH data set from Snijders *et al.* (2001). There are many gain/losss regions in the whole genome.

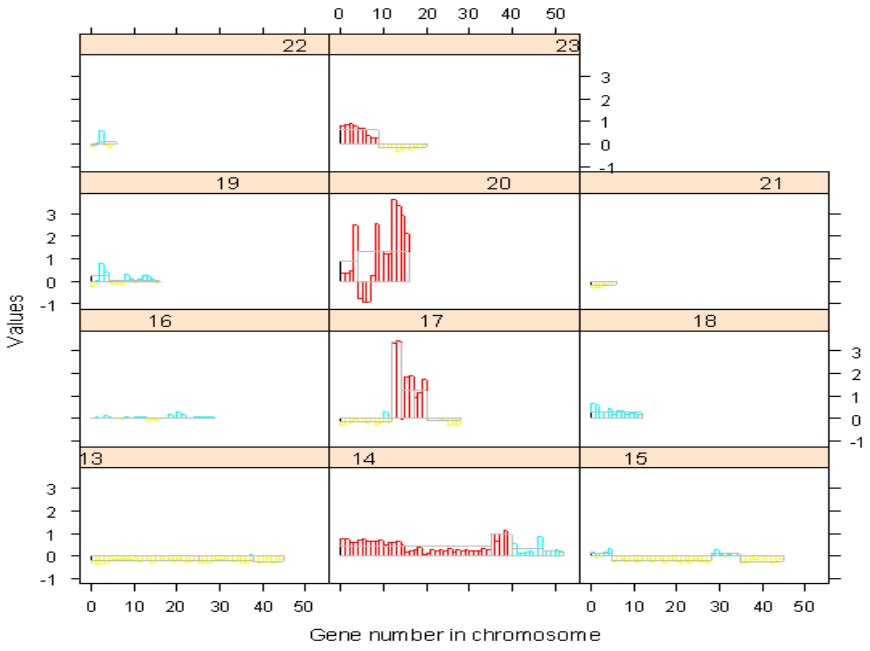

Figure 2.12: Representation of gain/loss regions in last 11 chromosomes. There are presence of abnormal regions in chromosome number 14, 17, 20 and 23.

2.4.2 Application-2

We apply the proposed method to one of the examples found in R library `clac`. The package has data set `BACArray` and the column `DiseaseArray` has 9980 observations containing 4 arrays, one of which is analyzed for comparison. Wavelet method detected two gain regions colored as red in Figure 2.13. Figures 2.14 and 2.14 of individual chromosome explicitly show that the chromosome 18 and 23 are the regions with copy number amplification. One normal array from the `clac` package is picked and then CLAC method is applied to the array. The outcome, presented in Figure 2.16, also indicates that chromosome 18 and 23 refer to the amplified regions for DNA copy number.

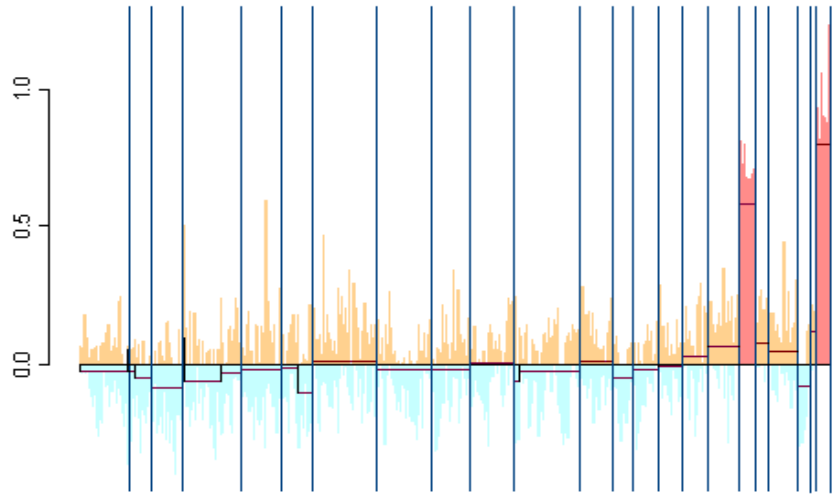

Figure 2.13: Plot of CGH Array taken from R package `clac`. The wavelet method detects only two gain regions in this data set.

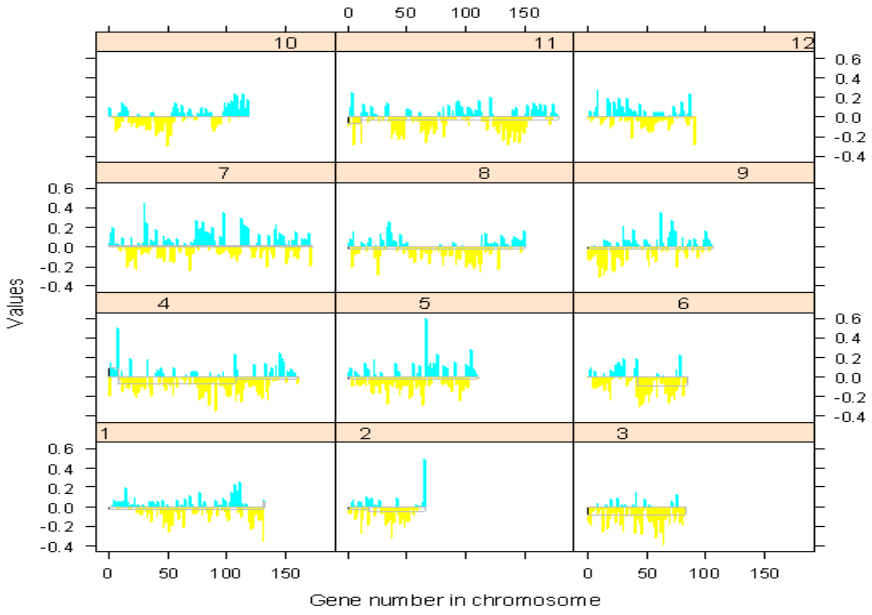

Figure 2.14: Gain or loss regions in first 12 individual chromosomes analyzed from CGH data `BACArray`. There are no abnormal regions present in these chromosomes.

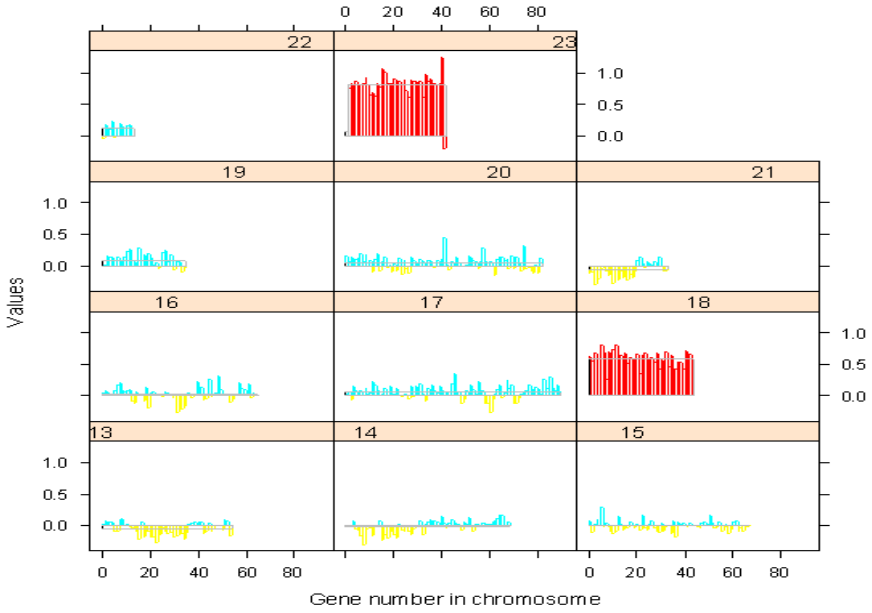

Figure 2.15: Gain or loss regions in 13 to 23 individual chromosomes analyzed from CGH data `BACArray`. Chromosomes 18 and 23 refer to regions of abnormal gain in DNA copy numbers.

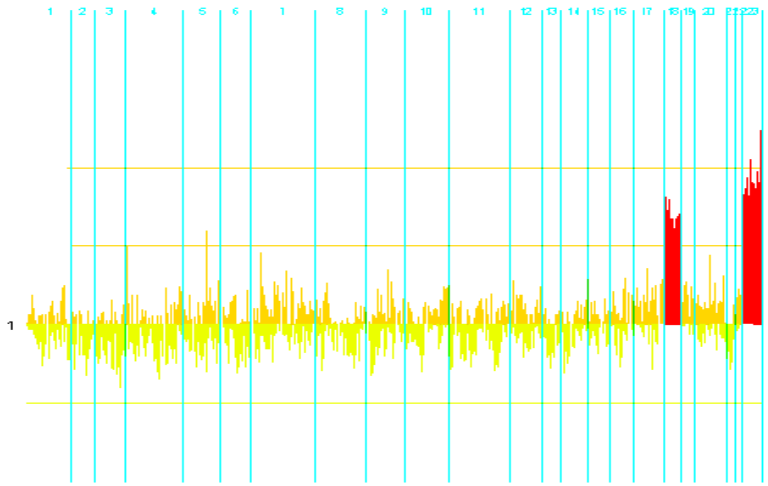

Figure 2.16: Representation of gain/loss regions using CLAC method in `BACArray`. It shows that there are two gain regions in 18th and 23rd chromosomes.

2.5 DISCUSSION

In this chapter we have proposed wavelet method to identify the abnormal DNA copy number positions on genome. Discrete wavelet transform has two limitations; namely dyadic length requirement and sensitivity of the starting of the time series. To overcome such limitations, we use *maximum overlap discrete wavelet transformation* (MODWT). The positions of the break points were detected using Wang's threshold. Calling a region to be gain, loss or normal depends on the selection of another threshold T_2. Through the simulated examples we demonstrate that the method performs quite well in selecting the break points and hence the abnormal regions in a time series sequence. Moreover, the procedure reports several abnormal regions in two real CGH arrays.

CLAC algorithm, proposed by Wang *et al.* (2005), uses some normal array for detecting deletion and amplification regions. Independence and normality of the clones are two strong assumptions; but the procedure of Jong *et al.* (2003) depends on these assumptions. ACF plots of the estimated errors from the fitted model are presented

in Appendix. It is evident from the plots that consideration of IID observations in the sequence would not be realistic. Our propose method does not assume that the observations be IID Through a short simulation example we show how the detection of the change points shifts when a moving average smoothing is used before applying the wavelet method. An R package, `WaveletCGH`, will be made available which implements the wavelet detection methods described in this chapter.

2.6 APPENDIX

Autocorrelation function (ACF) is useful in detecting the presence of correlation among the successive observations. In this study, we observe the residuals by subtracting the mean of any selected region from the observations in that region. That is, $e_t = z_t - \mu_t$ is the residual for t-th clone. ACF plots are presented for the residuals obtained from the application in CGH array described in Section 2.4.

2.6.1 ACF Plot from Application-1

Figure 2.17 is constructed to show the autocorrelation behavior of the error process for each chromosome. It seems that the residuals are not quite IID within each of the chromosome. The residuals in chromosome numbers 1, 8, 10, 14 and 23 demonstrate the presence of strong autocorrelation. This can be a justification to use a simulation study in Example-3 of Section 2.3.

2.6.2 ACF Plot from Application-2

Here the residuals are obtained from CGH array mentioned in Section 2.4.2. The ACF plots in Figure 2.18 indicate the presence of high autocorrelation in residuals for chromosome numbers 1, 4, 7, 8, 9, 10, 11, 13, 14 and 21. Therefore, considering the residuals to be IID would not be realistic in detecting the abnormal regions in this CGH array.

2.6.3 ACF Plot from Normal Array

The normal array described in Section 2.4.2, is analyzed for the presence of autocorrelation in the error term. Figure 2.19 reveals that there is presence of dependence characteristic in residuals within many of the chromosomes; for example, we can note the presence of high autocorrelation in chromosome numbers 1, 4, 7, 8, 9 and 14.

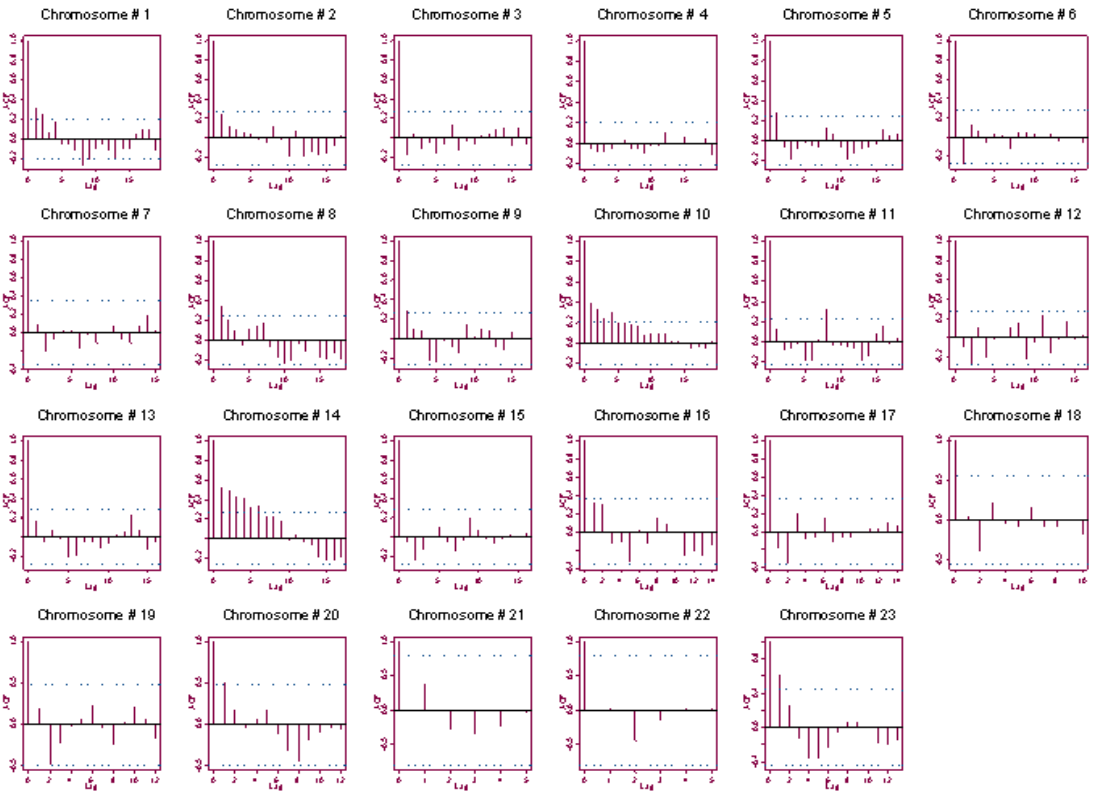

Figure 2.17: ACF plots for the residuals obtained for chromosome 1 to 23 using the data set in subsection 2.4.1. The residuals in few of the chromosomes indicate the presence of high autocorrelation.

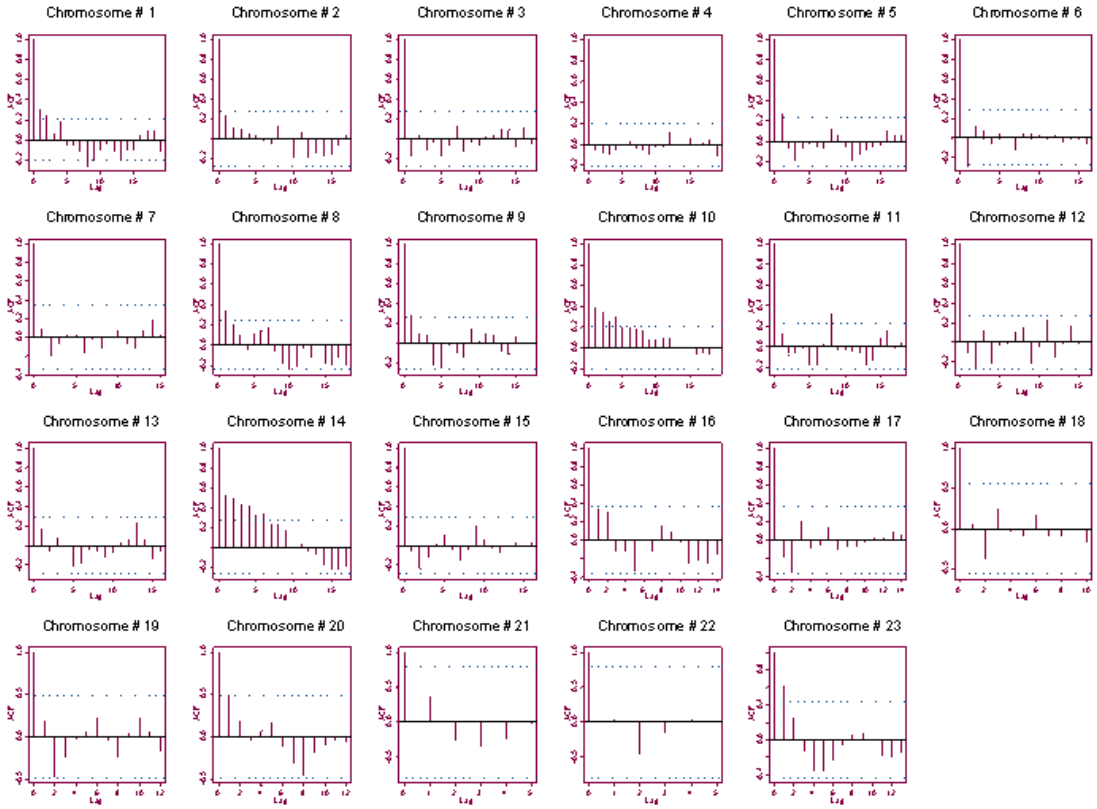

Figure 2.18: ACF plots for the residuals obtained for chromosome 1 to 23 using data set in subsection 2.4.2. The residuals in few of the chromosomes indicate the presence of high autocorrelation.

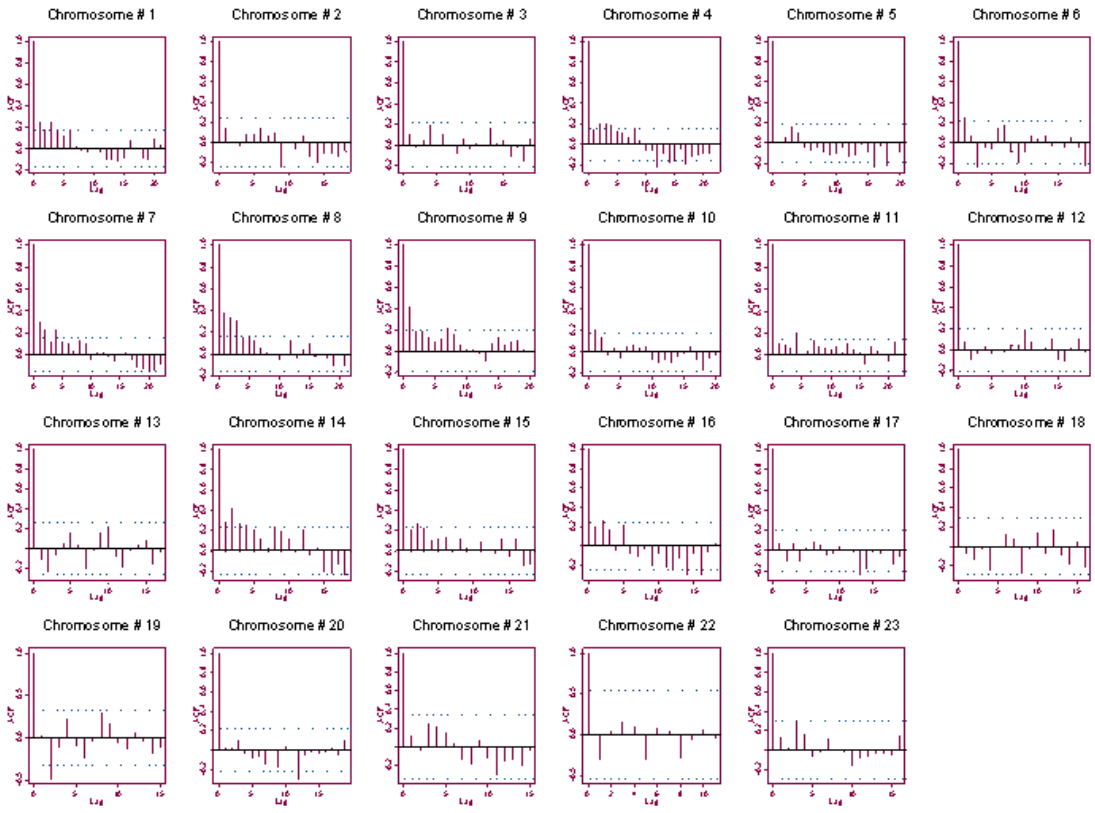

Figure 2.19: ACF plot for the residuals obtained from the normal array described in Section 2.4.2. There exist highly autocorrelated residuals within many chromosomes.

REFERENCES

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, **57**, 289-300.

Clevelend, R. B., Clevelend, W. S., McRae, J. E. and Terpenning, I. (1990) STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *J of Official Statistics*, **6**, 3-73.

Constantine, W. L. B. and Percival, D. B. (2003) S+Wavelets 2.0. *Insightful Corporation*, Seattle, WA.

. Daubechies, I. (1992) Ten Lectures on Wavelets. *Philadelphia: Society for Industrial and Applied Mathematics*.

Fisher, R. A. (1921). Studies in crop variation. I. An examination of the yield of dressed grain from Broadbalk. *J. Agric. Sci.*, **11**: 107-135.

Hodgson, G., Hager, J., Volik, S., Hariono, S., Wernick, M., Moore, D., Nowak, N., Albertson, D., Pinkel, D., Collins, C., Hanahan, D. and Gray, J.W. (2001). Genome scanning with array CGH delineates regional alterations in mouse islet carcinomas. *Nature Genetics* **29**, 491.

Jong, K., Marchiori, E., Vaart, A., Ylstra, B., Weiss, M. and Meijer, G. (2003). Chromosomal breakpoint detection in human cancer. *In LNCS*, (2611), Springer.

Lai, W. R., Johnson, M. D., Kucharlapari, R. and Park, P. J. (2005) Comparative analysis of algorithm for identifying amplications and deletions of rray CGH data. *Bioinformation*, **21**, 3763-3770.

Lingjaerde, O. C., Baumbusch, L. O., Lisestol, K., Glad, I. K. and Borrsen-Dale, A. (2005) CGH Explorer: a program for analysis of array-CGH data. *Bioinformation*, **21**, 821-822.

Matlab (2007). Detecting Discontinuities and Breakdown Points. *In Wavelet Toolbox: Wavelet Applications*.

Percival, D. B. and Walden, A. T. (2000). *Wavelet Methods for Time Series Analysis*. Cambridge University Press.

Pinkel, D. and Albertson, D. G. (2005) Array comparative hybridization and its applications in cancer. *Nature Genetics*, **37**, S11-S17.

Pollack J.R., Sorlie, T., Perou, C. M., Rees, C. A., Jeffry, S. S., Lonning, P. E., Tibshirani, R., Botstein D., Borrsen-Dale, A. and Brown, P. O. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences*, USA, **99**, 1296312968

Pounds, S. and Cheng C. (2004) Improving false discovery rate estimation, *Bioinformatics*, **20**, 1737-1745.

Snijders, A. M., Nowak, N., Segraves, R., Blackwood, S., Brown N., Conroy, J., Hamilton, G., Hindle, A. K, Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J. P., Gray, J. W., Jain, A. N., Pinkel, D. and Albertson, D. (2001)Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics* **29**, 263 - 264.

Storey, J. D.(2002) A direct approach to false discovery rates. *J. R. Stat. Soc. B*, **64**, 479-498.

Wang, Y. (1995). Jump and Sharp Cusp Detection by Wavelets. *Biometrika*, **82**, 385-397.

Wang, P., Kim, Y., Pollack, J., Narasimhan, B. and Tibshirani, R. (2005). Studies in crop variation. I. A method for calling gains and losses in array CGH data. *Biostatistics*, **6**, 1, 4558

Willenbrock, H. and Fridlyand, J. (2005). A comparison study: applying segmentation to array CGH data for downstream analyses. *Bioinformatics*, **21**, 40844091.

Chapter 3

IMPROVED CLASS PREDICTION IN GENE EXPRESSION MICROARRAY DATA

3.1 INTRODUCTION

The advancement of cDNA microarrays and high-density oligoneucleotide chips in biotechnology has drawn much interest of statistical analysis in cancer research. One of the primary areas of focus is the classification of tumors using the gene expression data. A better understanding of the molecular variation among the tumors can be studied thanks to the possibility of simultaneously analyzing thousands of gene expression profiles. However, the fruitful endeavor for this understanding depends on the selection of proper statistical approach.

Dudoit *et al.* (2002) and Simon *et al.* (2004) provided an extensive comparison of different classification methods. For the implementation of most of the methods, there needs to be an initial gene selection to make the number of genes to be less than the number of samples. Although their analyses show that the diagonal *linear discriminant analysis* (DLDA) maintains one of the top-ranking classifiers, the implementation of this method to other data sets does not seem to be appealing (Fort and Lambert-Lacroix, 2005). The big challenge of dealing with the microarray data is that the number of covariates is in thousands whereas the number of samples is usually not more than one hundred. Similar to regression method, the traditional discriminant analysis methods are not efficient in such situation. The method with principal

components, partial least squares, ridge regression with their penalized forms are discussed in several articles as ways to solve the problem of classification (Ghosh, 2003; Fort and Lambert-Lacroix, 2005).

Nearest neighbour algorithm is one of the most frequently used techniques in classification problem. This algorithm is also known as instance-based learning. Holmes and Adams (2003) proposed a method which takes into account multiple nearest neighbors as a set of covariates in contrast to traditional method where only single nearest neighbor is selected on the basis of cross-validation error rate. The authors also proposed that the optimization of k can be done by maximum pseudolikelihood instead of using cross-validation for misclassification rate. In a logistic regression setting, the theory is flexible as it can take the original covariates as well as multiple nearest neighbor covariates (NNC). Original covariates capture the linear effects and multiple NNC capture nonlinear effects present at different scales within the data. The presence of thousands of genes as covariates will lead to a problem in variable selection in their method. This is because the traditional step-wise regression will no longer be feasible in such circumstances. Although the procedure can be reformed in terms of tens of genes selected by some procedure, the performance of classification method depends on initial gene selection process (Lee *et al.*, 2005). Also, may researchers feel it is best to include as many genes as possible and are reluctant to use subset approaches (Guo *et al.*, 2007).

Fort and Lambert-Lacroix (2005) put their suggestion against using k-nearest neighbor method for some of the data sets due to many occurrences of indecision. Still the analysis shows that the performance of this method is much better than many other methods (Dudoit *et al.*, 2002). It was noticed that the presence of high positive correlation of the gene expression observations within the same group and high negative correlation between different groups brings about the nearest neighbor

classifier to perform as a good classifier in several data sets.

The estimation of regularized parameters involved in any model can be performed in several ways. *Bayesian Information Criterion* (BIC) is one of the popular criteria in selecting best model. Subset selection is highly variable as it is a discrete process, which either takes a variable or discard it (Hastie *et al.*, 2001). Tibshirani (1996) proposed *Least Absolute Shrinkage and Selection Operator* (LASSO), that shrinks some regression coefficients and sets other to zero, and thus works as a variable selection method. The $L1$ lasso penalty can be used in logistic regression framework when we have quite a large number of covariates. However, we found that the misclassification rate gets higher when all the nearest neighbor covariates are included in variable selection stage.

Nguyen and Rocke (2002) used partial least squares method for the purpose of classification in gene expression data. Recent methods include *Support Vector Machine* (SVM) and *Shrunken Centroid Regularized Discriminant Analysis* (SCRDA) (Hastie *et al.*, 2001; Guo *et al.*, 2007). SVM works in classification by producing linear boundary in the feature space and thus refers to non-linear boundary in input space. SCRDA is an extension of Fisher Linear Discriminant Analysis. This solves the non-singularity problem and provides a gene selection during the process.

Including NNC prior to running any of the method gives an augmented form. This provides some extra information to the classifier. First NNC can lead to capture non-linear relationship which might be ignored otherwise. Thus an improved version of many sophisticated methods can be achieved using this kind of augmentation.

3.2 METHODS

Suppose that we have expression levels for p genes over a size of n samples. The data matrix is given by $X = (x_{ij})$, a matrix of dimension $n \times p$. The value x_{ij} refers to the expression level for j-th gene in i-th sample. The response variable is a categorical variable taking values as $y_i = \{A_1, A_2, \ldots, A_g\}$, where g is the number of classes. In the present work we discuss only two-class prediction problem. Hence we can express y_i as taking values $\{-1, 1\}$. Predictions are built on the training set and the performances are evaluated using the test set. In an one-leave-out validation process, successively all but one observations are considered as training set and the error rate is measured.

3.2.1 K-Nearest Neighbor

The nearest neighbor method is based on the distance function; for example, correlation or Euclidian distance for pairs of observations. In a k-nearest neighbor method, predictions of new observations are made through the training set $\{y_i, x_i\}$ for $i = 1, 2, \ldots, n$. For a new observation, we find the k closest observations in the training set and then predict the class to be the one where the majority of the k-neighbours belong to. The process is run for each specified values of k and then the selection of k is done using cross-validation. However, Holmes and Adams (2003) proposed a new method for finding optimum value of k. Instead of using cross-validation method, optimum value of k is derived by maximizing pseudolikelihood from a logistic regression.

After the initial selection of a number of genes, say P, we have our set of variables as $\{x_1, x_2, \ldots, x_p\}$. Corresponding to the i-th observation, *k-nearest neighbor*

autocovariate (NNC) is defined as:

$$\nu_{i(k)}(A_1) = \frac{1}{k}\sum_{j\sim i}[I(y_j = A_1) - I(y_j = A_0)] \tag{3.1}$$

The indicator variable $I(x = \omega)$ takes the value 1 if $x = \omega$ and 0 otherwise; $\sum_{j\sim i}$ denotes that the summation is over the k-nearest neighbors of x_i in the set $x_1, x_2, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p$ The autocovariate $\nu_{i(k)}$ refers to the proportion of class A_1's to class A_0's within the k nearest neighbors of x_i. Therefore, if all the k nearest neighbors of x_i are in A_1 then the autocovariate $\nu_{i(k)}$ is 1; if all the k nearest neighbors of x_i are in A_0 then the autocovariate $\nu_{i(k)}$ is 0. Then a logistic regression model containing the covariates $\nu_{i(k)}$ can be written as

$$\Pr(y_i = A_1) = \eta_i = \frac{\exp(\alpha_k \nu_i(k))}{1 + \exp(\alpha_k \nu_i(k))} \tag{3.2}$$

The pseudolikelihood function is therefore,

$$L(\alpha_k; \nu_{(k)}) = \prod_{i=1}^{n} \eta_i^{\tilde{y}_i}(1-\eta)^{1-\tilde{y}_i} \tag{3.3}$$

where

$$\tilde{y}_i = \begin{cases} 0, & \text{if } y_i = A_0 \\ 1, & \text{if } y_i = A_1 \end{cases} \tag{3.4}$$

Optimal value of k is selected by maximizing the likelihood function. That is,

$$\hat{k} = \text{argmax}_k L(\alpha_k; \nu_{(k)}) \tag{3.5}$$

3.2.2 Efficient NNC Computation in R

We can use R package `class` to compute the nearest neighbour covariates. For calculating K-th NNC corresponding to test set, built-in function `knn()` can be used

to obtain predicted class ($\hat{y}$) and proportion (p) of votes for that winning class. Then $p \times K$ measures the number of votes for the winning class. We can define a proportion, called *consensus proportion* (C), as $(pK - (K - pK))/K$. Then K-th NNC would be $\hat{y} \times C$. To find the NNC corresponding to training set, we can consider successively each observation as test sample and do the above steps. The function `knn.cv()` performs the procedure automatically.

3.2.3 DLDA and DQDA

Let $f_l(x)$ be the conditional density of $\mathtt{x}$ in class $y = A_l$ and assume that this follows multivariate normal distribution of the form:

$$\mathtt{x}|y = A_l \sim \text{MVN}\,(\mu_l, \Sigma_l)$$

Let π_l be the prior probability of class l. Then the discriminant function is expressed as

$$Ł_l(x) = -\frac{1}{2}\log|\Sigma_l| - \frac{1}{2}(x - \mu_l)^T\Sigma_l^{-1}(x - \mu_l) + \log\pi_l \tag{3.6}$$

This is called quadratic discriminant function as it does not assume equal covariances throughout the classes. In a two class setting, the decision boundary between two classes can be given by a quadratic equation $\{x : Ł_1(x) = Ł_2(x)\}$. If the class density has diagonal covariance matrix of the form $\Sigma_l = \text{diag}\,(\sigma_{l1}^2, \sigma_{l2}^2, \ldots, \sigma_{lP}^2)$, then the discrimination rule is called *diagonal quadratic discriminant analysis* (DQDA).

If we assume that class density has same covariance matrix for all the classes; that is if $\hat{\Sigma}_l = \hat{\Sigma}$, this leads to *linear discriminant analysis* (LDA). When covariance matrix in LDA is diagonal of the form $\Sigma = \text{diag}\,(\sigma_1^2, \sigma_2^2, \ldots, \sigma_P^2)$, then this is called *diagonal linear discriminant analysis* (DLDA).

We predict an observed value x_0 to a class which maximizes the discriminant function in Equation 3.6; that is, $y(x) = \text{argmax}_l Ł_l(x)$.

3.2.4 Shrunken Centroid RDA

Shrunken Centroid Regularized Discriminant Analysis (SCRDA) was introduced by Guo *et al.* (2007). This is a modified version of LDA. After estimating the parameters, we can write the discriminant function from equation 3.6 as:

$$Ł_l(x) = x^T \hat{\Sigma}^{-1} \bar{x}_l - \frac{1}{2} \bar{x}_l^T \hat{\Sigma}^{-1} \bar{x}_l + \log \pi_l \tag{3.7}$$

where $\bar{x}_l$ represents the mean vector in l-th class. In high-dimensional setting, the estimates in LDA will be unstable and therefore cannot provide optimal results (Guo *et al.*, 2007). In order to overcome the singularity problem in such situation, the authors proposed using regularized form of the covariance estimate:

$$\tilde{\Sigma} = \alpha \hat{\Sigma} + (1 - \alpha) I_P \tag{3.8}$$

where α is a non-negative value in the range $0 \leq \alpha \leq 1$. Using the equation 3.8, we can redefine the discriminant function as:

$$\tilde{Ł}_l(x) = x^T \tilde{\Sigma}^{-1} \bar{x}_l - \frac{1}{2} \bar{x}_l^T \tilde{\Sigma}^{-1} \bar{x}_l + \log \pi_l \tag{3.9}$$

Then the SCRDA can be constructed as classifying an observation x in a group that minimizes:

$$(x - \bar{x}'_{l'})^T \tilde{\Sigma}^{-1} (x - \bar{x}'_{l'}) - \log \pi_{l'} \tag{3.10}$$

where $\bar{x}'_{l'}$ is the vector of shrunken centroid for group l. A shrunken centroid $\bar{x}'$ is defined as

$$\bar{x}' = \text{sgn}\,(\bar{x})(|\bar{x} - \Delta)_+ \tag{3.11}$$

where $\Delta > 0$ is shrinkage parameter.

To estimate the tuning parameter pair (α, Δ), we use the Min-Min rule in the analysis. The first step is to find all the pairs that yield minimum cross-validation error in training set. Finally, optimum pair of (α, Δ) refers to that values which correspond to minimum number of selected genes.

3.2.5 Support Vector Machine

The space that $\mathtt{x} = \{x_1, x_2, \ldots, x_p\}$ takes is called input space. A space obtained after transforming $\mathtt{x}$ to $\tau(\mathtt{x})$ is called feature space. An SVM is a technique that separates classes through non-linear boundary by creating linear boundary in transformed feature space.

Let $u'x + a = 0$ is the separating hyperplane between the groups. There exists two other bounds - the distance between which is sought to be maximum for separating the classes. This distance is called margin and denoted as $m = 1/||u||$. We can define the decision boundary through the optimization problem

$$\text{minimize } \tfrac{1}{2}||u||^2$$
$$\text{subject to } y_i(u'x + a) \geq 1 \text{ or } 1 - y_i(u'x + a) \leq 0$$

The Lagrangian is

$$L = \frac{1}{2}u'u + \sum_{i=1}^{n} \alpha_i(1 - y_i(u'x_i + a)) \tag{3.12}$$

where α_i is Lagrange multiplier. Setting gradient of L w.r.t u and a to zero and then substituting $u = \sum_{i=1}^{n} \alpha_i y_i x_i$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$, we get

$$L = -\frac{1}{2} \sum \sum \alpha_i \alpha_j y_i y_j x_i' x_j + \sum \alpha_i \tag{3.13}$$

Therefore, the optimization problem becomes

$$u(\alpha) = -\tfrac{1}{2} \sum \sum \alpha_i \alpha_j y_i y_j x_i' x_j + \sum \alpha_i$$
$$\text{subject to } \alpha_i \geq 0 \text{ and } \sum \alpha_i y_i = 0$$

However, if the classes overlap in feature space, there will arise some non-negative slack variables $\xi = \{\xi_1, \xi_2, \ldots, \xi_n\}$. Thus, we get modified optimization problem as

$$\text{minimize } \tfrac{1}{2}||u||^2 + c \sum_{i=1}^{n} \xi_i$$
$$\text{subject to } y_i(u' x_i + a) \geq 1 - \xi_i,\ \xi_i \geq 0$$

where c is tradeoff parameter between error and margin. This corresponds to

$$u(\alpha) = -\tfrac{1}{2} \sum \sum \alpha_i \alpha_j y_i y_j x_i' x_j + \sum \alpha_i$$
$$\text{subject to } c \geq \alpha_i \geq 0,\ \sum \alpha_i y_i = 0$$

As mentioned before, linear operation in the feature space is equivalent to non-linear operation in input space. Thus we reach to another SVM optimization problem through substituting the inner product $x_i' x_j$ by

$$K(x_i, x_j) = \tau(x_i)' \tau(x_j) \tag{3.14}$$

There are different types of kernals for SVM optimization; however the popular ones (Hastie *et al.*, 2001) are:

- Radial basis function kernel with width σ:

  $K(x_i, x_j) = \exp(-||x_i - x_j||^2 / 2\sigma^2)$

- Polynomial kernal with degree l: $K(x_i, x_j) = (1 + \langle x_i, x_j \rangle)^l$
- Neural network: $K(x_i, x_j) = \tanh(\kappa_1 \langle x_i, x_j \rangle + \kappa_2)$

3.3 IMPLEMENTATION

For each nearest neighbour $\{K_1, K_2, \ldots, K_l\}$, we can obtain the covariates as $\nu_{(K_1)}$, $\nu_{(K_2)}$, ... , $\nu_{(K_l)}$. Then the augmented set of inputs would be $\{x_1, x_2, \ldots, x_p, \nu_{K_1}, \nu_{K_2}, \ldots, \nu_{K_l}\}$. After inclusion of unit column vector, the design matrix is of the form $D = (1, X, V)$. We found that implementation of all the covariates in V lead to high misclassification rate. In such case, any method picks some unnecessary covariates that deters the optimization of the classification rate. Practical implementation reveals that 1-NN can provide good result in bioinformatic applications. In present work, we investigate the performance of four methods; namely DLDA, DQDA, SVM and SCRDA when first NNC is added to the original set of inputs.

3.3.1 Assessing Prediction Accuracy

Cross-validation is a simple but widely used method for assessing prediction accuracy. In a K-fold cross-validation, we randomly divide the data into K segments. We leave one part out, say j-th part, and fit the model for the remaining parts. Then estimate the error rate for that j-th part. We repeat the process for each of K segments, and finally find the overall misclassification error. In our analysis, we use $K = N$ which leads to *leave-one-out* (LOO) cross validation. We also perform re-randomization analysis. In this case we randomly divide the data into learning and validation part. The size of the validation part is considered as one fifth of the total sample size. A model is tuned from the learning part and prediction error is estimated from the validation part. We repeat the process for 300 times and find overall error rate.

3.3.2 Computation

We use Beowulf cluster computing environment with 58 nodes for doing all the analyses. Yu (2002) developed the package `Rmpi`, which is an interface to *Message Passing Interface* (MPI). This package allows to implement R codes cooperatively in parallel across multiple machines. Some of the microarray data sets are very large and so running the leave-one-out or re-randomization procedure demands lots of computation time. We enjoy very good computational savings using this Beowulf cluster computing facility.

3.4 SIMULATION RESULT

To discuss the motivation of proposed method, we use a simulated data set. The concept of this simulation is similar to what was discussed by Guo *et al.* (2007) as two-group dependent structure. We assume that the conditional densities of `x` in two classes are $\text{MVN}(\mu_1, \Sigma_1)$ and $\text{MVN}(\mu_1, \Sigma_1)$. There are $P = 2000$ input variables. The mean, μ_1, for first group is a $P \times 1$ vector of elements 0. The mean vector in another group has first 100 elements as 0.5 and rest 1900 as 0. The covariance for both groups is block diagonal but with different block sizes. Both the densities have covariance structure as:

$$\begin{pmatrix} \Sigma_\rho & 0 & 0 & \cdots & \cdots & \cdots \\ 0 & \Sigma_{-\rho} & 0 & 0 & \cdots & \vdots \\ 0 & 0 & \Sigma_\rho & 0 & \cdots & \vdots \\ \vdots & 0 & 0 & \Sigma_{-\rho} & 0 & \vdots \\ \vdots & \vdots & \vdots & 0 & \ddots & \vdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \end{pmatrix} \tag{3.15}$$

Each block in the covariance matrix has autoregressive form. If ρ is the autocorrelation between successive genes and the block size is B, then Σ_ρ can be written as:

$$\Sigma_\rho = \begin{pmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{B-1} \\ \rho & 1 & \rho & \cdots & \rho^{B-2} \\ \rho^2 & \rho & 1 & \ldots & \rho^{B-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{B-1} & \rho^{B-2} & \ldots & \ldots & 1 \end{pmatrix} \tag{3.16}$$

We consider same autocorrelation, $\rho = 0.9$, for both groups; but take different block sizes $B = 40$ and $B = 100$. Training set contains 100 observations from each class. To evaluate the performance, 500 test samples are generated from each group using same procedure.

Performance of different methods with and without augmented covariates

Method	$k = 0$	$k = 1$
DLDA	27.8	21.5
DQDA	28.9	24.7
SVM	19.3	16.7
SCRDA	25.8	11.1

Table 3.1: Misclassification rate for different methods in simulated data. All the rates are measured in percentage. Here $k = 0$ refers to original set of covariates, and $k = 1$ refers to one NNC augmented to the original set. A total of 200 training samples and 1000 test samples, measuring $P = 1000$ variables, are generated.

We see from Table 3.1 that all methods are hugely improved through the use of NNC. SCRDA gained the most improvement as the error rate decreased from 25.8% to only 11.1%. This augmentation turned SCRDA to be the best performing method in this data set. The gain in SVM is minimum.

3.5 MICROARRAY DATA SETS

We assess the proposed method using four publicly available data sets. The data sets are (i) colon cancer data (Alon *et al.*, 1999), (ii) acute leukemia data (Golub *et al.*, 1999), (iii) prostate cancer data (Singh *et al.*, 2002) and (iv) breast cancer data (vant Veer it et al., 2002). An overview of the data sets is given in Table 3.2. All of the data sets were either originally divided into groups of training and test sets, or by the aforementioned authors. However, for an extensive comparison we merge all the training and test samples and thereafter find leave-one-out as well as re-sampling error rates.

Summary of the microarray data sets used in the analysis.

Name	Description	P	n_1	n_2
Alon	Colon cancer	2000	40	22
Golub	Acute leukemia	7129	47	25
Singh	Prostate cancer	12600	59	77
Veer	Breast cancer	24188	51	46

Table 3.2: Summary table of four data sets that we analyze to evaluate the performance of proposed method. P refers to the number of genes in corresponding data. n_1 and n_2 are the number of samples available for class 1 and 2 respectively.

3.5.1 Colon Cancer Data

This data set contains 62 tissue samples with 40 tumor and 22 normal samples (Alon *et al.*, 1999). An Affymetrix oligonucleotide array complementary to more than $6,500$ human genes was used to analyze expression levels for these samples. Finally 2000 genes are finally included in the data, which are not readily preprocessed. We follow the pre-processing steps mentioned by Dudoit *et al.* (2002):

- thresholding at floor of 100 and ceiling of 16000,

- filtering to exclude the genes with $\max / \min \leq 5$ and $(\max - \min) \leq 500$
- transformation using logarithm of base 10.

3.5.2 Acute Leukemia Data

Acute leukemia data set contains 72 bone marrow samples obtained from adults with acute leukemia (Golub *et al.*, 1999). Expression levels for 7129 genes are measured using Affymetrix high-density oligonucleotide arrays. There are 47 samples of acute lymphoblastic leukemia (ALL) and 25 samples of myeloid leukemia (AML). The data is not preprocessed and so same procedure as that of Colon data is applied here.

3.5.3 Prostate Cancer Data

In this data set total of 12600 gene expression levels are measured for 136 tissue samples (Singh *et al.*, 2002). Expression profiles were derived from 77 prostate tumors and 57 nontumor prostate samples from patients undergoing surgery. The objective here is to separate tumor tissues from normal tissues. The pre-processing steps mentioned by Singh *et al.* (2002) are applied to the data set (Fort and Lambert-Lacroix, 2005):

- thresholding at floor of 10 and ceiling of 16000,
- filtering to exclude the genes with $\max / \min \leq 5$ and $(\max - \min) \leq 50$.
- transformation using logarithm of base 10 is used.

3.5.4 Breast Cancer Data

The data contains 24188 expression profiles for 97 breast cancer patients. They are divided into two groups - (i) who developed metastases within 5 years and (ii) who remained disease-free within 5 years (vant Veer it et al., 2002). 46 patients developed

distant metastases and 51 did not. The objective is to predict the presence of subclinical metastases in order to provide a strategy to select patients who would benefit from adjuvant therapy. The data set is preprocessed and so no further preprocessing step is applied.

3.6 GENE SELECTION

Selecting a subset of best differential genes provides better classification result for different methods in some microarray data sets (Fort and Lambert-Lacroix, 2005). Lee *et al.* (2005) compared different classification methods for three different types of initial gene selection. It was showed that process of initial gene selection makes difference in the performance. We use a criterion that is based on ratio of between to within group sum of squares of the genes (Dudoit *et al.*, 2002). The ratio for gene j is

$$\frac{\mathrm{BSS}\,(j)}{\mathrm{WSS}\,(j)} = \frac{\sum_i \sum_l I(y_i = l)(x_{lj} - \bar{x}_{.j})^2}{\sum_i \sum_l I(y_i = l)(x_{ij} - \bar{x}_{lj})^2}$$

where $\bar{x}_{.j}$ is the average expression level of gene j across all samples and $\bar{x}_{1j}$ is the average expression level of gene j across samples in class l. A selection of P genes are made by considering the genes having largest BSS /WSS ratios. Although SCRDA can automatically select the genes during the process, we use BSS /WSS criterion to select primarily 500, 1000 and 2000 genes for comparison with other methods.

Performance of different methods for $P = 500$

		LOO		OS	
	Data set	$k = 0$	$k = 1$	$k = 0$	$k = 1$
DLDA	Alon	12.90	12.90	13.83	13.83
	Golub	2.78	1.39	2.80	2.40
	Singh	24.26	22.79	24.74	23.81
	Veer	35.05	35.05	32.82	32.82
DQDA	Alon	12.90	12.90	13.63	13.57
	Golub	1.39	1.39	1.80	1.77
	Singh	34.56	35.29	33.75	33.35
	Veer	30.93	30.93	29.94	29.84
SVM	Alon	14.51	14.51	13.97	13.97
	Golub	1.39	1.39	1.49	1.49
	Singh	5.88	5.88	6.32	6.20
	Veer	35.05	35.05	31.66	31.58
SCRDA	Alon	12.90	11.29	14.63	14.23
	Golub	5.56	5.56	6.37	6.01
	Singh	5.88	5.88	5.88	5.55
	Veer	35.57	29.47	32.89	31.84

Table 3.3: *Leave-one-out* (LOO) and *out of sample* (OS) misclassification rates (in %) of different methods with and without the augmented nearest neighbour covariates (NNC). Here $k = 0$ refers to no NNC and $k = 1$ refers to first NNC included in the initial covariate set. A selection of best 500 genes was made for all data sets.

Performance of different methods for $P = 1000$

		LOO		OS	
	Data set	$k = 0$	$k = 1$	$k = 0$	$k = 1$
DLDA	Alon	12.90	12.90	13.72	13.66
	Golub	4.17	1.39	2.81	2.28
	Singh	29.41	28.68	28.43	27.83
	Veer	32.99	32.99	32.14	32.07
DQDA	Alon	12.90	12.90	14.25	14.13
	Golub	1.39	1.39	1.95	1.90
	Singh	36.76	36.76	36.17	36.03
	Veer	31.96	30.93	29.04	28.56
SVM	Alon	12.90	12.90	14.72	14.72
	Golub	1.39	1.39	1.59	1.59
	Singh	5.88	5.88	7.22	7.07
	Veer	30.93	30.93	31.17	31.14
SCRDA	Alon	12.90	9.68	13.87	13.67
	Golub	6.94	2.78	6.03	5.53
	Singh	8.38	5.15	6.01	5.87
	Veer	33.84	33.60	30.88	30.41

Table 3.4: *Leave-one-out* (LOO) and *out of sample* (OS) misclassification rates (in %) of different methods with and without the augmented nearest neighbour covariates (NNC). Here $k = 0$ refers to no NNC and $k = 1$ refers to first NNC included in the initial covariate set. A selection of best 1000 genes was made for all data sets.

Performance of different methods for $P = 2000$

		LOO		OS	
	Data set	$k = 0$	$k = 1$	$k = 0$	$k = 1$
DLDA	Alon	12.90	12.90	13.96	13.90
	Golub	2.78	1.39	2.91	2.34
	Singh	29.41	23.53	28.56	27.96
	Veer	30.93	30.93	32.40	32.32
DQDA	Alon	14.52	14.52	14.33	14.23
	Golub	1.39	1.39	1.94	1.88
	Singh	36.76	35.29	36.43	36.27
	Veer	29.90	27.84	29.28	29.26
SVM	Alon	11.29	11.29	14.70	14.70
	Golub	1.39	1.39	1.57	1.57
	Singh	6.62	5.89	7.08	6.95
	Veer	28.87	29.90	31.30	31.26
SCRDA	Alon	12.90	12.90	13.73	13.79
	Golub	5.56	2.78	6.82	6.13
	Singh	5.88	5.88	6.13	5.84
	Veer	24.74	23.20	30.74	30.84

Table 3.5: *Leave-one-out* (LOO) and *out of sample* (OS) misclassification rates (in %) of different methods with and without the augmented nearest neighbour covariates (NNC). Here $k = 0$ refers to no NNC and $k = 1$ refers to first NNC included in the initial covariate set. A selection of best 2000 genes was made for all but colon data for comparison. In colon data P is taken to be maximum possible genes after filtering.

3.7 CONCLUSION

We have discussed the plausibility of using a modified classification procedure to improve prediction accuracy in existing methods. Performance of the approach is evaluated through one simulated and four microarray data sets. The method is flexible and provides better results in most situation.

The simulation is constructed such a way that the decision boundary between two classes is non-linear. It is found that all methods got substantial improvement through the use of NNC approach. Table 3.3-3.5 demonstrate the misclassification error rates using different methods for a selection of 500, 1000 and 2000 genes. In colon data the number of genes is much lower than 2000 after applying filtering and thresholding steps; so maximum available genes are used in such case. We see from the result of LOO cross-validation that introducing NNC improves classification accuracy in almost all methods for each of the gene selection. In colon data, SCRDA attains the best performance thanks to adding NNC when $P = 500$ or 1000. DLDA enters the set of best classifier in leukemia data when NNC is used. In prostate cancer data, DLDA experiences improvement for all gene selections. The gain in prediction power due to NNC in other methods depends on the selection of P for this data. SCRDA is the best prediction method in breast cancer data. This method still gains some more accuracy due to NNC. Introducing NNC also improves the classification performance of DQDA in breast cancer data for larger number of genes. We see from the tables that SVM does not gain improvement for most of the instances. The application of re-sampling technique shows that some systemic decrease in misclassification rate can be gained through the use of first NNC.

Investigation showed that some other dimension reduction techniques; for example, *principal component regression* (PCR) or *partial least squares regression* (PLSR)

with augmented NNC can provide very good result. This approach can be extended to any classification rule for plausible improvement.

3.8 FUTURE WORK

We will extend this approach to study the performance in multi-class problem. The procedure can take multiple number of *nearest neighbour covariates* (NNC). We will develop some adaptive selection procedure for the optimal number of NNC to be finally added in the model. Another problem of interest would be to study if class prediction can be improved by combining predictors as has been found for time series forecasting.

REFERENCES

Alon, A., Barkai, N., Notterman,D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999). Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays. *Proc. Natl. Acad. Sci.*, **96**, 6745-6750.

Dudoit, S., Fridlyand, J. and Speed, T. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Stat. Assoc.*, **97**, 7787.

Efron, B., Hastie, T., Johnstone, I.M. and Tibshirani, R. (2002). Least Angle Regression. *Technical report, Department of Statistics, Stanford University.*

Fort, G. and Lambert-Lacroix, S. (2005). Classification using partial least squares with penalized logistic regression. *Bioinformatics*, **21** , 1104-1111.

Ghosh, D. (2003). Penalized discriminant methods for the classification of tumors from gene expression data. *Biometrics*, **59**, 992-1000.

Golub, T., Slonim, D., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J., Coller, H., Loh, M., Downing, J., Caligiuri, M., Bloomfield, C. and Lander, E. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286 (5439)**, 531-537.

Guo, Y., Hastie, T. and Tibshirani, R. (2007) Regularized linear discriminant analysis and its application in microarrays. *Biostatistics*. **8**, 1, 86-100.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*. New York: Springer-Verlag.

Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M., Yakhini, Z., Ben-Dor, A., Dougherty, E., Kononen, J., Bubendorf, L., Fehrle, W., Pittaluga, S., Gruvberger, S., Loman, N., Johannsson, O., Olsson, H., Wilfond, B., Sauter, G., Kallioniemi, O.-P., Borg, A. and Trent, J. (2001). Gene-Expression Profiles in Hereditary Breast Cancer, *The New England Journal of Medicine*, **344**, 539-548.

Holmes, C. C. and Adams N. M. (2003). Likelihood inference in nearest-neighbour classification models. *Biometrika*, **90**, 1, 99-112.

Iizuka, N., Oka, M., Yamada-Okabe, H., Nishida, M., Maeda, Y., Mori, N., Takao, T., Tamesa, T., Tangoku, A., Tabuchi, H., Hamada, K., Nakayama, H., Ishitsuka, H., Miyamoto, T., Hirabayashi, A., Uchimura, S. and Hamamoto, Y. (2003) Oligonucleotide microarray for prediction of early intrahepatic recurrence of hepatocellular carcinoma after curative resection. *The Lancet*, **361**, 923-929.

Lee, J. W., Lee, J. B., Park, M. and Song, S. H. (2005). An extensive comparison of recent classification tools applied to microarray data. *Computational Statistics & Data Analysis*, **48**, 869-885.

Nguyen, D. and Rocke, D. (2002). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, **18(1)**, 39-50.

Nutt, C. L., Mani, D. R., Betensky, R. A., Tamayo, P., Cairncross, J. G., Ladd, C., Pohl, U., Hartmann, C., McLaughlin, M. E., Batchelor, T. T., Black, P. M., von Deimling, A., Pomeroy, S. L., Golub, T. R. and Louis, D. N. (2003). Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. *Cancer Research*, **63(7)**, 1602-1607.

Simon, R. M., Korn, E. L., McShane, L. M., Radmacher, M. D., Wright, G. W., Zhao, Y. (2004). *Design and Analysis of DNA Microarray Investigations*. New York: Springer.

Singh, D., Febbo, P., Ross, D., Jackson, G., Manola, J., Ladd, C., Tamayo, A., Renshaw, A., DAmico, A. V., Richie, J., Lander E., Loda, M., Kantoff, P., Golub, T. and Sellers, W. (2002). Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell.* **1**, 203209.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B.*, **58**, 1, 267-288.

vant Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A. M., Mao, M., Peterse, H. L., van der Kooy, K., Marton, M. J., Witteveen, A. T., Schreiber, G. J., Kerkhoven, R. M., Roberts, C., Linsley, P. S., Bernards, R. and Friend, S. H. (2002). Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer. *Nature*, **415**, 530-536.

Yu, H. (2002). Rmpi: Parallel Statistical Computing in R. *R News*, **2(2)**, 10-14.

Chapter 4

CONCLUSION

Topics in microarray data analysis are investigated and improved methodology developed. A method using Pearson curve fitting to calibrate the null distribution is developed for finding periodic genes in short time time-regulated microarray experiment. This method outperforms Fisher's g test in case of short series. Moreover, this method is robust to missing observations in the series. Change point detection in CGH arrays is addressed using wavelet analysis. Maximum overlapping discrete wavelet transform is employed to detect loss and gain regions in different chromosomes. Several simulation experiments confirm the superior performance of the proposed method. We have studied the usefulness using the nearest neighbour covariate in addition to the original inputs in order to attain improved class prediction. In the presence of nonlinearity, the proposed method produces substantial gains in prediction accuracy.

CURRICULUM VITAE

MOHAMMAD SHAHIDUL ISLAM

Post-secondary Education and Degrees

- **Ph.D.** (*Statistics*), The University of Western Ontario, London, Canada, 2008.
- **M.Sc.** (*Statistics*), McMaster University, Hamilton, Ontario, Canada, 2002.
- **M.Sc.** (*Statistics*), Shahjalal University of Science and Technology, Sylhet, Bangladesh, 1997.
- **B.Sc.** (*Statistics*), Shahjalal University of Science and Technology, Sylhet, Bangladesh, 1996.

Honours and Awards

- Western Graduate Research Scholarship, 2003 – 2007
- Special University Scholarship, 2003 – 2004
- Chancellor Gold Medal in B.Sc. (Hons), 1996 .

Presentation

- Poster presentation at *Western Engineering & Science Research Showcase* presented by the University of Western Ontario on Friday January 25, 2008.

Articles for Journal Publications

- Islam, M. S. and McLeod, A. I. (2008, to be submitted). Testing periodicity and application to gene expression data.
- Islam, M. S., McLeod, A. I. and He, W. (2008, to be submitted). A method for analysis of CGH microarray data.
- Islam, M. S. and McLeod, A. I. (2008, to be submitted). Improved class prediction in gene expression microarray data.