Statistical Methods for Analyzing Genomic Data with Consideration of Spatial Structures

Xuesong Yu

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Washington

2007

Program Authorized to Offer Degree: Public Health and Community Medicine - Biostatistics

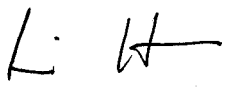
University of Washington
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Xuesong Yu

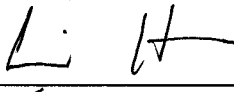
and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of the Supervisory Committee:

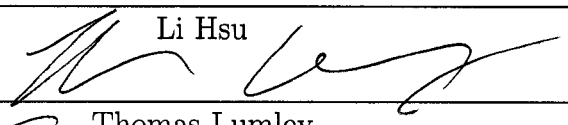


Li Hsu

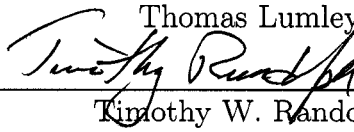
Reading Committee:



Li Hsu



Thomas Lumley



Timothy W. Randolph

Date: 11/18/2007

In presenting this dissertation in partial fulfillment of the requirements for the doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to Proquest Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, 1-800-521-0600, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature Xuesong Yu

Date 11/18/2007

University of Washington

Abstract

Statistical Methods for Analyzing Genomic Data with Consideration of Spatial Structures

Xuesong Yu

Chair of the Supervisory Committee:
Affiliate Professor Li Hsu
Department of Biostatistics

High-dimensional genetic data, such as DNA copy number and single nucleotide polymorphism (SNP), enable researchers unprecedented capabilities for studying genetic basis of diseases. In this dissertation, we develop statistical methods for analyzing two types of high-dimensional genomic data with consideration of spatial structures. In part one, we consider detecting DNA copy number changes using multi-scaled wavelet transformation. Genomic instability, such as copy number losses and gains, occurs in many genetic diseases. Studies of such genomic instability can help us understand the underlying mechanism of disease occurrences and progression. Array-based Comparative Genomic Hybridization (array-CGH) is a powerful technology for measuring copy numbers at thousands of loci simultaneously. We propose a wavelet-based non-parametric approach for detecting copy number changes. The maximum of 2-scale wavelet products across scales, as a novel test statistic, is motivated by combining information across scales to improve power. We explore two non-parametric approaches for estimating the null distribution, including permuting wavelet coefficients at finest scale and permuting residuals after lowess smoothing. Adjusted p-values are estimated using step-down maxT permutation algorithm by controlling the family-wise error rate. To avoid the false positives caused by autocorrelations between adjacent

wavelet coefficients, we propose to test locations at which only local maxima occur. Finally, we illustrate our method using two real data sets and perform a simulation study to investigate the finite sample performance of our method compared with two existing methods– a sequential testing method and a model selection method.

In part two, we consider genetic association studies with tightly linked SNP markers using family data. The transmission/disequilibrium test (TDT) is a popular approach in assessing the linkage disequilibrium between a marker locus and candidate disease locus. To account for local dependency in the presence of phase ambiguity, the TDT has been extended to multiple tightly linked markers by constructing haplotypes statistically. As an alternative, we propose a locally weighted TDT approach that weighs the contribution of multiple SNPs within a prespecified neighborhood according to their association with the locus of interest. We illustrate our method using GAW14 data.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank my committee members Peggy Porter, Bruce Weir, Thomas Lumley, Timothy Randolph and Hua Tang for their valuable feedback and help on my work. Timothy Randolph and Hua Tang have met with me regularly and have given me many excellent ideas and suggestions. In particular, I would like to thank my advisor, Li Hsu, for her support, enthusiasm and encouragement. She has set an exceptional example in both my professional and personal life. Without her guidance, this work would not have been possible. I also want to thank Carl Ton who has given me several helpful suggestions for the data analysis. Many thanks are due to Ellen Wijsman and Deborah Donnell for guiding me through an incredible learning experience in my research assistantship. Finally, I want to acknowledge the constant support, the unconditional love and the happiness that my husband, Jiangning Li and my daughters, Claire and Michelle, have brought to my life.

DEDICATION

To my parents

Chapter 1

INTRODUCTION

1.1 Genomic Instability and Array CGH Technology

Human cells have 23 pairs of chromosomes encoding our genetic information. The integrity and stability of chromosomes enable the cell to transmit accurately its genetic information and function properly physiologically. Genomic instability refers to the propensity for aberrations in chromosomes such as rearrangements, deletions, amplifications and other copy number changes. It occurs in many genetic diseases including for example, Down syndrome, a well-known developmental abnormality which is caused by the trisomy (triplication) of the 21st chromosome. Genomic instability is also involved in many complex diseases. For example, mental retardation may be caused by submicroscopic telomeric chromosome rearrangements (Veltman et al. 2002). Recently genomic instability has been actively studied in cancer. It is believed that cancer develops as a result of an accumulation of genetic aberrations at chromosomal loci that are critical to maintaining normal function. Studies of such genomic instability in these diseases can help us understand the underlying mechanism of disease occurrences and progression.

1.1.1 Array CGH Technology

A popular approach for assessing genomic instability is to measure copy numbers at thousands of loci throughout the whole genome by using the array-based comparative genomic hybridization (array-CGH) (Pinkel et al. 1998; Snijders et al. 2001). Typi-

cally, genomic DNA of a test sample (e.g. tumor sample) is labeled with red (Cy5) dye and genomic DNA of a reference genomic DNA (normal sample) is labeled with green (Cy3) dye. The two differentially labeled genomic DNA are pooled together and co-hybridized onto an array spotted with thousands of genomic clones. Then the relative fluorescent intensities between the reference sample and the test sample are captured by an image scanner. These fluorescent intensity ratios of two samples correspond to the relative DNA copy numbers in the test sample at these markers compared with the reference sample. The schematic representation of array CGH technology is shown in Figure 1.1. Two examples of typical array CGH profiles are plotted in Figure 1.2, where the X-axis is the chromosomal locations of thousands loci from chromosome 1 to 23 and the Y-axis is the corresponding log-relative intensities of a breast tumor versus a normal sample.

Gains or losses in copy number are shown by log-relative intensities greater or less than zero. Though not necessarily always the case, they often occur by segment rather than by single locus, exhibiting a strong local spatial correlation. Then the regions or genes that show consistent copy number gains or losses across samples are further interrogated for their possible involvement in the disease occurrence (Pinkel et al. 2005). Interestingly, it seems that there is also a consistency in the locations where chromosomes break such that the copy numbers are different at the break points. For example, recurrent chromosome break points were observed in breast cancer (Huang et al. 2004) and in neuroblastoma (Stallings et al. 2006).

1.1.2 Issues with Array CGH

Under an ideal situation, with a single cell and no measurement error, the intensity ratio between the two channels is 3/2 for a single copy gain, and 1/2 for a single

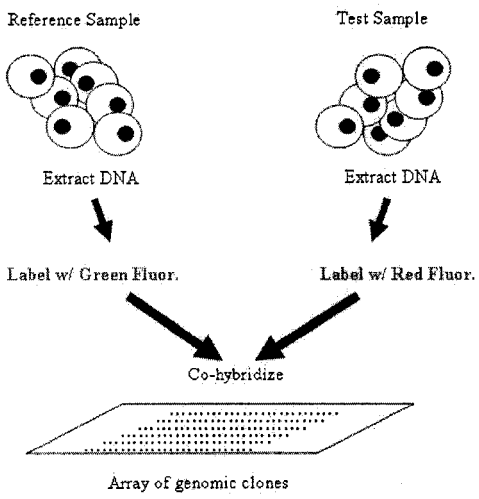


Figure 1.1: Schematic representation of array CGH. The test and reference samples are labeled with two different fluorescent dyes and competitively hybridized to a microarray constructed using Bacterial Artificial Chromosome (BAC) clones. The ratio of the fluorescence intensity for each spot indicates the relative DNA copy number in the test sample to the reference sample.

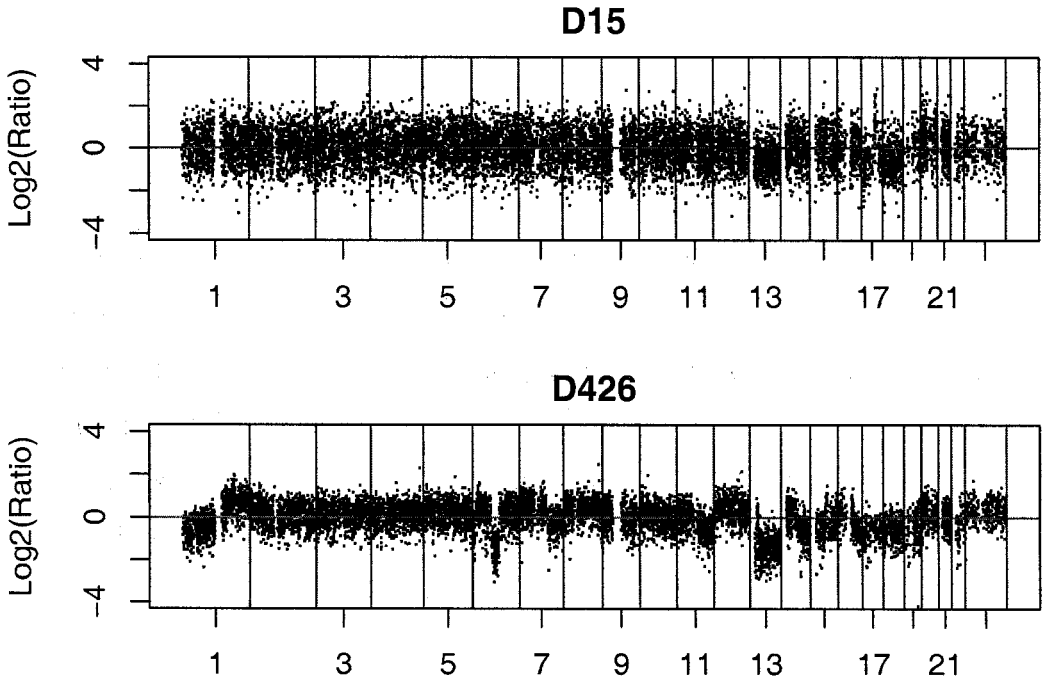


Figure 1.2: Visualization of array CGH profiles for two tumor samples. The x-axis is labeled with the chromosome number separated by blue vertical lines. The y-axis is log2 ratios of intensities between test and reference samples. The red horizontal lines indicate no copy number change ($\log_2$(ratio)=0). Array CGH profiles could range from having little copy number change (top) to having multiple aberrations (bottom).

copy loss, and 2/2 for no copy change for the autosomal chromosomes when the copy number for the reference sample is 2. However, due to stochastic variations, the measured intensity ratios of two dyes are often continuous, with measurements clustering around the underlying true DNA copy numbers. In addition, the current technology can only handle a collection of cells, the underlying true DNA copy number for the test sample may be continuous, not fall exactly at 3/2 for a single copy gain or 1/2 for a single copy loss, due to heterogeneity of cells within a sample. In other words, the underlying true DNA copy number for the test samples is an average of copy numbers from a mixture of heterogeneous cells. This actually makes estimation of underlying discrete DNA copy number intractable without further assumption or information.

Since we donot know the truth, we'll use the total spatial correlation to infer the underlying copy number. We take the approach by first detecting change points at which the copy numbers change, then segmenting the chromosomes into regions where clones within one region share the same underlying DNA copy number from the observed intensity ratios before any subsequent statistical analysis such as comparing copy number changes among tumor subtypes and finding common combinations of changes in tumor samples. We call this procedure "segmentation". The simplest approach to identify aberrations is by visual examination. However the manual identification of DNA copy number changes is not only time-consuming, but also subjective and not reproducible. For example suppose we are given array CGH data on four chromosomes as shown in Figure 1.3. While it is straightforward to detect manually the aberrations for the profiles (a) and (b), it becomes more difficult for the profiles (c) and (d). Though the existence of aberration regions in the profiles (c) and (d) is not arguable, different people are very likely to give different results about the change points at which aberrations occur. So an automated objective and systematic method is needed for segmentation of array CGH data.

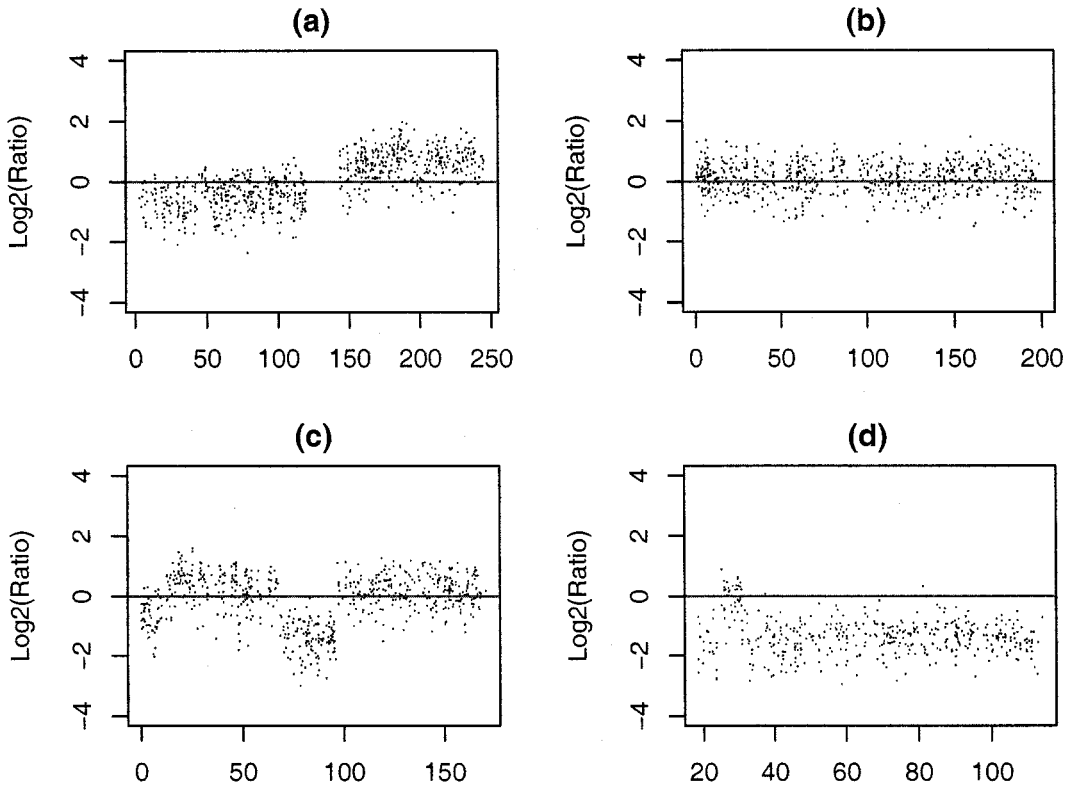


Figure 1.3: Scatterplots of the observed log2 ratios of intensities for different chromosomes of a tumor sample. The x-axis is the chromosomal location in megabases (Mb). The y-axis is the log2 ratios of intensities. The red horizontal lines indicate no copy number change ($\log_2$(ratio)=0).

1.2 *Existing Methods for Segmentation of Array CGH Data*

1.2.1 Overview

An array CGH profile can be viewed as a sequence of segments, ordered by their genomic locations, such that markers within each segment have the same underlying true copy number. In other words the true copy numbers are considered to be a piece-wise constant function with discrete steps. Many methods and algorithms have been developed in the past few years for segmenting array CGH data and can also be used to identify change points, at which copy numbers differ in the neighboring clones. Generally speaking, all these methods can be broadly categorized into three main approaches.

The first approach is to use model selection procedures to balance trade-offs between maximizing a likelihood function with increasing number of parameters in the model and the penalty from overfitting. For example, these include a Gaussian likelihood function based on a piece-wise constant model using either a fixed penalty parameter (Jong et al. 2003), a data-driven penalty function (Picard et al. 2005) or a modified BIC (Zhang et al.2007), an unsupervised Hidden Markov Model using Bayesian information criterion (BIC) or Akaike information criterion (AIC) (Fridlyand et al. 2004). More recently, Bayesian techniques were used to estimate the posterior probabilities of markers being changing points. Methods include Wen et al. (2006), Engler et al. (2006), and Guha et al. (2006).

The second approach is based on function estimation by considering the measured log-relative intensities as log-relative true copy numbers measured with error. Non-parametric function estimation techniques could be used to infer underlying true copy numbers. For example, Eilers et al. (2005) proposed a quantile smoothing method to account for the square wave features of the copy number data while smoothing

the function . Other methods also include a weighted likelihood method with weights determined adaptively (Hupe et al. 2004), a wavelet-based denoising method by Hsu et al. (2005) and a "fused lasso" smoothing method by Tibshirani et al. (2007). Even though some nonparametric techniques such as wavelets are much suited to detect sharp changes in the function, the smoothed or denoised function is typically not a step function. An additional step may be taken by clustering adjacent markers with similar smoothed values into one segment and estimating the copy number with the empirical average of log-relative intensity values over the marker loci within the segment (Hsu et al. 2005).

The third approach is to select clusters by controlling the overall type I error rate. For example, Olshen et al. (2004) proposed a circular binary segmentation (CBS) method using a sequential testing procedure. They considered all possible locations and sizes of a square shaped function, calculated t-test alike statistics to measure the mean differences between two segments for each possible combination of segments, and compared the maximum test statistics against a null distribution. The genome is partitioned by the segments with the significant test statistics and each segment is then applied the same testing procedure. This procedure continues until no test statistic in each segment exceeds the critical value at a pre-specified significant level. Wang et al. (2005) proposed another approach to select "significant" clusters by controlling false discovery rate (FDR), where the clusters are hierarchically formed along the chromosome based on the "similarity" measurement between adjacent markers or clusters.

Lai et al. (2005) compared the performance of 11 publicly available segmentation methods under different scenarios using both simulated and real data sets. They found that when the noise level is high, smoothing methods (such as, wavelets denoising (Hsu et al. 2005) and quantile regression (Eilers and de Menezes 2005) appeared to work well. The homoscedastic model in Picard et al. (2005) and Olshen et al. (2004)'s change-point method performed consistently well in most of cases. In terms

of the speed of the algorithms, the Hidden Markov Model (Fridlyand et al. 2004) and Olshen et al. (2004)'s method are the most computational intensive and smoothing algorithms are among the fastest ones. (More recently, in order to speed up the CBS algorithm, Venkatraman and Olshen (2007) adopted a sequential testing approach to reduce the number of permutation when there exists strong evidence for a change point.) Almost at the same time, Willenbrock and Fridlyand (2005) reported a comparison study for methods developed by Olshen et al.(2004), Fridlyand et al. (2004) and Hupe et al. (2004) using both simulated and real data sets. They concluded that the method by Olshen et al.(2004) has the best performance in terms of sensitivity and FDR. In what follows we'll give further details of some of these methods starting with the model selection approach.

1.2.2 Model Selection Approach

Consider n markers on one chromosome or genome ordered by their physical locations. Denote by Y_i the observed log intensity ratio for the ith marker, for $i = 1, \ldots, n$. Suppose there are R change points which lead to $R+1$, $0 \leq R \leq n-1$, nonoverlapping segments. Let c_r denote the index of the last marker in the rth segment, $c_1 < c_2 < \ldots < c_R$, and μ_r be the constant mean in the rth segment. By convention, define $c_0 = 0$ and $c_{R+1} = n$. We call $\{c_r : r = 1, \ldots, R\}$ the change points. Then in the rth segment we have

$$Y_i = \mu_r + \epsilon_i \qquad \text{if } c_{r-1} < i \leq c_r, \qquad r = 1, \ldots, R+1 \tag{1.1}$$

where ϵ_i is independently and identically Gaussian distributed with mean 0 and variance σ^2. Although the i.i.d. normality assumption is arguable, most existing methods rely on this assumption. Fridlyand et al. (2004) stated that the i.i.d. normality assumption is supported by the self-to-self hybridizations data (i.e. a sample is hybridized to itself) where $Y_i = \epsilon_i$. Picard et al. (2005) compared two models: heteroscedastic model $M_1 : \epsilon_i \sim N(0, \sigma_r^2)$ for $c_{r-1} < i \leq c_r$ and homoscedastic model

$M_2 : \epsilon_i \sim N(0, \sigma^2)$ and found that model M_2 with constant variance tends to estimate more change points than model M_1 without assuming same variance. Under a homoscedastic model M_2, one can compute the log-likelihood by $\mathcal{L}_R = \sum_{r=1}^{R+1} \ell_r$, with

$$\ell_r = -\frac{1}{2} \sum_{i=c_{r-1}+1}^{c_r} \left\{ \log(2\pi\sigma^2) + \left(\frac{y_i - \mu_r}{\sigma} \right)^2 \right\} \tag{1.2}$$

Given the number of change points R and the locations of the change points $(c_1 < c_2 < \ldots < c_R)$, it is straightforward to estimate the mean for each segment and the variance using the maximum likelihood method, giving

$$\hat{\mu}_r = \frac{1}{c_r - c_{r-1}} \sum_{i=c_{r-1}+1}^{c_r} y_i \tag{1.3}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{r=1}^{R+1} \sum_{i=c_{r-1}+1}^{c_r} (y_i - \hat{\mu}_r)^2 \tag{1.4}$$

Plugging estimates $\hat{\mu}_r$ (1.3) and $\hat{\sigma}^2$ (1.4) into log likelihood function (1.2), we can obtain estimates for change point locations by maximizing the likelihood function (1.2) with respect to $\{c_1, c_2, \ldots, c_R : c_1 < c_2 < \ldots < c_R\}$ for given R.

Now the key problem is how to select the optimal number of change points R. The likelihood function $\mathcal{L}_R$ increases with the number of change points and reaches the maximum when $R = n - 1$. That is, each segment includes only one marker. Selecting the number of change points purely based on maximizing the likelihood will lead to a saturated model. To avoid this over-fitting problem, a penalized version of the likelihood that balances between a good fitting and a reasonable number of parameters in the model is adopted. A general form of a penalized likelihood function is,

$$\tilde{\mathcal{L}}_R = \hat{\mathcal{L}}_R - \beta \times pen(R) \tag{1.5}$$

where $pen(R)$ is a penalty function that increases with the number of change points, and β is a factor. The estimated number of change points is then $\hat{R} = \text{argmax}_R(\tilde{\mathcal{L}}_R)$.

In this model formulation, it is clear that choosing appropriate β and $pen(R)$ becomes critical for correctly detecting change points. For example, Jong et al. (2003) proposed to set $\beta = 10/3$ and $pen(R) = 3R$ as they considered the number of parameters to be estimated equal to the number of means and variances plus the number of change points. Picard et al. (2005) demonstrated the most commonly used model selection criteria BIC where $\beta = 1/2\log(n)$ and Akaike information criterion (AIC) where $\beta = 1$ tended to overestimate the number of change points. So instead of choosing universal β, Picard et al. (2005) proposed a data-driven penalty function by defining β as a decreasing sequence with respect to the number of segments $R + 1$. Briefly let $\beta = \{\beta_0, \ldots, \beta_{R_{\max}+1}\}$, where $R_{\max}$ is the maximum number of change points, such that $\beta_0 = \infty$ and for $i \geq 1$

$$\beta_i = \frac{\hat{\mathcal{L}}_{R_{i+1}} - \hat{\mathcal{L}}_{R_i}}{pen(R_{i+1}) - pen(R_i)}$$

for $R_i = 1, 2, \ldots, R_{\max} + 1$. If one views $\hat{\mathcal{L}}_R$ as a response variable and $pen(R)$ as a predictor variable, the sequence of β_i represent the slopes between two adjacent points. The method aimed at selecting R for which $\hat{\mathcal{L}}_R$ ceases to increase significantly. In other words, the estimated number of change points is the largest R such that the second derivative (difference) of the likelihood is smaller than a given threshold, i.e.

$$\hat{R} = \max\{R \in (1, \ldots, R_{\max}) : \hat{\mathcal{L}}_{R-1} - 2\hat{\mathcal{L}}_R + \hat{\mathcal{L}}_{R+1} < s \times n\}$$

where s is a constant. The performance of the method relies on the correct choice of s. The limitation of this method is that the choice of s is arbitrary and there is lack of theoretical justification.

Recently, Zhang and Siegmund (2007) pointed out that the classic Bayesian information criterion (BIC) can not be appropriately applied for many change point problems due to irregularities in the likelihood functions at change points. Specifically, the Taylor expansion used in the derivation of the BIC does not hold at the change points since the function is not differentiable at the change points. So they

derived a modified version of the BIC by asymptotically approximating the Bayes factors. The modified BIC also consists of two terms in which first term consists of the log-likelihood, just like BIC, but the penalty term is different and determined by the model and data.

Once the number of change points is determined, the burden lies in the computation for searching $\{c_1, \ldots, c_R\}$. Picard et al. used dynamic programming strategy to globally search for the maximum likelihood and significantly reduced the computational cost from $O(n^R)$ to $O(n^2)$ for any R. This algorithm ensures that change-point estimators indeed converge to the true change points. Jong et al. (2003) proposed a local search algorithm to find the locations of change points by maximizing the penalized likelihood locally. However, the local search algorithm is not guaranteed to find the consistent estimators of the change points.

Another interesting method in this category is by Fridlyand et al. (2004). They proposed to use a Hidden Markov Model (HMM) treating the underlying true but latent copy numbers as hidden states with certain transition probabilities along the chromosome and form a likelihood function based on this model. As in Picard et al. (2005) and Jong et al. (2003), Fridlyand et al. (2004) also used the penalized likelihood function in choosing the number of states. They used the AIC or BIC followed by a merging step to reduce the number of false positives. Since the hidden states are "missing", an EM algorithm was used to maximize the likelihood function. Initial estimates for HMM parameters are critical for obtaining optimal results and moreover the threshold for merging needs to be tuned appropriately. For more details on the algorithm and estimating of initial parameters, see Fridlyand et al. (2004).

1.2.3 Function Estimation Approach

Another approach for segmenting array CGH data is to first smooth or denoise the data and then apply a reasonable threshold to assign the smoothed data into several states. Instead of estimating the location of change points first, the smoothing-based

methods start at estimating μ_i, for each $i = 1, \ldots, n$, assuming

$$Y_i = \mu_i + \epsilon_i \qquad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2), \tag{1.6}$$

then segment the whole sequence to form the model given by equation (1.1). In general, smoothing techniques do not provide estimators of the change points since the primary goal is to uncover the underlying true function. There are several methods in this category and they differ in the ways of estimating μ_i and the segmentation algorithm.

A quantile smoothing method based on the minimization of penalized L_1 norms (sum of absolute errors) rather than L_2 norms (sum of mean squared errors) was proposed to fit the array CGH data (Eilers and de Menezes 2005). It involves minimization of an objective function

$$Q = \sum_{i=1}^{n} |y_i - \mu_i| + \lambda \sum_{i=2}^{n} |\mu_i - \mu_{i-1}| \tag{1.7}$$

where λ is the penalty parameter, the larger λ is, the smoother μ will be. It is not trivial to find the minimization of Q. So they borrowed an idea from the quantile regression to minimize

$$S = \sum_{i}^{n} h_\theta \left(y_i - \sum_{j}^{m} b_{ij} \alpha_j \right) \tag{1.8}$$

where b_{ij} is an element of a matrix of regression basis and α_j's are the regression coefficients to be estimated. Here, $h_\theta(u)$ is the so-called check function which is θu if $u > 0$ and $(1 - \theta)u$ if $u \leq 0$. The right side of (1.8) is a summation of weighted absolute values of residuals, such that it assigns weight θ for positive residuals and $1 - \theta$ for negative residuals. When $\theta = 0.5$, equation (1.7) and (1.8) are equivalent. This is called median regression. When $\theta = 0.25(0.75)$ the lower (upper) quantile curve is fitted. For in-depth discussion on how to choose tuning parameter λ and why equations (1.7) and (1.8) are equivalent, see Eilers and de Menezes (2005). This approach does not provide segmentation after giving a smoothing fit for the data.

Hupe et al. (2004) proposed to estimate μ_i by using a weighted likelihood function with weights determined adaptively. Based on these estimates, the whole sequence can be split into $R+1$ segments in a way such that within a segment adjacent μ_i are not different by a small constant, say 0.01. This adaptive weights smoothing (AWS) procedure requires many tuning parameters and it is not clear how to choose these tuning parameters based on the data. In addition, the threshold criterion (0.01) is somewhat arbitrary and it is not clear how it is related to the data. Furthermore, this procedure needs an additional filtering step which is based on a penalized likelihood in which the penalty function is a kernel function whose role is to reduce the false positive rate.

In Hsu et al. (2005), a nonparametric regression by wavelet method is used to denoise the data and the partition-around-medoids algorithm is used to segment the denoised data into different states. There are a few advantages of the wavelet method (Donoho and Johnstone 1994; Percival 2000): (1) it handles abrupt changes well; (2) it has sound theoretical results, e.g. good time-frequency localization; and (3) it is computationally efficient.

Neither the model-selection based nor the smoothing based methods provide a statistical significance level for the segmentation. Some of these methods require choosing tuning parameters, which can be cumbersome and subjective when dealing with a large number of arrays. All these make it hard to compare among these methods.

1.2.4 Controlling Type I Error Rate

Instead of using the model selection framework, Olshen et al. (2004) proposed a circular binary segmentation (CBS) method which essentially uses the hypothesis testing procedure to select optimal change points. This CBS method is a modification of the classic binary segmentation method proposed by Sen and Srivastava (1975) to allow for testing one or two change points. In the original binary segmentation

method, the likelihood ratio test is used for testing the null hypothesis of no change point against the alternative of one change point at an unknown location i. Let $S_i = y_1 + \ldots + y_i$, $1 \leq i \leq n$, be the partial sum up to i markers. The test statistic is given by $Z = \max_{1 \leq i < n} |Z_i|$, where

$$Z_i = \frac{1}{\sqrt{\sigma^2/i + \sigma^2/(n-i)}} \{S_i/i - (S_n - S_i)/(n-i)\}$$

with $\hat{\sigma} = \text{MAD}(\text{diff}\ (y_i))/\sqrt{2}$ and MAD stands for median absolute deviation and diff (y_i) is the set of adjacent difference of y_i. Note that Z_i is a two sample t-test with equal variance, except $\hat{\sigma}$ is estimated by MAD not by sample variance for robustness reason. The location of the change point is estimated to be i such that $|Z_i| = Z$ exceeds the critical value obtained for Z under the null hypothesis of no change point. The procedure then continues by splitting the segment into two parts at the change point and apply the same test statistic to each of the two parts separately. The procedure will stop if there is no further change points detected in any of the segments obtained from the estimated change points. The limitation of this procedure is that it has little power to detect a short segment buried in the middle of a long segment.

To overcome this, Olshen et al. (2004) modified it to test the null against the square wave alternative with two change points, the test statistic is given by $Z_c = \max_{1 \leq i < j \leq n} |Z_{ij}|$, where

$$Z_{ij} = \frac{1}{\sqrt{\sigma^2/(j-i) + \sigma^2/(n-j+i)}} \{(S_j - S_i)/(j-i) - (S_n - S_j + S_i)/(n-j+i)\}.$$

Without the normality assumption, a permutation approach permuting the locations of intensity ratios may be used to generate the null distribution. Although it is appealing that normality is not required , the trade-off is the intensive computational cost for generating the null distribution by permutation. The CBS method controls the overall type I error rate only if there is no change point on the entire chromosome. Since it searches sequentially within each segment, the number of spurious change points actually increases with the number of true change points in the data.

Wang et al. (2005) proposed a hierarchical clustering-based algorithm, "cluster along chromosomes" (CLAC), for the analysis of array CGH data. Clustering is one of the most important unsupervised learning techniques which deals with finding a structure in a collection of data. It aims to group observations which are similar among themselves within a group but dissimilar to those that belong to other groups. There are a few potential drawbacks with clustering algorithms, some of which this approach inherits: (1) high dimension due to a large number of observations; (2) computational intensity; (2) the effectiveness of clustering depending on the distance measure; (3) possibly different ways to interprete the results of the clustering algorithm.

Array CGH data fit in the clustering framework well if one considers a segment with same copy number as a cluster. Agglomerative clustering algorithm is a bottom-up clustering method that generates a binary tree to represent the similarities in the data. It begins with every observation representing a singleton cluster. Then, in each successive iteration, it agglomerates (merges) the closest two clusters by some similarity measure and produces one less cluster, until no two clusters meet the similarity criteria. Since the order of the markers along the chromosome is fixed, the "similarity" between two clusters no longer refers to the spatial distance but to the similarity of the intensity ratios between the two contiguous clusters. So for the CLAC algorithm only contiguous clusters are merged when building the tree bottom-up.

There are three commonly used methods to determine which two clusters are sufficiently similar to be linked (merged) together:

- Single Linkage method - cluster objects based on the minimum distance between them (also called the nearest neighbor rule)

- Complete Linkage method - cluster objects based on the maximum distance between them (also called the furthest neighbor rule)

- Average Linkage method - cluster objects based on the average distance between

all pairs of objects

Wang et al. used the complete linkage and the single linkage methods. After a hierarchical tree was built, the "significant" clusters with copy number gains/losses were selected by controlling the false discovery rate (FDR), and the null distribution is determined by an independent set of normal-normal data. However, it tends to overestimate the number of change points (Lai et al. 2005, Tibshirani et al. 2007).

Both Olshen et al. (2004) and Wang et al. (2005) provided approaches that controls the overall type I error rate, that is, the probability of finding one or more false change points under the global null hypothesis of no copy number changes in the chromosome. Wang et al. (2005)'s approach requires an external set of normal samples in order to obtain the null distribution. The downside of Olshen et al. (2004)'s approach is that the false positive rates increase with the number of true change points. Moreover, both methods require one to commit a particular level of significance as traditional hypothesis tests do. To detect change points at a different significant level, one would need to re-run the segmentation procedures. Obviously it is not very efficient, especially when individual change points are of interest. An alternative way to making such commitment is to estimate the p-value for each marker under the null hypothesis of the marker is not a change point. This allows investigators to examine change points at their own significant levels without re-running programs again and again.

1.3 Outline of Dissertation

In this dissertation, we develop statistical methods for analyzing two types of high-dimensional genomic data with consideration of spatial structures. In the first part, we consider detecting DNA copy number changes using multi-scaled wavelet transformation. We aim to develop a test statistic for examining the likelihood of a locus being a change point and provide an approach to estimate an individual p-value for

each of these candidate change points while accounting for the multiple comparison. Estimating an individual p-value by controlling family-wise error rate provides some flexibility for scientists to call change points at their chosen significant level. We'll use the wavelet transform as a basis for our test statistics as it is particularly suitable for investigating the local irregularity of functions. The rest of the dissertation is organized as follows. Chapter 2 reviews the wavelet transform. In Chapter 3, we propose a test statistic for detecting change points and provide non-parametric methods of estimating p-values for change points. We also provide multiple comparison adjustment procedures. The estimated change points are shown to be asymptotically consistent. Two real data sets are used to illustrate the proposed method and the results are shown in Chapter 4. Chapter 5 describes the results from a simulation study for examining the performance of the proposed approach and methods proposed by Olshen et al. (2004) and Picard et al. (2005). Finally, Chapter 6 summarizes the proposed methods and provides a discussion of directions for the future research.

In the second part, chapter 7, we consider genetic association studies with tightly linked SNP markers using family data. We illustrate our method using GAW14 data and conduct a simulation study to compare the proposed method with the transmission/disequilibrium test (TDT).

Chapter 2
WAVELET TRANSFORM

Wavelet transform methods have proven to provide a powerful mathematical tool with wide-ranging applications in many areas such as signal and image processing, pattern recognition, and more recently peak identification in proteomics, due to its capability of decomposing a function of interest at both time (or space) and frequency scale. The general applications of wavelet methods can be found in Meyer (1993), Young (1993) and Mallat(1998). For the comprehensive reviews of wavelet applications in statistics, interested readers are referred to Ogden (1997), Vidakovic (1999) and Abramovich et al. (2000). In this Chapter, we will review the concept of the wavelet transform and application of the wavelet transform on detecting change points. In-depth coverage on mathematical theory and underlying ideas of wavelets can be found in Chui (1992), Daubechies (1992) and Meyer (1992).

2.1 Wavelet Transform

2.1.1 The Continuous Wavelet Transform (CWT)

A function $\psi(\cdot)$ is said to be a (mother) wavelet if its Fourier transform $\Psi(f)$ satisfies the admissibility condition, that is

$$0 < C_\psi = \int_0^\infty \frac{|\Psi(f)|^2}{f} df < \infty$$

with $\Psi(f) = \int_{-\infty}^{+\infty} \psi(u) e^{-i2\pi f u} du$. This condition implies that a wavelet is absolutely and square integrable: $\int_{-\infty}^{\infty} |\psi(u)| du < \infty$ and $\int_{-\infty}^{\infty} |\psi(u)|^2 du < \infty$ and furthermore $\int_{-\infty}^{\infty} \psi(u) du = 0$ and $\int_{-\infty}^{\infty} \psi^2(u) du = 1$ which is necessary for a stably invertible transform (Daubechies 1992). It is worth noting that condition $\int \psi(u) du = 0$ implies

that a wavelet ψ is oscillatory around 0, which in fact gives its name. The mother wavelet can then be dilated and translated by $\psi_{s,x}(u) = \frac{1}{\sqrt{s}}\psi(\frac{u-x}{s})$, for any scale $s \in \mathbb{R}^+$ and time (or location) $x \in \mathbb{R}$. The constant $\frac{1}{\sqrt{s}}$ is for energy normalization, i.e.

$$\int_{-\infty}^{\infty} \psi_{s,x}^2(u)du = \int_{-\infty}^{\infty} \psi^2(u)du = 1.$$

Figure 2.1 displays the dilated and translated versions of the Haar wavelet which $\psi(u)$ takes value $-1/\sqrt{2}$ when $-1 < u \leq 0$, $1/\sqrt{2}$ when $0 < u \leq 1$, and 0 otherwise. Three examples of translation $\psi_{1,x}(u)$ by varying time x are at the top row and three examples of scaling $\psi_{s,0}(u)$ by varying scale parameter s are at the bottom row.

For a given wavelet family $\{\psi_{s,x} : s \in \mathbb{R}^+, x \in \mathbb{R}\}$, the continuous wavelet transform (CWT), W, of a function $f \in L^2(\mathbb{R})$ at given time x and scale s is defined as $f \mapsto Wf$, where

$$Wf(s,x) \equiv f * \psi_s(x) = \int \psi_{s,x}(u)f(u)du$$

with $*$ denotes convolution. Grossmann and Morlet (1984) proved a fundamental fact about the CWT - it preserves all the information in f, so f can be recovered by inverting the CWT

$$f(x) = \frac{1}{C_\psi}\int_0^{+\infty}\int_{-\infty}^{+\infty} Wf(s,x)\frac{1}{\sqrt{s}}\psi(\frac{u-x}{s})du\frac{ds}{s^2}.$$

Another important property of wavelets is the regularity condition which is necessary for localization in the frequency domain (Sheng 2000). By taking the Taylor expansion of function $f(x)$ at 0 up to order N, the wavelet transform of f at $x = 0$ can be written as

$$Wf(s,0) = \frac{1}{\sqrt{s}}\left[f(0)M_0 s + \frac{f^{(1)}(0)}{1!}M_1 s^2 + \ldots + \frac{f^{(N)}(0)}{N!}M_n s^{N+1} + O(s^{N+2})\right] \quad (2.1)$$

where $f^{(p)}(0)$ denotes the pth derivative of f and $M_p = \int u^p\psi(u)du$ denotes the pth moment of the wavelet $\psi(\cdot)$. So the first nonzero moment of the mother wavelet $\psi(\cdot)$ determines how fast the wavelet transform $Wf(s,x)$ decays. For a wavelet having

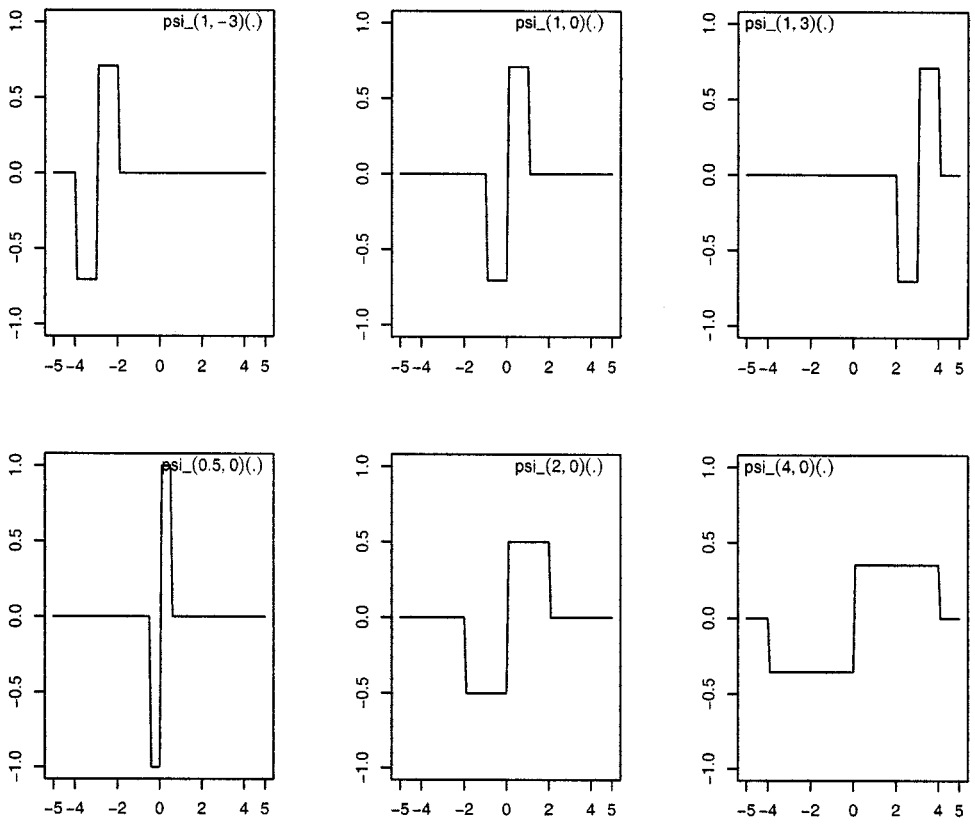


Figure 2.1: The dilated and translated versions of the Haar wavelet. The top row of plots shows the effect of translation (shifting), $\psi_{1,x}(u)$; the bottom row shows the effect of dilation, $\psi_{s,0}(u)$.

$N+1$ vanishing moments, i.e. $M_p = 0$ for $p \leq N$, the wavelet transform $Wf(s,x)$ convergences to zero at a rate of $s^{(N+1/2)}$ for a smooth function f at x according to equation (2.1). This property gives theoretical explanation why the wavelet transform is suitable for detecting irregular features such as change points, peaks or edges as the wavelet transform is influenced by the local regularity of the function f.

Mallat and Hwang (1992) showed that the wavelet transform $Wf(s,x)$ can be interpreted as a first derivative of f smoothed by a differentiable smoothing function $\theta_s(u)$ where the degree of smoothing depends on scale s, that is $Wf(s,x) = s\frac{d}{dx}(f * \theta_s)(x)$. The local extrema of a wavelet transform correspond to sharp variations in the signal, such as change points. Therefore, one could detect the change points by testing local extrema of a wavelet transform. We will denote $W_{s,x} \equiv Wf(s,x)$ whenever there is no confusion. The wavelet transform therefore provides an efficient representation for functions which have similar features to the functions in the wavelet basis.

2.1.2 The Discrete Wavelet Transform (DWT)

The continuous wavelet transform is computed over every possible scale s and time x. It is therefore very demanding in both computation and analysis of all wavelet coefficients because of their redundancy, making it difficult in practical applications. Instead of considering every scale and time, the discrete wavelet transform is computed only at dyadic time and dyadic scale (Mallat 1989). Let n be the number of loci or data points, the discrete wavelet transform can be represented by $W_{j,k} = Wf(2^j, 2^j k)$, where $j = 1, \ldots, J-1$, $k = 1, \ldots, n/2^j$. So the number of data points n needs to be 2^J for some positive integer. The DWT yields a total of $n-1$ wavelet coefficients and one scale coefficient, a much reduced number to work with.

The DWT is an orthogonal transformation of the discrete sample Y. To be specific, let $\mathcal{W}$ denotes the $n \times n$ orthonormal matrix with each row defined by the wavelet basis dilated and translated from the mother wavelet and scaling function, then the DWT can be represented by $w = \mathcal{W}Y$ where w is a vector of length $n = 2^J$. This

implies that $Y = \mathcal{W}^T w$. The DWT can be computed using fast pyramidal algorithm (Mallat 1989). One drawback with the DWT is that it is not invariant to translation; that is, if the signal is slightly shifted, the wavelet coefficients of the shifted signal can be completely different from those of the original signal. In our copy-number data, if a change point occurs at a location x, the localization of the estimated change point using the wavelet coefficients at the coarser scales might not be necessarily consistent with the location x. See Percival et al. (2000) (Page 160) for illustration examples. Hence, one can not consistently locate the change point using the DWT.

2.1.3 The Maximal Overlap Discrete Wavelet Transform (MODWT)

To overcome the problem of inability to precisely locate the break point using the DWT, a modified version of the DWT was introduced to subsample the CWT only at dyadic scales, but not time. This is called the Maximal Overlap Discrete Wavelet Transform (MODWT) (Percival et al. 2000) or translation invariant wavelet transform (Coifman et al. 1995). The MODWT coefficients are computed at all possible locations (time) at dyadic scales. As a result, it eliminates alignment artifacts as seen in the DWT and is translation invariant, i.e. $Wf_u(j,x) = Wf(j,x-u)$ if f_u denotes a translation of f, $f_u(x) = f(x-u)$. This allows us to precisely locate the variations in the signal. As a trade-off, there is a redundancy of wavelet coefficients along the time axis. Compared to the DWT, the MODWT yields a total of $n \log_2 n$ wavelet coefficients. For the purpose of detecting change points, we will use the MODWT so that we don't miss or misalign the locations of the change points.

For illustration, Figure 2.2 shows the wavelet coefficients W_j from MODWT (left panel) and DWT (right panel) across different scales using Haar wavelet. The local maximum and minimum in MODWT align the locations of the change points very well while there is no such alignment for DWT.

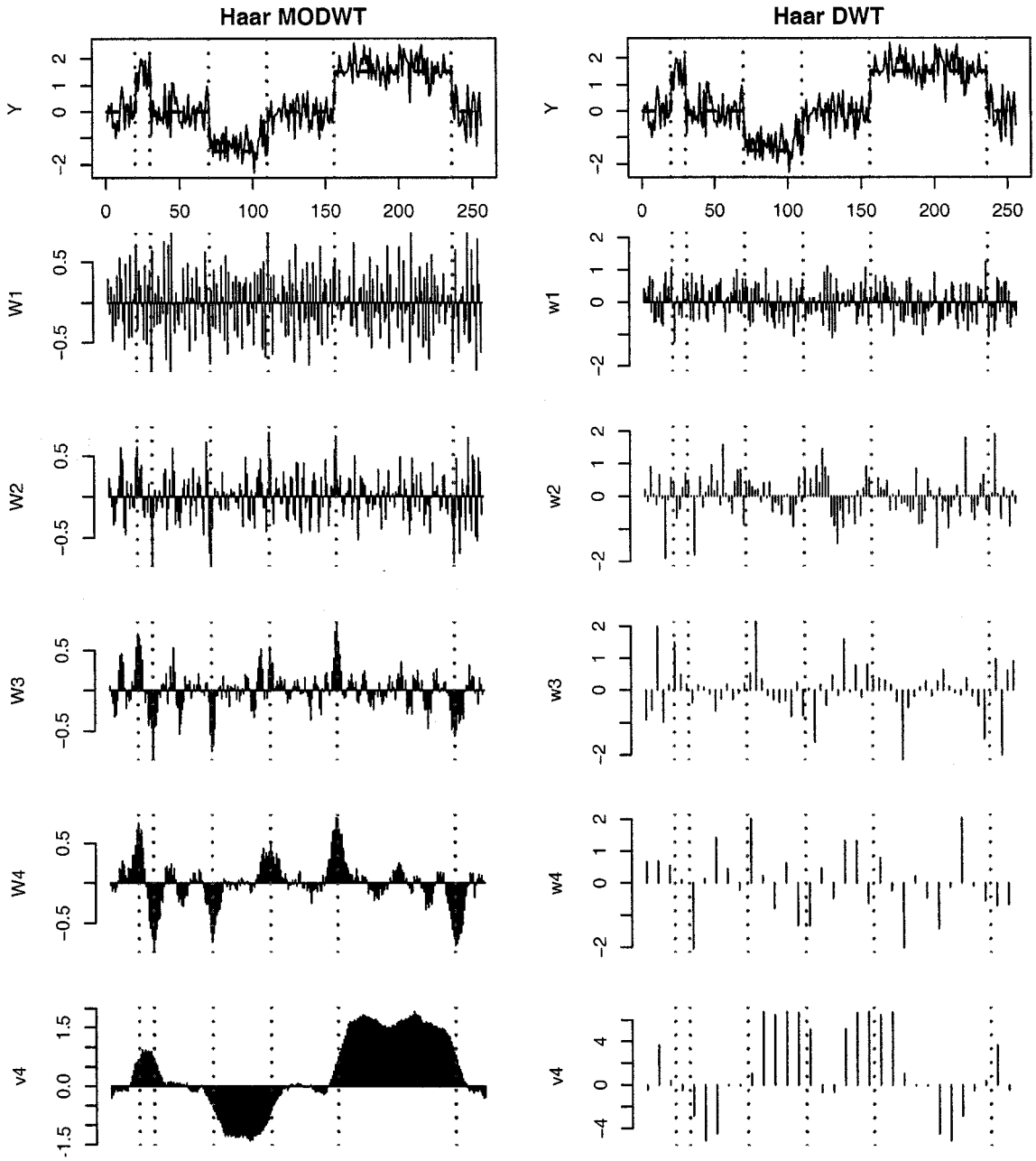


Figure 2.2: Wavelet coefficients W_j using Haar wavelet by MODWT (left panel) and DWT (right panel). The top two plots are the simulated signal Y with the dashed horizontal lines indicating the f. The vertical lines indicate the change points.

2.2 *Detecting Change Points*

Since 1990s the developed theory of wavelets has made the wavelet transform a popular tool in many areas such as function estimation and signal or image analysis. One of the seminal work is the wavelet thresholding methods proposed by Donoho and Johnstone (1994, 1995). Their work inspired a great amount of research in this area. Recently, by treating the wavelet thresholding as a multiple hypothesis testing problem, Tadesse et al. (2005) proposed a method of extracting wavelet coefficients that resulting from true signals by controlling the false discovery rate in a Bayesian framework.

The wavelet transform replaces the Fourier transform as an attractive alternative for detecting change points mainly due to its good time-frequency localization. The Fourier transform is a global representation which is not well suited for finding the locations of the change points. Although there have been many methods proposed for change-point detection using the wavelet transform, the idea behind these methods is nearly always the same, which is to detect change points by finding large values of wavelet coefficients at certain scales as the wavelet transform $Wf(s,x)$ characterizes the local irregularity of f. The methods differ mainly in defining how large is large enough for a coefficient to indicate a change point.

Wang (1995) proposed a test statistic based on the maximum of absolute values of wavelet coefficients at a particular scale using DWT. The critical values at significant level α was given by assuming a Gaussian process under the null. He proved that the number of estimated change points $\hat{R}$ is asymptotically consistent and the location of each estimated change point is consistent with the true locations as well. However, as mentioned by the author, this method was not very powerful when the noise level was high (SNR$<$ 5) and the detection was less precise at coarser scales mainly because this method uses the DWT which is not translation invariant. Instead of using asymptotic approximation for the maxima of Gaussian processes, Raimondo and Tajvidi (2004)

modeled the extreme values over certain threshold based on the generalized Pareto distribution. The choice of the threshold is a trade off between bias and variance of the function estimator and it becomes very critical when the sample size is small. The translation-invariant Haar wavelet was used to improve the power of detection. Antoniadis and Gijbels (2002) proposed a discontinuity locator using the absolute value of wavelet coefficients and showed that the number of detected change points was consistent. The cut-off value for detecting change points involves two parameters, the optimal scale and the exponent of the Hölder condition. However these two parameters depend on the unknown function f and hence are unknown.

The above methods were developed in the frequentist's framework. Ogden and Lynch (1999) proposed a Bayesian approach to estimate the change points by assuming a prior distribution for the change point and derived the posterior distribution of the change point given the empirical wavelet coefficients.

2.3 Some Issues to Consider for Wavelet Analysis

In application, there are several important issues to consider when applying wavelet methods. Here we will address some of the issues that need to be considered in order to make wavelet analysis useful, such as the choice of wavelet and choice of scale.

2.3.1 Choice of Wavelet

The first important practical issue in a wavelet analysis is which wavelet function to use. There are many choices of wavelet families, including the well-known Daubechies wavelets and Coiflet wavelets (Daubechies 1992). In general, different choices of wavelets may lead to different analysis results. The choices of wavelet cannot be made in any absolute sense and very much depends on the application and data. Percival and Walden (2000) suggest that it may be wise to choose the wavelet filter with smallest width L that still gives acceptable analysis since smaller L means fewer coefficients involved in boundary and also suggest to choose the wavelet filter that is

a good match to the underlying features in data.

The simplest is the Haar wavelet. Though simple, it actually fits well for our data for the following two reasons. First, the underlying copy numbers in the array-CGH data are discrete. In the absence of measurement error, the function of log-relative intensities is step-wise, a shape similar to the Haar wavelet. This makes the wavelet representation efficient. Second, since the Haar wavelet has one vanishing moment, the wavelet coefficient $Wf(j,x)$ is the difference of adjacent averages in the neighborhood of location x over the scale j. This property is particularly suitable for detecting change points since the magnitude of $Wf(j,x)$ directly reflects the size of a scale-based change in f at location x. For the same reasons, Hsu et al. (2005) also used the Haar wavelet for denoising the data.

2.3.2 Choice of Scale

A reasonable choice of scales for the wavelet transform is again very much application dependent. For function estimation, one should use all scales commensurate with the properties of the function. As the scale increases, there will be more coefficients affected by insufficient data at the boundary and so putting a limit on the largest scale is desirable. Now the question is how big is too big. For the MODWT, the general rule of thumb is to choose a scale J_0 such that $J_0 < \log_2(N/(L-1)+1)$ where N is the number of data points and L is the width of the wavelet filter which will be further defined in Section 3.2 (Percival and Walden 2000). So for N=500, the upper bound for J_0 is scale 8 if using the Haar wavelet.

In terms of change point detection, choosing appropriate scale is important but not a easy task since it highly depends on the unknown function f (Wang 1995; Antoniadis et al. 2002). In general, wavelet coefficients at coarser scales are more powerful at detecting broad changes as they involve averaging over longer segments. However wavelet coefficients at coarser scales are less sensitive to detecting narrower peaks because of inclusion of neighboring data points that are not relevant to the

peaks. So one can not simply use the coarsest scale to achieve the best results. Wang (1995) commented that the optimal scale is the scale s at which the empirical wavelet coefficient at change point x, $WY(s,x)$, is dominated by $Wf(s,x)$. So for a function with different step sizes, the optimal scales are very likely different. In this dissertation, we propose a test statistic that automatically utilizes all the information across scales to detect change points.

Chapter 3

DETECTING CHANGE POINTS BY MULTISCALE WAVELET PRODUCTS

3.1 Model and Notation

Some notation have been introduced in Section 1.2, but for completeness we will review them before describing the proposed method. Consider n marker loci with known physical locations on a chromosome. Although the locations of these loci may not be evenly spaced along the chromosome, we will only use the relative physical ordering instead of the exact physical locations of these markers. We refer, for example, to the study by Sardy et al. (1999) who observed that treating the sampled loci as if they were evenly spaced performed equally well as wavelet methods that attempted to account for potentially unequal physical distances. Let Y_i denote the observed log-relative intensity ratio for the ith marker locus, for $i = 1, \ldots, n$. Assuming an additive measurement error model for the log-relative intensities, the observed data can be written as

$$Y_i = f(x_i) + \epsilon_i, \tag{3.1}$$

where ϵ_i, $i = 1, \ldots, n$, are independent and identically distributed Gaussian random variable $N(0, \sigma^2)$, and f is a piece-wise constant function reflecting the discreteness in copy number. Suppose there are R change points which lead to $R+1$, $0 \leq R \leq n-1$, nonoverlapping segments. For the rth segment, let c_r denote the index of the end marker in the segment and μ_r be the corresponding copy number. By convention, define $c_0 = 0$, $c_{R+1} = n$, and $c_1 < c_2 < \ldots < c_R$ where the inequalities are strict. The collection of $\{c_r : r = 1, \ldots, R\}$ is called the set of change points and R is the number of change points, the key parameters in a segmentation. Rewriting model (3.1) in

terms of $\{c_r, \mu_r : r = 1, \ldots, R+1\}$, we have

$$Y_i = \mu_r + \epsilon_i \qquad \text{if } c_{r-1} < i \leq c_r.$$

For identifiability of the change points, we assume $\mu_r \neq \mu_{r+1}$. Under the null hypothesis of no change point, the function is then constant $f(x_i) = \mu_0$. Without loss of generality, we assume $\mu_0 = 0$.

The main goal of this research is to detect change points, which include estimating the number of change points, the locations of change points and the mean for each segment.

3.2 Multiscale Products

The wavelet transform measures the extent of local irregularity. A function f is said to be Lipschitz β at x_0 if there exists a constant $K > 0$ such that

$$|f(x_0 + h) - f(x_0)| \leq K|h|^\beta$$

as $h \longrightarrow 0$. Daubechies (1992, page 45-49) stated that if f is differentiable, i.e. Lipschitz β at x_0, then $|Wf(s, x_0)| \leq Ks^{3/2}$, otherwise if f is not Lipschitz β at x_0, $\max_{x \in x_0 + sv}\{|Wf(s,x)|\} \geq Ks^{\beta+1/2}$, where v is within the support of ψ_s. In this thesis, the local irregularity refers to jumps in a step-wise function: the larger the wavelet coefficients, the greater the jumps. It is therefore natural to have test statistics based on the wavelet coefficients and examine whether these wavelet coefficients are much greater than what would have been expected given the amount of noise or variability in the data. In this section we describe the proposed multiscale products starting by recasting the discussion of Section 2.1.3 into the notation of discrete linear filtering as used in Percival et al. (2000).

Let $h = \{h_l : l = -L/2, \ldots, -1, 0, \ldots, L/2 - 1\}$ be a wavelet filter, where L is the width of the filter. Define $h_l = 0$ for $l < -L/2$ and $l \geq L/2$. A wavelet filter must

satisfy the following three conditions:

$$\sum_{l=-L/2}^{L/2-1} h_l = 0, \sum_{l=-L/2}^{L/2-1} h_l^2 = 1 \text{ and } \sum_{l=-\infty}^{\infty} h_l h_{l+2m} = 0$$

for all nonzero integers m. The Haar wavelet filter is $\{h_0 = -1/\sqrt{2}, h_{-1} = 1/\sqrt{2}\}$ and $L = 2$. Then the empirical wavelet coefficient $W_{j,i}$ at scale j for the ith marker locus can be written as

$$W_{j,i} = \sum_{l=-L_j/2}^{L_j/2-1} h_{j,l} Y_{i-l},$$

where $\{h_{j,l}\}$ is the corresponding wavelet filter at scale j and L_j is the width of $\{h_{j,l}\}$, $L_j = (2^j - 1)(L - 1) + 1$. Figure 3.1 presents $\{h_{j,l}\}$ for Haar and Daubechies (D4) wavelet filters at scale $j = 1, 2, 3, 4$. Correspondingly, MODWT wavelet filters at scale j are defined as $\tilde{h}_{j,l} = h_{j,l}/2^{j/2}$. Hereafter, we will simply use $h_{j,l}$ when representing MODWT wavelet filters.

For wavelet coefficients at the two boundaries, we extend the data Y beyond its boundaries Y_1 and Y_n in a symmetric manner: $Y_{n+k} = Y_{n-k}$, where $k = 0, 1, \ldots, n$. This may be preferred over periodic boundary handling as it preserves the continuity of the function.

To determine whether a marker locus is a change point, we could estimate how "extreme" the wavelet coefficients are at marker locus i across various scales under the null distribution of no break points. Now the question is how to combine the wavelet coefficients across these scales to increase the power for detecting break points.

One choice is to take a multiscale product of wavelet coefficients by multiplying MODWT coefficients across adjacent scales. We have also considered other possibilities for combining wavelet coefficients, for example, taking the summation or the maximum, but none seems performing as well as the multiscale product in the simulation studies (Section 3.5). The reason is that signals are typically correlated across adjacent scales while noises are not and so if wavelet coefficients at a true change point propagate through several scales, depending on the size of an aberration. So

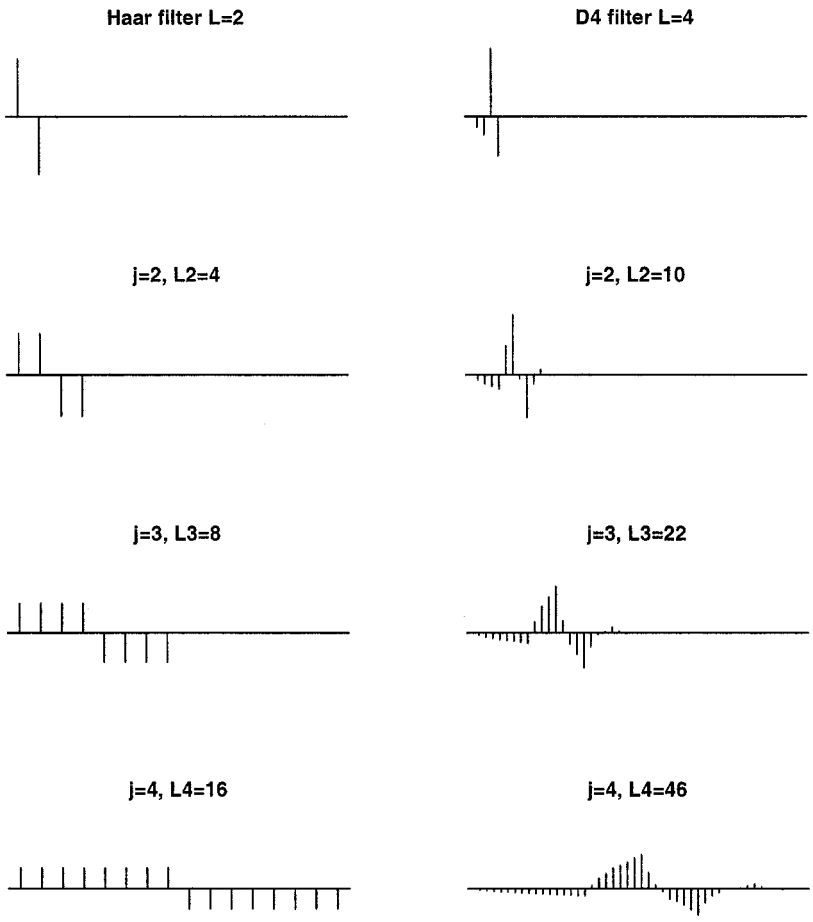


Figure 3.1: Haar and Daubechies (D4) wavelet filters and their $\{h_{j,l}\}$ at higher scales, where L_j is the width of $\{h_{j,l}\}$.

taking the product of wavelet coefficients at the same location across scales enforces the signal while reducing the noise. We will address this further in Section 3.3.

Actually, the multiscale products method had been used in edge detection even before wavelets had been developed (Rosenfeld et al. 1970). Sadler and Swami (1999) used wavelet products across three adjacent scales for detecting edges of steps but failed to control an overall type I error rate and did not adjust for multiple comparison. Bao and Zhang (2003) used two adjacent scale wavelet products for thresholding and recovering the true signal. However, to our knowledge, there is no existing work on how to conduct statistical inferences based on multiscale products, which will be the focus of this work.

A general form of the multiscale product at the ith marker locus is $U^*_{D,i} = \prod_{j\in D} W_{j,i}$, where D is a subset of all the possible scales $\{1,\ldots,J-1\}$ and $J = \log_2 n$. A useful choice for D is the (adjacent) two-scale product

$$U_{j,i} = W_{j,i}W_{j+1,i},$$

because if marker locus i is indeed a change point, the wavelet coefficients at the adjacent scales are most correlated. Now detecting change points is equivalent to determining how far the $U_{j,i}$'s are in the tail of the distribution under the null hypothesis of no change points. It essentially becomes a multiple hypothesis testing problem.

3.3 Test Statistics

In this section we will present a test statistic that will be used in examining whether each marker locus is a change point. First, defining the test statistics leads to address one of the most commonly asked question in the wavelet analysis: which scales to use (Wang 1995; Antoniadis et al. 2002). The truth is that the choice of scale j highly depends on the unknown function f. In general, wavelet coefficients at coarser scale are more powerful at detecting break points as they involve averaging over longer segments. Moreover the number of local maxima that are due to noises decreases

quickly with scale because of increased smoothness. However wavelet coefficients at coarser scales are less sensitive in detecting narrower regions of aberrations in CGH. So one can not simply use the coarsest scale to achieve the best power.

Figure 3.2 shows the 2-scale wavelet products U_j (right panel) and single-level wavelet coefficients W_j (left panel) from simulated signals with additive white noise. We can see that the wavelet transforms are getting smoother as scale increases. As we expected, the local maxima at the narrow aberration region, marker locus 100 and 110, start to diminish after scale 5, while the local maxima at wide regions stand further out as scale increases. Therefore, for this simulated data set, we see that the "optimal" scales for detecting these 6 change points are very likely different, such that scale 4 might be "optimal" for the first 2 change points at narrow region and the scale 5 or beyond for wide regions.

3.3.1 Test Statistic at One Scale

One possible solution is to look at one scale at a time then combine the detected change points across scales. The test statistics at scale j are the $U_{j,i}$'s and the hypotheses are $H_i : U_{j,i} = 0$ versus $K_i : U_{j,i} > 0$. Here we consider a parametric approach of detecting the change points based on an assumption that ϵ is i.i.d. normal with mean 0 and variance σ^2. Under model (3.1), we have

$$W_{j,i} = \sum_{l=-L_j/2}^{L_j/2-1} h_{j,l} f(i-l) + \sum_{l=-L_j/2}^{L_j/2-1} h_{j,l}\epsilon_{i-l}.$$

Let's consider the null situation that the ith marker locus is not a change point and there are no change points around the ith marker locus. In this situation, the first term is equal to 0 under the condition $\int_{-\infty}^{\infty} \psi(u)du = 0$; i.e. $\sum_{l=-L_j/2}^{L_j/2-1} h_{j,l} = 0$. Since $\epsilon_i \sim N(0, \sigma^2)$, $W_{j,i}$ will also be normally distributed with mean zero and variance σ_j^2, where

$$\sigma_j^2 \equiv Var(W_{j,i}) = \sum_{l=-L_j/2}^{L_j/2-1} h_{j,l}^2 Var(\epsilon_{i-l}) = \sum_{l=-L_j/2}^{L_j/2-1} h_{j,l}^2 \sigma^2 = \sigma^2/2^j.$$

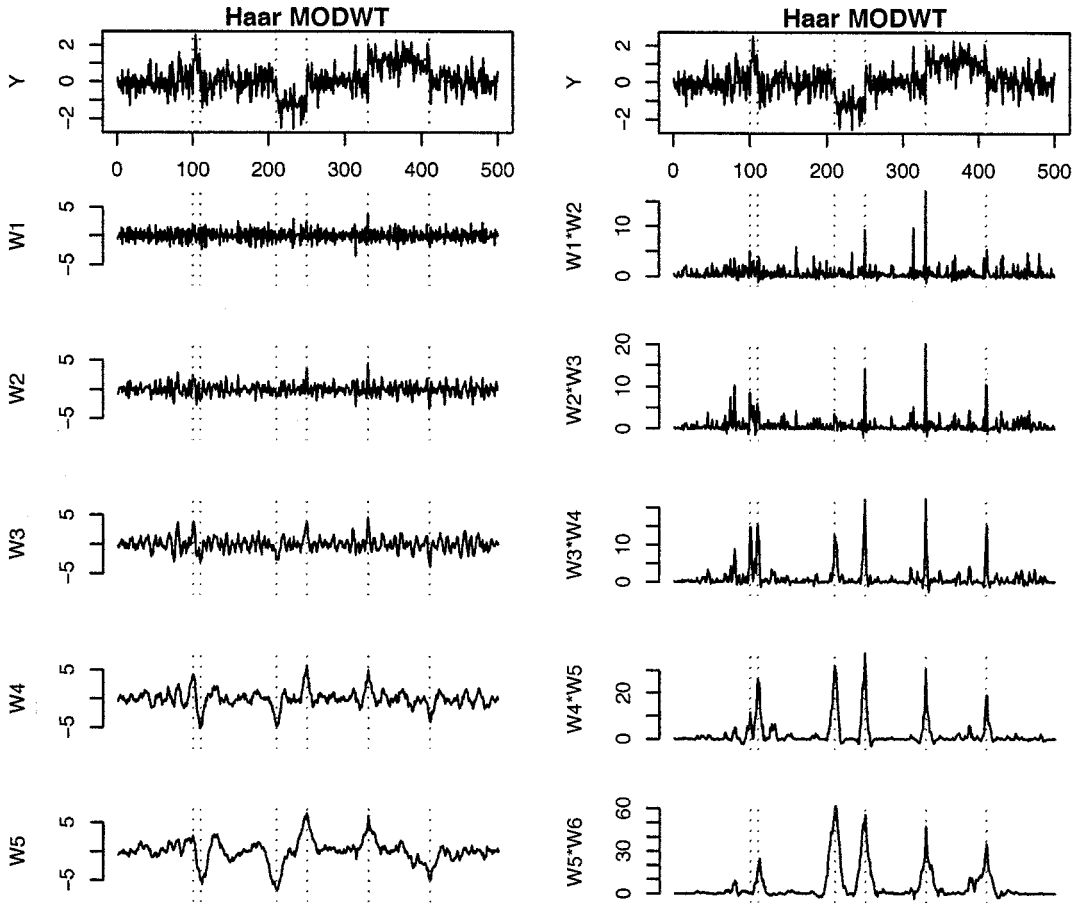


Figure 3.2: Wavelet coefficients W_j (left panel) and corresponding 2-scale products U_j (right panel) across scales. The top row is the simulated signal Y with the dashed lines indicating the step function f. The vertical dotted lines indicate the change points at locations 100, 110, 210, 250, 330, 410. $\sigma = 0.5$ and $n = 500$.

Then $W_{j,i}$ and $W_{j',i}$ are bivariate normals with mean 0 and variance-covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_j^2 & \rho_{j,j'}\sigma_j\sigma_{j'} \\ \rho_{j,j'}\sigma_j\sigma_{j'} & \sigma_{j'}^2 \end{pmatrix},$$

In general, we can derive that

$$\begin{aligned} Cov(W_{j,i}, W_{j',i}) &= Cov\left(\sum_{l=-L_j/2}^{L_j/2-1} h_{j,l}Y_{i-l} \sum_{l'=-L_j'/2}^{L_j'/2-1} h_{j',l'}Y_{i-l'}\right) \\ &= \sum_{l=-L_j/2}^{L_j/2-1}\sum_{l'=-L_j'/2}^{L_j'/2-1} h_{j,l}h_{j',l'}Cov(Y_{i-l}, Y_{i-l'}) \end{aligned}$$

Under assumption of Y being i.i.d., we have $\rho_{j,j'} = 2^{\frac{j+j'}{2}}\sum_{l=-L/2}^{L/2-1} h_{j,l}h_{j',l}$, with $L = \min(L_j, L_{j'})$.

The probability density function $g(u)$ of a product of bivariate normals with zero means is given by Miller (1964):

$$g(u) = \frac{1}{\pi\sigma_j\sigma_{j'}\sqrt{1-\rho_{j,j'}^2}} e^{\rho_{j,j'}\sigma_j\sigma_{j'}v} K_0(\sigma_j\sigma_{j'}|\Sigma|) \tag{3.2}$$

where $u = W_{j,i}W_{j',i}$, K_0 is the modified Bessel function of the second kind and order zero, and Σ is a covariance matrix for $W_{j,i}$ and $W_{j',i}$. Figure 3.3 shows the probability density functions for $N(0,1)$ and $g(u)$, the product of a bivariate normal with mean 0 and variance 1 and $\rho = 1/\sqrt{2}$. We can see that the $g(u)$ is highly right-skewed.

An advantage of this approach is the computational efficiency as the pdf of the test statistic under the null is known. So we could compute the tail probability using numerical integration, i.e. $p_i = \int_{u_i}^{\infty} g(u)du$, where u_i is the observed 2-scale wavelet product at location i. Then the unadjusted p-values can be converted to adjusted p-values by taking into account the multiple comparison using either single-step procedures such as the Bonferroni correction, Sidak procedure, or step-down procedures such as the Holm procedure or min P procedure (Dudoit et al. 2003; Ge et al. 2003). One limitation of this approach is that the method tends to be

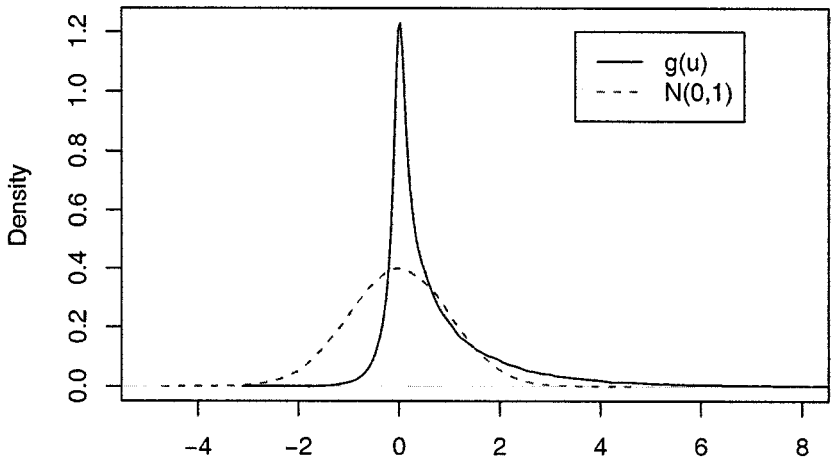


Figure 3.3: Probability density function of N(0,1) (dotted line) and product of bivariate normal (0, 1) with $\rho = 1/\sqrt{2}$ (solid line).

conservative because of a large number of tests across scales. For example, when n=500 markers, for each scale j, we have 500 test statistics $U_{j,i}$'s. Suppose we look at four scales from $j = 3$ to 6. Then we will need to account for a total of 2000 tests to obtain adjusted p-values which can be very conservative especially for single-step procedures, where correlation among test statistics is not accounted for.

3.3.2 Test Statistics Over Multiple Scales

In order to be able to accurately control the FWER and estimate adjusted p-values while having the capacity to detect both narrow and wide aberration regions, we combine $U_{j,i}$ across scales by taking a maximum of standardized $U_{j,i}$ over $j = 2, \ldots, J-1$, i.e. $M_i = \max_{j=2,\ldots,J-1}\{U_{j,i}\}$. So the problem of detecting the change points becomes

testing

$$H_i : M_i = 0 \text{ versus } K_i : M_i > 0. \tag{3.3}$$

For each location i, we will estimate the probability of observing a value equal to or greater than the observed M_i given the null hypothesis of no change points is true. Note that for the 2-scale wavelet product, the test statistic M_i is always positive regardless whether the mean changes are positive or negative at change point i.

3.4 Multiple Testing

In this section we will describe a procedure for obtaining the adjusted p-values accounting for multiple comparisons, for which research has been very active recently. Ge et al. (2003) gave a comprehensive review of multiple testing methods. Among all the multiple testing procedures, controls of family wise error rate (FWER) or false discovery rate (FDR) are perhaps the most commonly used. The FWER is defined as the probability of at least one type I error, i.e., FWER=Pr(number of false positives > 0). The FDR is defined as the expected proportion of type I errors among the rejected hypotheses (Benjamini et al. 1995). Under global null hypotheses, i.e., no change point in the function f, FDR=FWER. Generally speaking, controlling the FWER tends to yield more conservative results than controlling the FDR.

Since both the marginal and joint distributions of the test statistics M_i are unknown, we use resampling methods to estimate both raw and adjusted p-values. However, generating a null distribution is not trivial for our data, because the function f is unknown. We consider two approaches for generating the null distribution.

3.4.1 Approach I: Permuting $\hat{\epsilon}$

The second approach is to use the resampling method, which avoids the normal assumption for the error distribution. A simple permutation of the observed Y_i works if the non-zero segments are only a small proportion of the whole region. Otherwise,

the permutated Y's, a mixture of noises and aberrations with non-zero means, would not yield a proper null distribution that reflects the true error distribution. We thus propose to permute $\hat{\epsilon}_i = Y_i - \hat{f}_i$ to generate the null distribution where $\hat{f}_i$ is a robust estimator using lowess, a locally weighted regression (Cleveland 1979). It is worth noting that the empirical distribution of error ϵ is highly dependent on the estimated $\hat{f}$; that is, the closer $\hat{f}$ to the observed $\{Y_i\}$, the smaller the errors. The key parameter in lowess smoothing is the size of the smoothing window given as the proportion of the data that is included in the smoothing window. The larger the window size the smoother $\hat{f}$. In Section 5.3 we test the performance of the wavelet method using two different window size 0.05 and 0.1 by simulation and find that the two window sizes give very close results , although size 0.1 is slightly better than size 0.05.

3.4.2 Approach II: Permuting W_1

Alternatively we could permute the wavelet coefficients at the finest level W_1 to estimate the null distribution of the test statistics. Since the W_1 are the scaled differences of two adjacent Y_i's, this would give a close approximate estimation of the null distribution if ϵ is normally distributed given that the number of change points is relatively small compared to n. However, when ϵ has heavy tail with large kurtosis, permuting W_1 will be anti-conservative, in other words it will overestimate R, the number of change points. To see this, suppose X and Y are i.i.d. with mean 0 and variance σ^2. Let $Z = X - Y$, then

$$f_Z(z) = \int_{-\infty}^{\infty} f(y)f(z+y)dy$$

In general the distribution of variable Z is not identical to the distribution of X unless X is normally distributed. To see the relationship between Z and X, we look at the first four moments: mean, variance, skewness and kurtosis for X and Z. We can show that the skewness, α_1, of Z is 0 and the kurtosis, α_2, of Z is exactly half of the kurtosis of X. Below is the derivation. Let μ_n and $\mu_{n,z}$ denote the nth central

moment of a random variable X and Z, respectively. We have

$$\mu_{3,Z} = EZ^3 = E(X-Y)^3 = E(X^3 - 3X^2Y + 3XY^2 - Y^3) = 0,$$

so $\alpha_1 = \frac{\mu_{3,Z}}{(\mu_{2,Z})^{3/2}} = 0$

$$\mu_{4,Z} = EZ^4 = E(X-Y)^4 = E(X^4 - 4X^3Y - 4XY^3 + 6X^2Y^2 + Y^4) = 2\mu_4 + 6\mu_2^2,$$

so

$$\alpha_2 = \frac{\mu_{4,Z}}{\mu_{2,Z}^2} - 3 = \frac{2\mu_4 + 6\mu_2^2}{4\mu_2^2} - 3 = \frac{\mu_4 - 3\mu_2^2}{2\mu_2^2} = \alpha_{2,X}/2$$

In Table 3.1, we list the skewness and kurtosis of both X and Z for normal, t and double exponential distributions. Figure 3.4 shows the density functions for t distributions with df=3 and 5 for scaled Z and X. The right panel is the zoomed right tails of density functions. Scaled Z has heavier tails than X when the degree of freedom is 3 or less.

Table 3.1: Skewness and Kurtosis

	X		Z	
Distribution of X	Skewness	Kurtosis	Skewness	Kurtosis
$N(\mu, \sigma^2)$	0	0	0	0
$t(\nu\ ^a)(\nu > 4)$	0	$\frac{6}{\nu-4}$	0	$\frac{3}{\nu-4}$
Double Exp$(0, \sigma^2)$	0	3	0	3/2

[a] ν is degree of freedom

Since the accuracy of the approximation of the null distribution by permuting W_1 highly depends on the kurtosis of the null distribution, we recommend to use this approach when sample kurtosis is relatively small, i.e. close to 3. For the t distribution, when the degree of freedom is greater than 4, permuting W_1 is a good choice. Otherwise, we need to use approach II by permuting $\hat{\epsilon}$.

After the null distribution of ϵ is estimated, we compute adjusted p-values using step-down maxT permutation algorithm proposed by Westfall and Young (1993). Following is the algorithm for calculating adjusted p-values. First for observed data, we compute the test statistic for each marker locus, i.e., $M_i = \max_j(U_{j,i})$ for $j = 2, 3, \ldots J_0 - 1$. Order the observed test statistics such that $M_{s_1} \geq M_{s_2} \geq \ldots \geq M_{s_n}$. For the bth permutation, $b = 1, \ldots, B$:

1. Permute the W_1 or $\hat{\epsilon}_i = Y_i - \hat{f}_i$, where $\hat{f}$ is the lowess estimator.

2. Compute test statistics $M^*_{i,b}$ for each marker based on permuted data.

3. Compute the successive maxima of test statistics by

$$\begin{aligned} u_{n,b} &= M^*_{s_n,b} \\ u_{i,b} &= \max(u_{i+1,b}, M^*_{s_i,b}) \qquad \text{for } i = n-1, \ldots, 1 \end{aligned}$$

   such that $u_{i,b} = \max_{l=i,\ldots,n} M^*_{s_l,b}$.

4. Repeat the above steps B times and the adjusted p-values are estimated by

$$p^*_{s_i} = \frac{\#\{b : u_{i,b} \geq M_{s_i}\}}{B} \qquad \text{for } i = 1, \ldots, n$$

   the raw p-values are estimated by

$$p_i = \frac{\#\{b : M^*_{i,b} \geq M_i\}}{B} \qquad \text{for } i = 1, \ldots, n.$$

Note that $\{p^*_i\}$ is a set of adjusted p-values for testing 3.3 in the sense that

$$pr(p^*_i \geq \alpha | H_0, 1 \leq i \leq n) \geq 1 - \alpha$$

where H_0 denotes global null– no change point.

3.5 Power Comparison for Wavelet Coefficients, 2-scale Sum and 2-scale Product

As mentioned in Section 3.2, although it is believed that multiscale products tend to enforce the signal and reduce the noise, no existing research has examined the power of multiscale products for detecting change points. Hence it is unknown exactly how powerful a 2-scale product is compared to two other alternatives: 2-scale sum and single-level wavelet coefficients. In this section, we will compute the power functions for wavelet coefficients, 2-scale sum and 2-scale product using the Haar wavelet. The same algebra could also be done for other wavelet families. Since our focus is on the Haar wavelet, we'll not pursue other wavelet families further. Here we consider two cases for power calculations. The first case is to calculate the power of detecting the change point at locus i at scale j, so we consider only a single test at location i. The second case is that we have two change points far apart (square wave) and would like to calculate the power of detecting these two change points given the fixed overall significant level. The second case is really the main interest here, however, it is intractable to obtain power analytically since it involves the joint distributions of n correlated non-Gaussian variables for 2-scale products. Instead we used Monte Carlo simulations to obtain power for the second case.

3.5.1 Case I: A Single Test

Consider the case in which there is only one change point in the middle of the region. This is perhaps the simplest situation for change point detection. For this situation, the distributions for all three test statistics, 2-scale sum, 2-scale product and single-level wavelet coefficients, can be easily derived, so we can obtain power analytically or using numerical integration.

In what follows, we'll derive power functions for three test statistics. Let d be the mean of the aberration region. As before, the baseline is assumed equal 0. At

the change point, the expectations of the wavelet coefficients are $d/2$ for all scales assuming the number of marker loci is large enough. First we derive the power for individual-level wavelet coefficients. Under the null hypothesis $H_0 : W_{j,i} = 0$. Under the alternative $H_1 : W_{j,i} \neq 0$, so it is a two-sided test. Then the power function of single-level wavelet coefficients $W_{j,i}$ is

$$\begin{aligned}
\beta(d) &= P_d(\frac{2^{\frac{j}{2}}W_{j,i}}{\sigma} > z_{\alpha/2}) \\
&= P_d(\frac{2^{\frac{j}{2}}W_{j,i}}{\sigma} - \frac{2^{\frac{j-2}{2}}d}{\sigma} > z_{\alpha/2} - \frac{2^{\frac{j-2}{2}}d}{\sigma}) \\
&= P_d(Z > z_{\alpha/2} - \frac{2^{\frac{j-2}{2}}d}{\sigma}) \\
&= 1 - \Phi(z_{\alpha/2} - \frac{2^{\frac{j-2}{2}}d}{\sigma}) \qquad (3.4)
\end{aligned}$$

where Φ is the cumulative distribution function of the standard normal distribution, and Z is standard normal with $\alpha/2 = P(Z > z_{\alpha/2})$. So in order to detect the change point at location i with power β, the finest resolution for individual-level wavelet coefficients is

$$j_w = 2 + 2\log_2 \frac{(z_{\alpha/2} + z_{1-\beta})\sigma}{d}$$

Next we derive the power function for the 2-scale sum test statistic. Define a standardized version of 2-scale sum

$$S_j(i) = \frac{W_{j,i}}{\sigma_j} + \frac{W_{j+1,i}}{\sigma_{j+1}}$$

It is straightforward to obtain $Var(S_j(i)) = 2+\sqrt{2}$, where $\sqrt{2}$ is the extra variance due to the correlation of the wavelet coefficients between the two adjacent levels. Under the null hypothesis of no change point, we have $S_j(i) = 0$, under the alternative $E(S_j(i)) = \frac{\sqrt{2^{j-2}}(1+\sqrt{2})d}{\sigma}$. Then the power function of 2-scale sum at scale j is

$$\begin{aligned}
\beta(d) &= P_d(\frac{S_j(i)}{\sqrt{2+\sqrt{2}}} > z_{\alpha/2}) \\
&= P_d(\frac{S_j(i)}{\sqrt{2+\sqrt{2}}} - \frac{\sqrt{2^{j-2}}(1+\sqrt{2})d}{\sqrt{2+\sqrt{2}}\sigma} > z_{\alpha/2} - \frac{\sqrt{2^{j-2}}(1+\sqrt{2})d}{\sqrt{2+\sqrt{2}}\sigma})
\end{aligned}$$

$$
\begin{aligned}
&= P_d(Z > z_{\alpha/2} - \frac{\sqrt{2^{j-2}}(1+\sqrt{2})d}{\sqrt{2+\sqrt{2}}\sigma}) \\
&= 1 - \Phi(z_{\alpha/2} - \frac{\sqrt{2^{j-2}}(1+\sqrt{2})d}{\sqrt{2+\sqrt{2}}\sigma}) \qquad (3.5)
\end{aligned}
$$

The finest resolution for 2-scale sum is

$$j_s = 2 + 2\log_2 \frac{(z_{\alpha/2} + z_{1-\beta})\sigma c}{d}$$

where $c = (\sqrt{2}-1)\sqrt{2+\sqrt{2}} = 0.765$. We observe that $S_j(i)$ is more powerful than individual-level wavelet coefficients $W_{j,i}$ for all σ and d, and $j_s = j_w - 0.77$. Note that the power for both $W_{j,i}$ and $S_j(i)$ are functions of d/σ which gives a justification of using d/σ to define SNR. We can easily see that the power is greater with larger SNR.

Last we derive the power for 2-scale wavelet product. We donot have a closed form for the distribution function of a 2-scale product $U_j(i)$ under the alternative. Instead we use numerical integration to calculate the power. Define a standardized version of 2-scale product

$$U_j(i) = \frac{W_{j,i}}{\sigma_j}\frac{W_{j+1,i}}{\sigma_{j+1}}$$

The pdf of $u = xy$ can always be written as

$$f(u) = \int_{-\infty}^{\infty} f(x, \frac{u}{x})\frac{1}{|x|}dx$$

with $x = \frac{W_{j,i}}{\sigma_j}$ and $y = \frac{W_{j+1,i}}{\sigma_{j+1}}$. We have already shown in Section 2.2 that x and y are bivariate normals with mean vector$(\frac{2^{\frac{j-2}{2}}d}{\sigma}, \frac{2^{\frac{j-1}{2}}d}{\sigma})$ and variance-covariance matrix

$$\begin{bmatrix} 1 & \rho_{j,j+1} \\ \rho_{j,j+1} & 1 \end{bmatrix}$$

The power function is

$$
\begin{aligned}
\beta(d) &= P_d(U_j(i) > u_\alpha) \\
&= \int_{u_\alpha}^{\infty}\int_{-\infty}^{\infty} f(x, \frac{u}{x})\frac{1}{|x|}dxdu \qquad (3.6)
\end{aligned}
$$

where $\alpha = P(U > u_\alpha)$ with pdf of U is given in (3.2).

In Table 3.2 we list the finest scale to obtain 90% power at 0.01 significant level using 3 test statistics. 2-scale sum and 2-scale product have the same results while single-level coefficient W_j need larger scale. Since the scale j can only be integer from 1 to J, j might be rounded to the smallest integer greater than the computed value.

Table 3.2: Finest scale for wavelet coefficients W_j, 2-scale sum S_j and 2-scale product U_j at significant level 0.01 and power 0.90.

SNR	scale		
	j_w	j_s	j_u
1.0	5.9	5.1	5.1
1.5	4.7	4.0	4.0
2.0	3.9	3.1	3.1
2.5	3.3	2.5	2.5
3.0	2.7	2.0	2.0
3.5	2.3	1.5	1.5

Figure 3.5 displays the power functions for all three methods at scales 2, 3 and 4, respectively. Again we can see that the power curves for 2-scale sum and 2-scale product are almost identical and are much higher than the power curve at lower level coefficient W_j. We also observe that the power curves for 2-scale sum and 2-scale product are slightly lower than the curve at higher level coefficient W_{j+1}. Hence without accounting for multiple testing, the 2-scale sum S_j and 2-scale product U_j have almost the same power of detecting change point but are slightly less powerful than W_{j+1}. So the combined test is close to the single scale test with higher power. This can be considered as paying a price of small power loss with gain of robustness in terms of which scale to use.

3.5.2 Case II: Multiple Tests

In the second case, we compare the uses of single-level wavelet coefficients, the two-scale sum and the two-scale product test statistics for finding change points while accounting for the multiple testing. We simulated a square wave signal encompassing 200 marker loci in the center of a total of 500 markers and find that the two-scale product test statistics yields better power. Here the power was defined as the probability of correctly detecting two change points. As we mentioned before, it is not tractable to obtain power analytically for 2-scale product. Instead we use Monte Carlo simulations to obtain the powers. The data is simulated using model(3.1) with $\epsilon \sim N(0, 0.5^2)$. The aberration mean μ is chosen from 0.5 to 2, increasing by 0.1 each time so that the SNR ranges from 1 to 4. The adjusted p-values are estimated by permuting scaled W_1. Figure 3.6 shows the power curves for all three methods at scale 4, 5 and 6. The 2-scale product test statistic has the highest power among three methods regardless of SNR values. The powers for the two-scale product $(W_5 W_6)$ and the two-scale sum $(W_5 + W_6)$, single level wavelet coefficients W_5 and W_6 are 0.78, 0.53, 0.12, 0.68, when the signal to noise ratio is 1.5, and 0.98, 0.94, 0.60, 0.98 when the signal to noise ratio is 2.0, respectively. The signal to noise ratio (SNR) is defined as mean of the aberration region divided by the standard deviation of the additive Gaussian noise, i.e. $SNR = \mu/\sigma$. Based on these results, we'll use 2-scale products across scales to detect change points for the rest of the dissertation.

3.6 Local Maximum

Given the 2-scale product statistics at each marker locus, it would be natural to simply apply some standard multiple testing procedures for all n test statistics to control the overall type I error rate. However due to the high redundancy of the MODWT wavelet coefficients, the wavelet coefficients are autocorrelated along the chromosome with the lag increasing with the scale. So a direct application of multiple testing procedures

detects not only the true change point but also several marker loci that are around the true change point, depending on the width of μ_r and the scales used in the test statistics. To avoid this, we propose to test locations at which local maxima occur. Figure 3.7 illustrates the idea of testing at local maxima. The bottom plot is the $-\log_{10}(p)$ vs location and we can see that the marker loci around change points have very small p-values which might be falsely estimated as change points if ignoring local maxima.

Defining a local maximum is not obvious for a noisy discrete signal, but one might proceed by using a wavelet transform. This connection can perhaps better described for a continuous wavelet transform first (Mallat and Hwang (1992);Mallat (1999)). Suppose that $\theta(u)$ is a differentiable smoothing function, where $\int \theta(u)du = 1$ and $\theta(u) \to 0$ as $u \to \infty$. Consider a wavelet function $\psi = d\theta/du$ and set $\theta_s(u) = \frac{1}{s}\theta(\frac{u}{s})$. Then the wavelet transform can be expressed as

$$Wf(s,x) = s\frac{d}{dx}(f * \theta_s)(x),$$

which is a first-order derivative of f when smoothed by θ_s where the degree of smoothing depends on scale s. Now set $s = 2^j$. Based on this concept, Mallet et al. (1992) defined a scale-j local extreme for f at x_0 when $dWf(s,x)/dx$ has a zero-crossing at $x = x_0$. Hence, this scale-based local extreme can be found by taking the wavelet transformation of wavelet coefficients $\{W_{j,1}, \ldots, W_{j,n}\}$ at scale j and looking for zero-crossings.

We apply this approach to the 2-scale product $U_{j,x}$ and define the local maximum at the point x_0 to be where the wavelet transformation of $U_{j,x}$ crosses zero from positive to negative. Note that $U_{j,x}$ has local maximum at t_0 if both $W_{j,x}$ and $W_{j+1,x}$ have local maxima at t_0. Simply speaking detecting change points becomes a search for local maxima across scales that pass a certain threshold. We show that the change points detected by using the local maximum of $U_{j,x}$ are consistent with the true change points with a rate of $1/n$. So one can think of multiscale products as

a way of simultaneously looking at wavelet coefficients across the scales. Figure 3.8 illustrates the idea of using a wavelet transform to detect local maxima.

change points are estimated to be the locations of local maxima such that the corresponding adjusted p-values are smaller than certain values, say 0.01. Let $\hat{R}$ be the number of those local maxima whose adjusted p-values, $p^* < \alpha$ and $\{\hat{c_r} : r = 1, \ldots, \hat{R}\}$ be their locations. So we simultaneously estimate R and $\{c_r : r = 1, \ldots, R\}$.

The following is the summary of the proposed algorithm for detecting the change points and segmentation.

1. Perform a MODWT on the observed data Y up to a given scale J_0, say 6, to obtain the corresponding empirical wavelet coefficients $W_{J_0,i}$, then estimate the standard deviation σ by median absolute deviation of the wavelet coefficients at the finest scale, i.e. $\hat{\sigma} = \sqrt{2}median(|W_{1,i}|)/0.6745$, adjusting 0.6745 for asymptotically normal consistency.

2. Then calculate point-wise products of empirical wavelet coefficients at two adjacent scales, $U_j = W_j W_{j+1}, j = 2, 3, \ldots J_0 - 1$. The test statistic for location i is the maximum of $U_j(i)$ over $j = 2, \ldots, J_0 - 1$, i.e. $M_i = \max_{j=2,\ldots,J_0-1}\{U_j(i)\}$.

3. Obtain the local maxima for M_i and denote by T_i the local maxima at location i.

4. Estimate the null distribution by permuting W_1 or $\hat{\epsilon}$ and obtain adjusted p-values for each M_i using step-down maxT algorithm.

5. change points are estimated to be the locations of local maxima such that the corresponding adjusted p-values are smaller than a prespecified threshold, say 0.01.

6. The whole chromosome is segmented using estimated change points. Estimated mean for each segment is the sample average of Y_i within the segment.

3.7 Consistency

In this section we will show the consistency of the estimated change points. To do so, we need to re-define model (3.1) so that we can apply the CWT. Model (3.1) is closely related to the following white noise model

$$dZ(x) = f(x)dx + \tau dV(x),\ x \in [0,1],$$

where V is a standard Brownian motion, τ is a noise parameter, and f is an unknown step function. To see the connection between the nonparametric regression model (3.1), which is $Y_i = f(x_i) + \epsilon_i$, and the white noise model, define a process $\{Z_n(x) : x \in [0,1]\}$ via $x_0 = 0, Z_n(0) = 0$ and $Z_n(x_i) = \sum_{j=1}^{i} Y_j$ for $i = 1, \ldots, n$. Then Z_n is a white noise process with the function $f_n(x) = \sum_{j=1}^{i} f(x_j)$ and the standard Brownian motion process for $x_i \leq x < x_{i+1}$, and $\tau = \sigma n^{-1/2}$ (Wang 1995). In other words, for $x_i \leq x < x_{i+1}$,

$$Z_n(x) = \frac{1}{n}\sum_{j=1}^{i} Y_j = \frac{1}{n}\sum_{j=1}^{i} f(x_j) + \frac{\sigma}{\sqrt{n}}\frac{1}{\sqrt{n}}\sum_{j=1}^{i}\epsilon_j^*,$$

where ϵ^* are iid standard normal. So we can see that the processes Z_n and Z converges as $n \to \infty$.

Wang (1995) proved the following theorem that with probability approaching to 1, the maximum of absolute values of wavelet coefficients at scale s, $|WY(s,x)|$, occurs only within $supp(\psi_{s,x})$, the support of $\psi_{s,x}$, around the true change point θ. We re-state his proof here for the completeness.

Theorem 1. Suppose f has one change point at θ and is differentiable elsewhere. Let ψ denotes a mother wavelet for family $\{\psi_{s,x}\}$. $Pr\{s^{-1}(\hat{\theta} - \theta) \in supp(\psi)\} \to 1$, as $n \to \infty$, where $supp(\psi)$ denotes the support of ψ, θ is the true change point in f and

$$\hat{\theta} = \arg\max_{0 \leq x \leq 1} \{|WY(s,x)|\}$$

Proof: Let K be a positive constant and $K < \infty$. Let η be any constant greater than 1 and s be of order $\tau^2|\log\tau|^\eta$ as $n \to \infty$. If f is differentiable at all points except at θ, by Theorem 2.9.1 of Daubechies (1992, p.45), we obtain that, for all (s, x) with $(\theta - x)/s \notin supp(\psi)$,

$$|Wf(s,x)| \leq Ks^{3/2} \tag{3.7}$$

For $(\theta - x)/s \in supp(\psi)$, by Theorem 2.9.3 and 2.9.4 of Daubechies (1992, p.49), we have

$$\max|Wf(s,x)| \geq Ks^{\beta+1/2}, \tag{3.8}$$

since f is not Lipschitz continuous at θ. Wang also proved that for all x and small s,

$$|WV(s,x)| \leq K|\log s|^{1/2}. \tag{3.9}$$

The above three inequalities (3.7),(3.8) and (3.9) together imply that

$$\begin{aligned} \max|WY(s,x)| &\geq K(s^{1/2} - \tau|\log s|^{1/2}) \\ &\geq K\tau(|\log\tau|^{\eta/2} - |\log\tau|^{1/2}) \\ &\geq K\tau|\log\tau|^{\eta/2}(1 - |\log\tau|^{(1-\eta)/2}) \end{aligned} \tag{3.10}$$

and, for all (s, x) with $(\theta - x)/s \notin supp(\psi)$, we have

$$|WY(s,x)| \leq Ks^{3/2} + K\tau|\log s|^{1/2} \leq K\tau|\log\tau|^{1/2} \tag{3.11}$$

By (3.10) and (3.11) we can see that as $\tau \to 0$ that is $n \to \infty$, $Pr\{s^{-1}(\hat{\theta} - \theta) \in supp(\psi)\} \to 1$.

So by Theorem 1, with probability approaching to 1, the local maximum of absolute values of wavelet coefficients at scale s, $|WY(s,x)|$, only occurs within the support of $\psi_{s,x}$ around true change points θ. However the exact location for local maximum will depend on the shape of the function $f(x)$ and the wavelet function ψ. In the following, we will show that for a step function f and Haar wavelet, the local maxima

actually converge to the true change points as $n \to \infty$. Furthermore, the change point estimator is consistent. First we need to assume for each $r = 1, \ldots, R, \mu_r \neq \mu_{r+1}$ for identifiability. Then assume for larger n, the change points are far enough from each other such that

$$\lim_{n\to\infty} c_r/n = \tau_r \text{ for } 1 \leq r \leq R$$

Let $\hat{\tau}_r = \hat{c}_r/n$, where $\hat{c}_r$ are the index locations of local maxima with adjusted p-values $p^* < \alpha$.

Corollary 1: For a step function f and Haar wavelet, $\hat{\tau}_r - \tau_r = o_p(1)$, for $r = 1, \ldots, R$. That is, $\hat{\tau}_r$ is a consistent estimator.

Proof: Let $d_r = \mu_{r+1} - \mu_r, d_r \neq 0$. By Theorem 1, it suffices to show that, $Pr(\max_{i\in I}(M_i) = M_{c_r}) \to 1$, as $n \to \infty$, for $I = \{i : c_r - 2^{J_0-1} \leq i \leq c_r + 2^{J_0-1}\}$. We have

$$\begin{aligned} \max_{i\in I}(M_i) &= \max_{i\in I} \max_{j=2,\ldots,J_0-1} \{U_j(i)\} \\ &= \max_{j=2,\ldots,J_0-1} \max_{i\in I} \{W_{j,i}W_{j+1,i}\} \end{aligned}$$

Now we need to show that for each level j, the local maximum of $\{U_j(i)\}$ within I converges to c_r in probability. Then the location of the local maximum of $\{M_i\}$ within I converges to c_r at some level j. First we show that the location of local maximum of $\{|W_{j,i}|\}$ converges to c_r as $n \to \infty$.

$$\begin{aligned} W_{j,i} &= Wf(x_i) + W\epsilon_i \\ &= h_j * f(x_i) + h_j * \epsilon_i \\ &= \sum_{l=-2^{j-1}}^{2^{j-1}-1} h_{j,l} f(i-l) + \sum_{l=-2^{j-1}}^{2^{j-1}-1} h_{j,l}\epsilon_{i-l} \\ &= \left\{ \sum_{l=0}^{2^{j-1}-1} f(i-l) - \sum_{l=1}^{2^{j-1}} f(i-l) + \sum_{l=0}^{2^{j-1}-1} \epsilon_{i-l} - \sum_{l=1}^{2^{j-1}} \epsilon_{i-l} \right\} 2^{-j} \\ &= \left(\frac{d_r}{2} - \frac{|i-c_r|d_r}{2^j} \right) 1_{|i-c_r|\leq 2^{j-1}}(i) + \frac{\sum_{l=0}^{2^{j-1}-1} \epsilon_{i-l} - \sum_{l=1}^{2^{j-1}} \epsilon_{i-l}}{2^j} \end{aligned}$$

$$= \left(\frac{d_r}{2} - \frac{|i-c_r|d_r}{2^j}\right) 1_{|i-c_r|\leq 2^{j-1}}(i) + o_p(1) \tag{3.12}$$

Therefore, $Pr(\max_{i\in I}(|W_{j,i}|) = |W_{j,c_r}|) \to 1$, for $|i-c_r| \leq 2^{j-1}$, as $n \to \infty$ for large j.

$$\begin{aligned} U_{j,i} &= W_{j,i}W_{j+1,i} \\ &= \left\{\left(\frac{d_r}{2} - \frac{|i-c_r|d_r}{2^j}\right) 1_{|i-c_r|\leq 2^{j-1}}(i) + o_p(1)\right\}\left\{\left(\frac{d_r}{2} - \frac{|i-c_r|d_r}{2^{j+1}}\right) 1_{|i-c_r|\leq 2^{j}}(i) + o_p(1)\right\} \\ &= \left(\frac{d_r}{2} - \frac{|i-c_r|d_r}{2^j}\right)\left(\frac{d_r}{2} - \frac{|i-c_r|d_r}{2^{j+1}}\right) 1_{|i-c_r|\leq 2^{j-1}}(i) + o_p(1) \end{aligned} \tag{3.13}$$

Hence, $Pr(\max_{i\in I}(U_{j,i}) = U_{j,c_r}) \to 1$, as $n \to \infty$.

Therefore it proves the location of the local maxima of M_i converges to c_r at some j. By the definition of $\hat{c_r}$, $Pr(\hat{\tau_r} = \tau_r) \geq 1-\alpha$. So as $\alpha \to 0$, $\hat{\tau_r} \to \tau_r$.

Next we will show that the number of estimated change points, $\hat{R}$, is $1-\alpha$ consistent to the true number R.

Corollary 2. $Pr(\hat{R} = R) \to 1-\alpha$ as $n \to \infty$, where .

Proof: By the definitions of $\hat{R}$ and Corollary 1 we can see that, $Pr(\hat{R} \geq R) \to 1-\alpha$ and $Pr(\hat{R} \leq R) \to 1-\alpha$. Therefore $Pr(\hat{R} = R) \to 1-\alpha$ as $n \to \infty$.

3.8 Summary

In this chapter we proposed a test statistic for detecting change points and provided two non-parametric approaches to estimate individual p-value for each of these candidate change points while accounting for the multiple comparison. The test statistic is the maximum of 2-scale wavelet products across different scales which is a powerful way of combining information across scales. The justification of using 2-scale product was made by comparing 2-scale product with 2-scale sum and single-level wavelet coefficients. The estimated change points were shown to be consistent. We developed non-parametric approaches for estimating the null distribution, including permuting wavelet coefficients at finest scale W_1 and permuting $\hat{\epsilon}$ through lowess smoothing. Fi-

nally, we illustrated the idea of testing at local maxima to avoid false positives around the change points due to autocorrelations among adjacent wavelet coefficients.

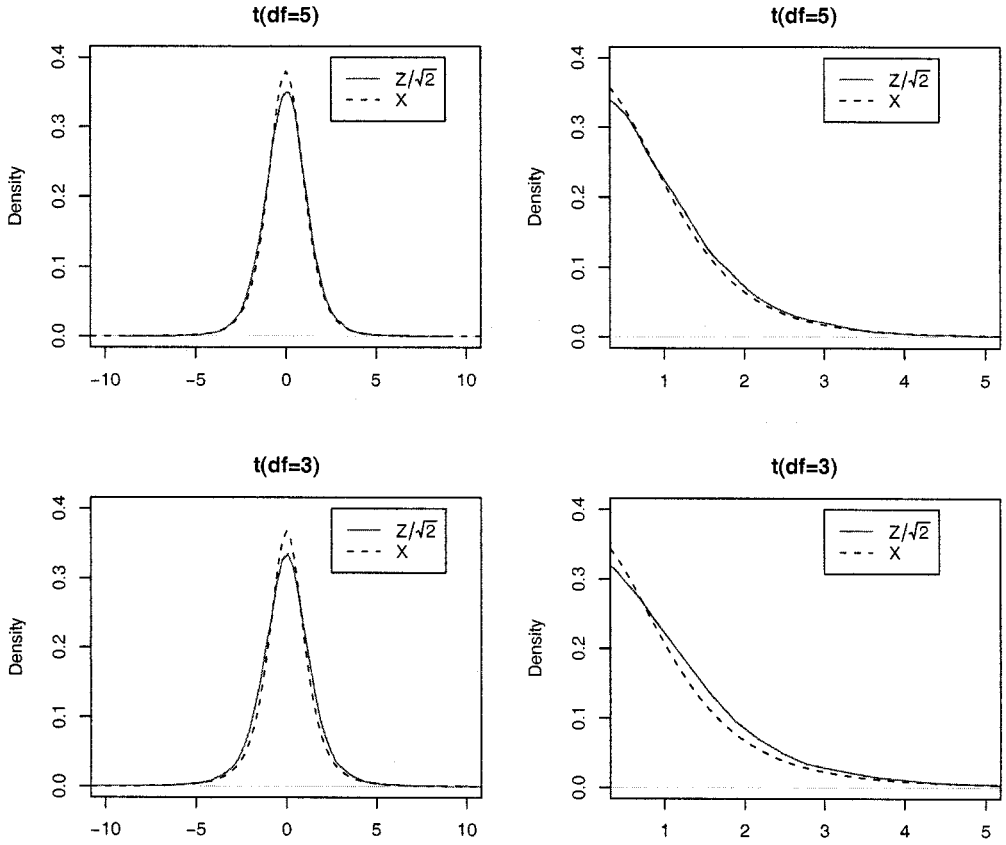


Figure 3.4: Density functions for $X \sim$ t(df=5) and t(df=3) in the left panel, the solid lines indicate the $Z/\sqrt{2}$ and dotted lines are for X. The right panel show the zoomed right tails.

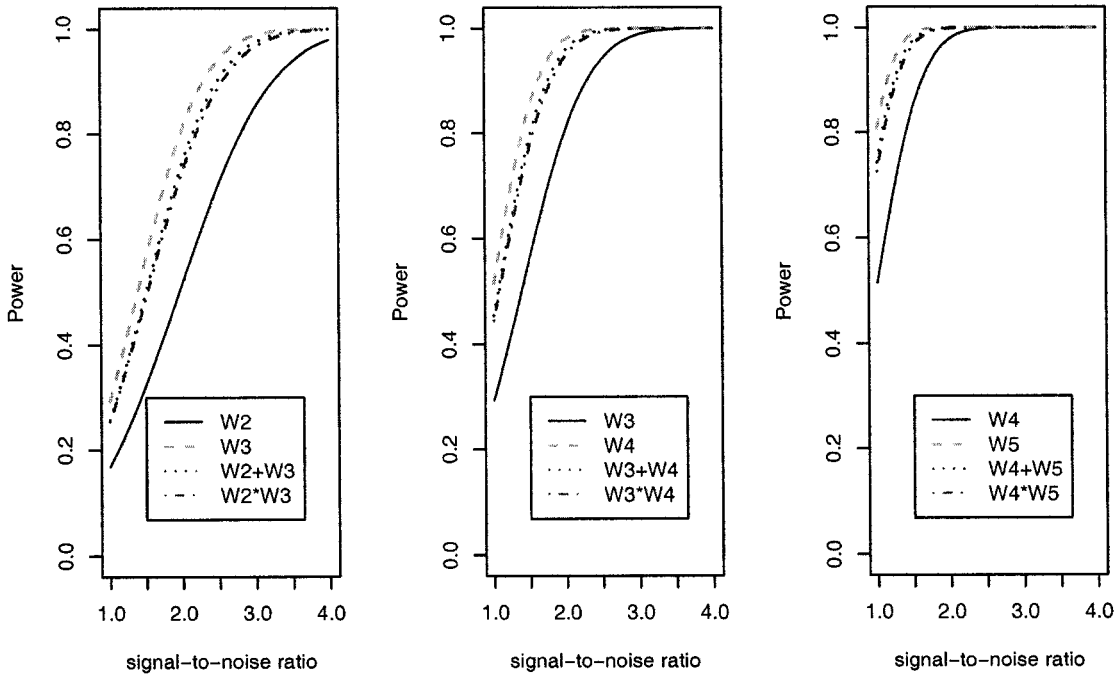


Figure 3.5: Power functions for wavelet coefficients, 2-scale sum and 2-scale product at scale 2(left panel), 3 (central panel) and 4 (right panel) using single test.

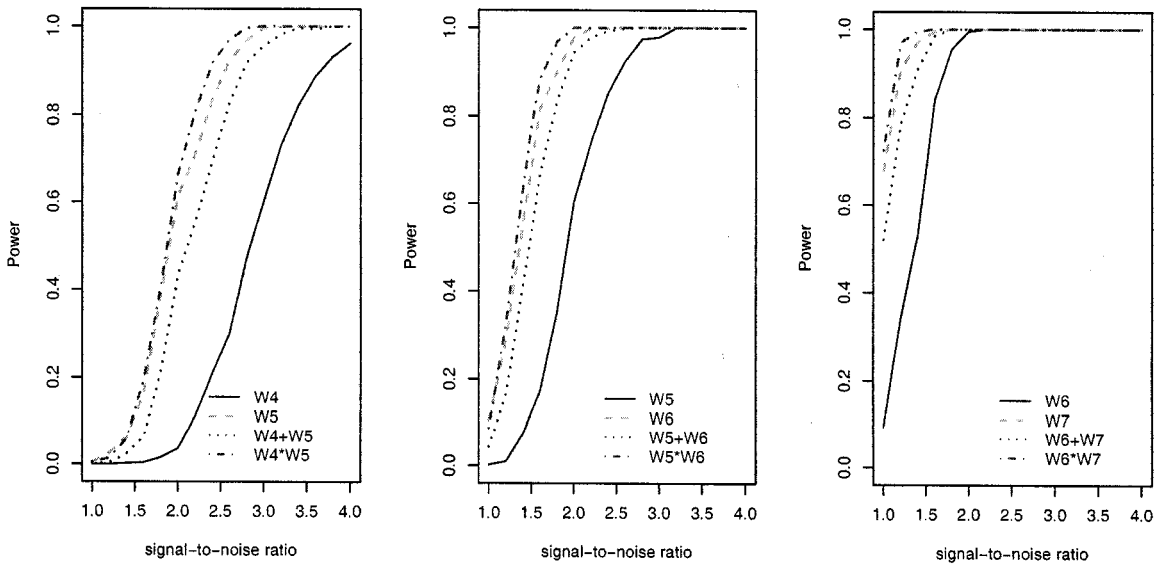


Figure 3.6: Power of detecting a square wave signal using wavelet coefficients, 2-scale sum and 2-scale product. The powers are calculated based on 500 simulated data sets.

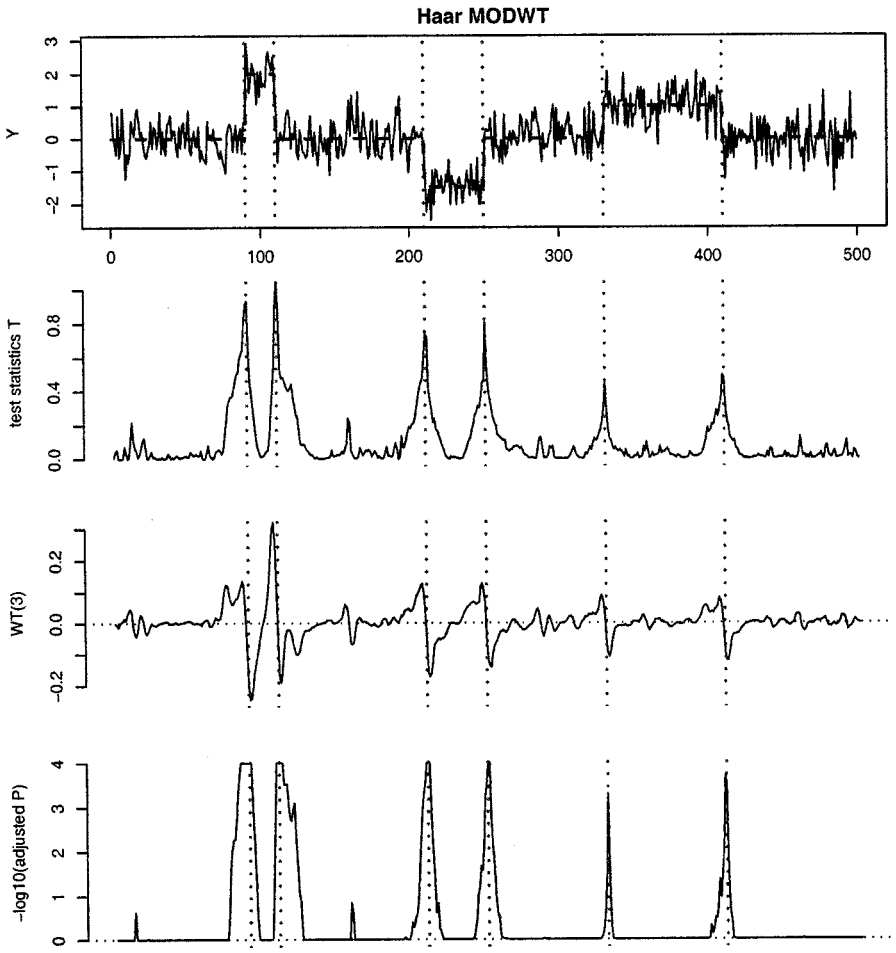


Figure 3.7: The adjusted p- values for test statistics. The top plot is the simulated signal Y with the dotted blue lines indicating the true step function f. The vertical red lines indicate the change points. The second plot represents the test statistics for each marker. The third plot is the wavelet transform of T at scale 3. The bottom plot is the $-\log_1 0$ of adjusted p-values truncated at 4.

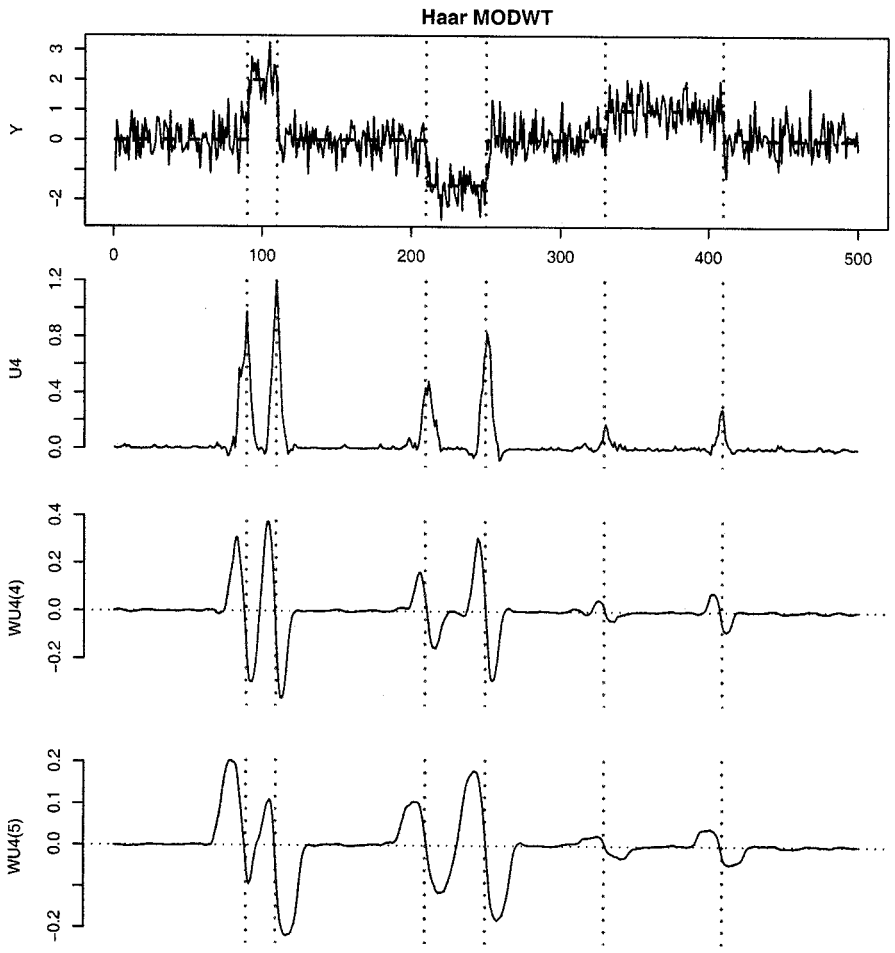


Figure 3.8: Wavelet transform of 2-scale product U_4. The top plot is the simulated signal Y with the dotted blue lines indicating the true step function f. The vertical red lines indicate the change points. The second plot represents $U_4 = W_4 W_5$ for each marker. The third plot is the wavelet transform of U_4 at scale 4. The bottom plot is the wavelet transform of U_4 at scale 5.

Chapter 4

ANALYSIS RESULTS FROM REAL DATA

In this chapter we illustrate our proposed methods using two real datasets from array CGH experiments, Coriel cell lines data by Snijders et al. (2001) and Map 10K breast cancer data by Loo et al. (in preparation). We will briefly describe the two studies in Section 4.1 and Section 4.2, respectively. The detailed description of the data can be found in the original papers. In both real data analysis and simulations in Chapter 5, we compared the proposed method with the CBS method (Olshen et al. 2004) and the penalized likelihood method implemented in CGHseg (Picard et al 2005). We chose these two methods for comparison for two reasons. First, the methods by Olshen et al. (2004) and Picard et al. (2005) performed consistently well based on a comprehensive comparison study performed by Lai et al. (2005). Second, both the CBS and the CGHseg methods report change points along with segmented regions, unlike some of other function estimation methods which only give smoothed or denoised profiles, making it difficult to compare with our method. The CBS method detects break points by controlling the overall type I error rate under the complete null hypothesis which makes it more comparable with our procedure.

When not specified, the default parameters are used in both the CBS and the CGHseg methods, even though some parameters could be tuned to fit a specific data set better. Particularly, we will use the CBS method without the extra pruning step as the number of estimated change points is very sensitive to the pruning parameters and choosing such pruning parameters is rather subjective.

4.1 Coriel Cell Lines Data

In 2001, Snijders et al. studied the DNA copy number changes for 15 Coriel cell lines using array CGH technology. Each array contained 2276 mapped BAC clones spotted in triplicate. The Coriel cell line data have been analyzed by many methods (Olshen et al. 2004; Fridlyand et al. 2004; Hsu et al. 2005; Zhang et al. 2007) and can be freely downloaded at http://www.nature.com/ng/journal/v29/n3/suppinfo/ng754_S1.html. This data set is mainly used as a proof of principles because the copy number alterations have been identified by an alternative technology, spectral karyotyping. Of 15 cell lines, 6 cell lines have whole chromosome amplifications only and 9 cell lines have partial chromosome amplifications or deletions. We applied the 2-scale product wavelet method to all 15 Coriel cell lines with known change points and compared the detected change points with the truth.

We analyzed the whole genome, i.e. all 23 chromosomes, simultaneously. There are several considerations for this. First, the density of clones varies greatly from chromosome to chromosome. For example, chromosomes 19, 21 and 22 have 35, 30 and 15 clones, respectively, whereas only 7 chromosomes have more than 100 clones. Second, it allows us to detect whole chromosome aberrations as the rest of genome provides a good baseline reference. Third, the estimation of variance will be more accurate since the estimation is based on the whole genome with many more data points. Finally, genome-wide analysis allow us to control type I error rate for the whole genome, instead of at the chromosome level.

We calculated kurtosis for each cell line to decide which multiple comparison approach to use for the wavelet method. Under the assumption that the number of change points is much smaller relative to the total number of markers, we could approximately estimate the skewness and kurtosis of the error distribution using W_1. The median skewness of W_1 was -0.02 with 1st quantile -0.11 and 3rd quantile -0.004. The median kurtosis of W_1 was 6.9 with 1st quantile 3.7 and 3rd quantile 8.6. Recall

that the kurtosis of W_1 is only half of the kurtosis of the error distribution, therefore, assuming normality might not be appropriate for this data set. Two multiple comparison approaches, permuting W_1 and $\hat{\epsilon}$, were used. Interestingly we found that both approaches gave the exact same results for all 15 cell lines except for GM02948. Permuting W_1 had an additional false positive on Chromosome 20 at 65.3Mb (Clone ID RMC20P071). Though some of the cell lines have very large kurtosis, permuting W_1 still performed reasonably well mainly because this data set has very low noise level. It is also because that W_1 is robust to distribution assumption. Chapter 5 will further investigate the validity of this approach under various non-normal error distributions. Since the Coriel data are generated from cell lines, the noise level is considerably low with mean of $\hat{\sigma}$ 0.067 and $\hat{\sigma}$ range from 0.052 to 0.079. Here only the results from permuting $\hat{\epsilon}$ will be presented.

Figure 4.1 shows the segmentated profiles of all 15 Coriel cell lines using the wavelet method, the CBS method and the CGHseg method. We observed that both the wavelet and CGHseg methods gave almost identical segmentations while the CBS method detected many small segments which are probably false positives. The SNRs were calculated for each aberration region based on segmentation results using the wavelet method. The mean SNR was 9.0 with range from 5.5 to 12.7. In Table 4.1, we list the number of false positives and false negatives for 15 cell lines at significant level 0.01. Note that the CBS method had many false positives, total 99, while the wavelet method and the CGHseg method had only 4 and 9 false positives, respectively. Of these false positives, all three methods detected the whole chromosome aberrations on sex chromosome 23 for GM00143, GM01535, GM05296 and GM07408. This suggested some of the false positives, such as the aberrations on chromosome 23, might be real but just not confirmed by cytogenetic experiment.

All three methods had two false negatives on cell lines GM01535 in which only one altered clone exists on chromosome 12qtel. The CBS method is not designed to detect singletons due to lack of variation. Though both the wavelet and CGHseg

method are capable of detecting singletons, it is difficult to detect singletons when SNR is not large enough. Moreover without external information, it is impossible to distinguish singletons from outliers. One approach may be to examine whether the singletons are recurrent over many arrays. However, technical errors , such as failure to hybridize, could also contribute to the recurrence. In this case, normal samples will be needed to help distinguish real alterations at single clones from outliers. Table 4.2 lists the change points along with their adjusted p-values for 9 Coriel cell lines as there are no change points (only whole chromosome aberrations) for other 6 cell lines. The adjusted p-values for the two false negatives on GM01535 are 0.231 and 1.000. Figure 4.2 shows the adjusted p-values for 4 cell lines.

Many existing segmentation methods, such as the CGHseg, Eilers et al. (2005) and modified BIC by Zhang et al. (2007), require the error being normally distributed. However, it is not known whether this normality assumption holds in general. We will show later in Chapter 5 by simulations that the parametric methods, such as the CGHseg method, can perform very poorly when the normality assumption is not met, therefore it is important to test the validity of this key assumption. In Figure 4.3 and 4.4, we plotted the QQplots for the fitted residuals using the wavelet method along with the sample kurtosis for the fitted residuals. We can see that some cell lines had heavy tails with very large kurtosis, for example GM03134 and GM05296. Therefore, a distribution-free method might be needed for analyzing array CGH data, such as the CBS method and the proposed wavelet method.

Since this data set has also been analyzed by Fridlyand et al. (2004) and Zhang et al. (2007), we could compare the proposed wavelet method with Fridlyand et al. (2004) and Zhang et al. (2007) methods indirectly by quoting their analysis results. It is interesting to note that methods by Fridlyand et al. (2004) and Zhang et al. (2007) detect all the change points even the singleton on GM01535 while the wavelet, the CBS and CGHseg failed to detect the singleton. The methods by Fridlyand et al. (2004) and Zhang et al. (2007) have few number of false positives. As stated by

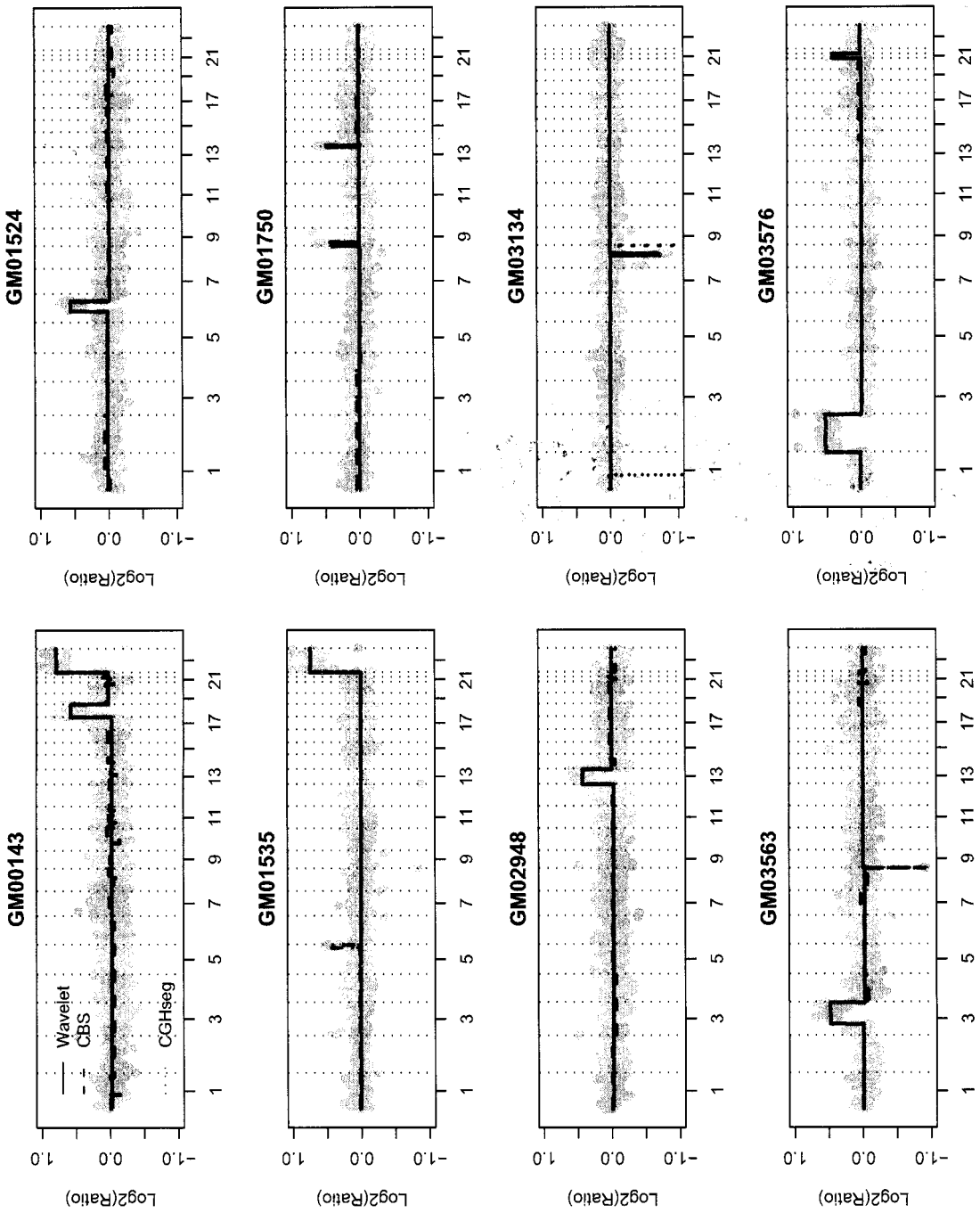
Zhang et al. (2007), the hidden Markov model of Fridlyand et al. (2004) involves user selection of training parameters and thresholds, so it might not be a fair comparison.

4.2 Breast Cancer Data

To further illustrate the application of the wavelet method, we use another more challenging data on breast cancer reported by Loo et al. (in preparation), because the noise level is high with the mean of $\hat{\sigma}$ 0.90 and 1st quantile 0.71 and 3rd quantile 1.07. The objective of this study is to examine the commonalities and differences of the genomic copy number profiles of two major histological types of breast cancer, infiltrating ductal carcinoma (IDC) and infiltrating lobular carcinoma (ILC). IDC and ILC differ both clinically and morphologically. IDC is the most common type of breast cancer and accounts for 86-95% of breast cancers but ILC incidence rate has increased rapidly recently. ILC is more likely to be estrogen receptor and progesterone receptor positive and is slightly larger than IDC on average. This report is the first large scale genome-wide study to identify genomic alterations in lobular tumors and contrast them with the more common ductal tumors. The genomic DNA of 89 IDC and 78 ILC formalin-fixed tumor samples were extracted and the copy numbers were measured for 9670 clones using the GeneChip Mapping 10K Assay (Affymetrix, Santa Clara CA). We analyzed one chromosome at a time due to a large number of clones in the genome.

To decide which multiple comparison approach to use for the wavelet method, we again calculated the skewness and kurtosis for each array using W_1. The median skewness was -0.006 with 1st quantile -0.021 and 3rd quantile 0.004. The median kurtosis was 0.546 with 1st quantile 0.419 and 3rd quantile 0.749. We analyzed the data using permuting W_1 and $\hat{\epsilon}$. Both approaches gave similar results except permuting W_1 detected slightly more change points for some arrays which might very likely be false positives. Hence, we will only present results using permuting $\hat{\epsilon}$ here.

For illustration, Figure 4.5 shows segmented profiles using the wavelet, the CBS


GM00143
GM01524
GM01535
GM01750
GM02948
GM03134
GM03563
GM03576
Wavelet
CBS
CGHseg
Log2(Ratio)
1.0
0.0
−1.0
1
3
5
7
9
11
13
17
21

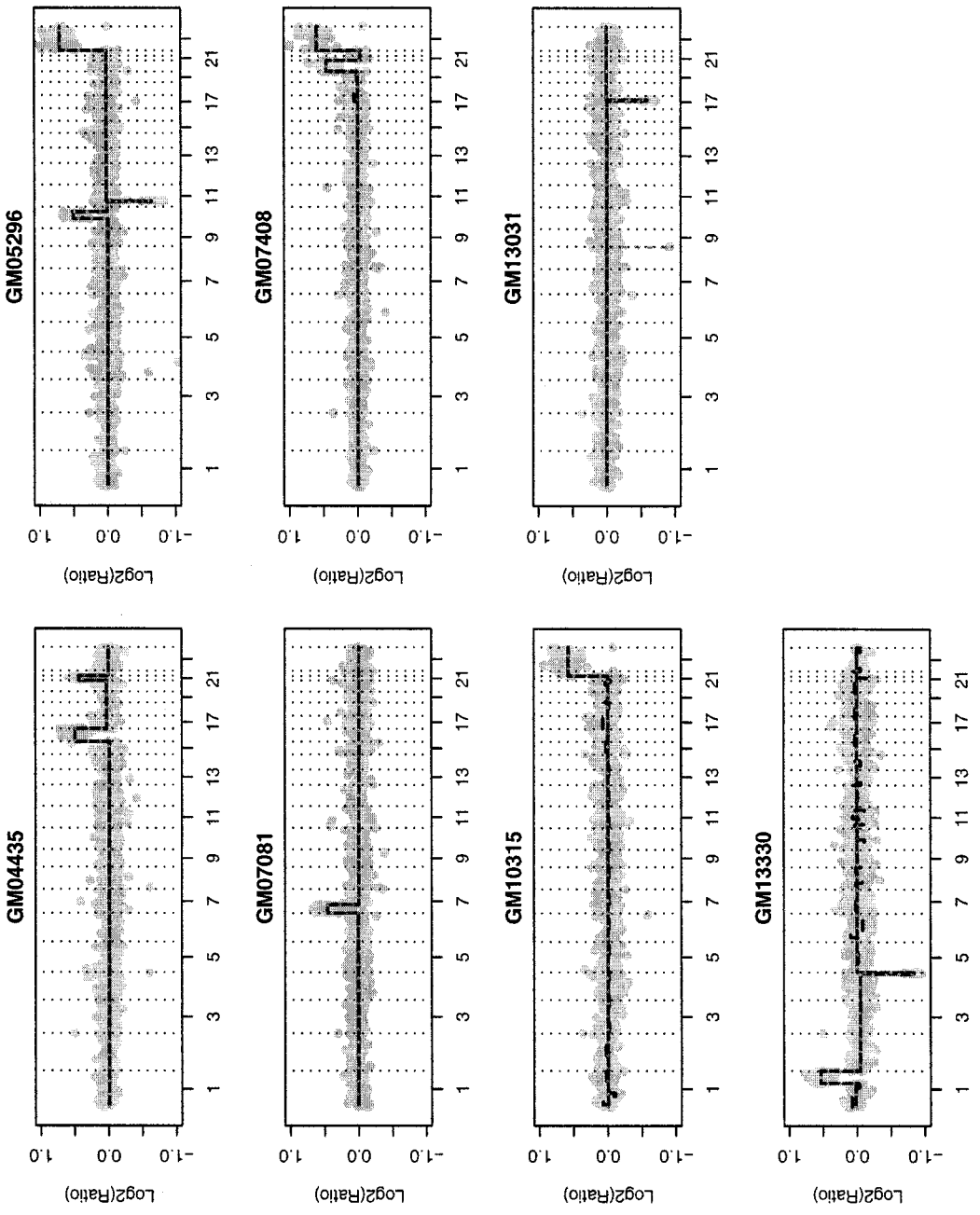


Figure 4.1: Segmentation of 15 cell lines using the wavelet, CBS and CGHseg methods. Solid lines indicate the wavelet method, the dash lines are for the CBS method and the dotted lines are for CGHseg.

Table 4.1: The number of false positives (FP) and false negatives (FN) on 15 Coriel cell lines using the wavelet, CBS and CGHseg methods. significant level 0.01 was used for both the wavelet and CBS methods. No pruning was performed for CBS.

	Wavelet		CBS		CGHseg	
Cell line	FP	FN	FP	FN	FP	FN
GM00143	1	0	27	0	1	0
GM01524	0	0	11	0	0	0
GM01535	1	2	1	2	1	2
GM01750	0	0	2	0	1	0
GM02948	0	0	7	0	0	0
GM03134	0	0	0	0	2	0
GM03563	0	0	10	0	0	0
GM03576	0	0	1	0	0	0
GM04435	0	0	2	0	0	0
GM05296	1	0	1	0	1	0
GM07081	0	0	0	0	0	0
GM07408	1	0	3	0	1	0
GM10315	0	0	10	0	0	0
GM13031	0	0	0	0	2	0
GM13330	0	0	23	0	0	0
Total	4	2	99	2	9	2

Table 4.2: change points and their adjusted p-values on 9 Coriel cell lines using the wavelet method.

Tumor	BAC clone	Chromosome location	kb	Adjusted p values
GM01524				
	CTD-2009c06	6q12	74205	< 0.001
	RP11-107m03	6q22.3	143303	< 0.001
GM01535				
	RP11-210k16	5q34	176824	< 0.001
	Vysis-5qtel_227A	5qtel	198500	< 0.001
	RP11-81g12	12q24.3	141551	0.231
	Vysis-12q_tel_241A	12qtel	142000	1.000
	Vysis-X/Y_p_tel_257A	23ptel	0	< 0.001
GM01750				
	RP11-33o15	9p21	24325	< 0.001
	RP11-125A05	14q12	9655	< 0.001
GM03134				
	RP11-117N14	8q21.1	84403	0.003
	RP11-27I15	8q21.3	95100	0.003
GM03563				
	Vysis-FHIT/HRAC1_318B	3p14.2	76349	< 0.001
	RP11-233n20	3q29	217902	< 0.001
	RP11-28n06	9p24	1367	< 0.001
GM05296				
	RP11-14i14	10q21.3	64187	< 0.001
	RP11-128h11	10q24-10q25	110000	< 0.001
	Vysis-WT1_371C	11p12	34420	< 0.001
	RP11-72a10	11p12	39623	< 0.001
	Vysis-X/Y_p_tel_257A	23ptel	0	< 0.001
GM07081				
	RP11-251I15	7ptel	57971	< 0.001
GM13031				
	RP5-1071i14	17q	50231	< 0.001
	RP11-670E13	17q22-q24	58122	< 0.001
GM13330				
	RP11-234m03	1q23	156276	< 0.001
	RP11-188a04	1q43-44	237341	< 0.001
	RP11-272O03	4q34.3	173943	< 0.001
	Vysis-4qtel_225A	4qtel	184000	< 0.001

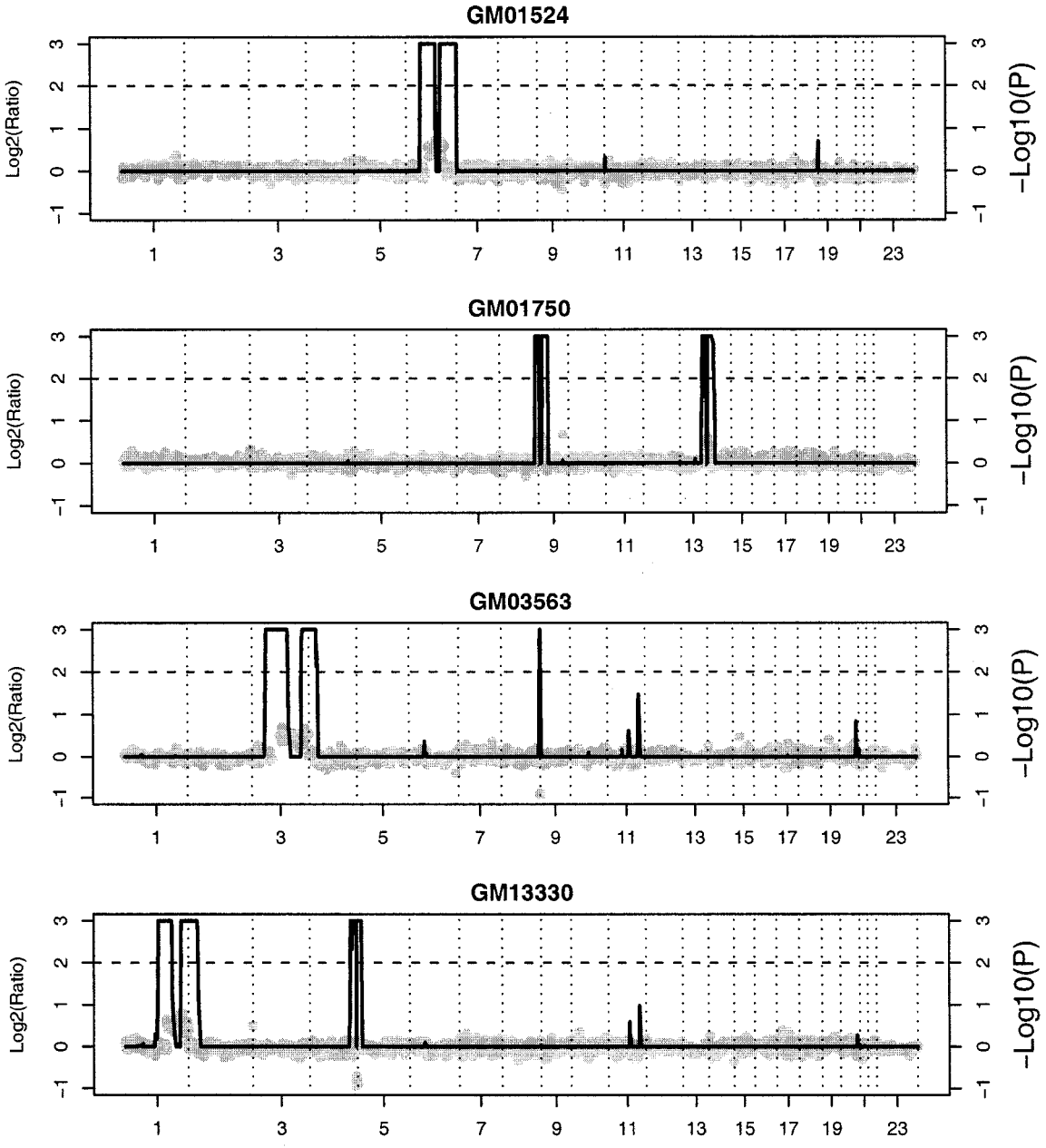


Figure 4.2: Plot of adjusted p-values vs marker loci. The y-axis (right side) is the $-\log_{10}(P)$. p-values are truncated at 0.001. The horizontal dash line is at the significant level $p = 0.01$.

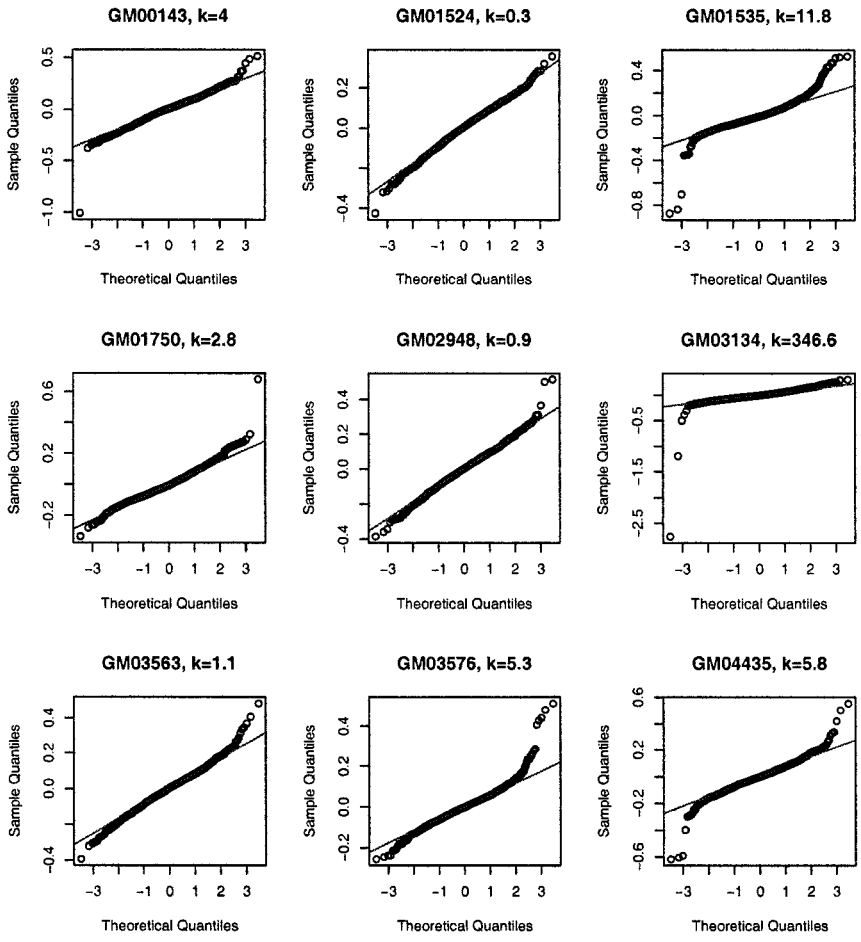


Figure 4.3: QQ-plots of $\hat{\epsilon}$ after segmented by the wavelet method. k denotes kurtosis.

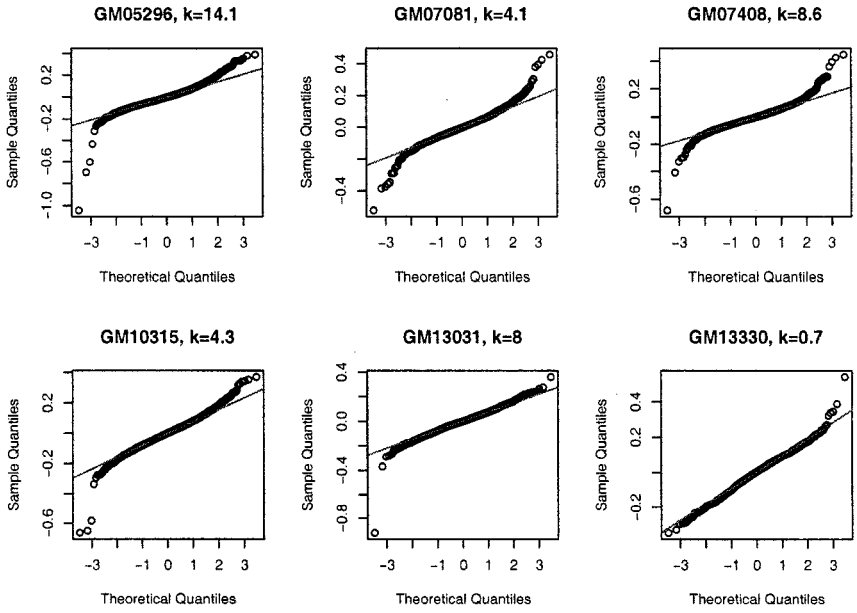


Figure 4.4: QQ-plots of $\hat{\epsilon}$ after segmented by the wavelet method. k denotes kurtosis.

and the CGHseg methods for tumors L141T, L165T, D170T and D098T. We used the approach of permuting residuals by fitting a lowess smoothing with window size 0.1. All three methods gave similar segmented profiles except that the CGHseg method tends to detect more very narrow aberration regions on chromosomes with no obvious abrupt edge, such as chromosome 14, 19 and 20. This is actually consistent with our simulation results that will be shown in Chapter 5 that the CGHseg method tends to yield more false positives when there is no change point. The chromosomes 8 and 17 were plotted individually in Figure 4.6 for these four tumors to have a better view of the segmentations. Since the true change points for these breast cancer tumors are not known, we can not report the overall number of false positives and false negatives. Instead we will present the change points hot spots found on chromosomes 8 and 17 stratified by breast cancer subtypes, IDC and ILC.

The change points hot spots are regions where change points occur in multiple tumors. Table 4.3 lists the change points hot spots found on chromosomes 8 and 17 by the wavelet, the CBS and the CGHseg methods. A total of six hot spots were found and three on each chromosome. The three hot spots on chromosome 8 were around 35 and 37 Mb. The three hot spots on chromosome 17 were found around 14 Mb and 23Mb. We then searched these regions on National Center for Biotechnology Information (NCBI) web site (http://www.ncbi.nlm.nih.gov) to see if any interesting genes are located on the regions. The results are listed in Table 4.4. The gene located on chromosome 17 around 23 Mb is known coded the galectin which is a family of beta-galactoside-binding proteins. This galectin is involved in modulating cell-cell interactions and is strongly overexpressed in Hodgkin's disease tissue.

We found that all three methods gave comparable results with the CGHseg method detecting slightly more change points at these regions. The number of tumors having copy number changes at particular marker was counted for subtypes IDC and ILC, then for each marker frequencies of copy number changes for IDC and ILC were tested using Fisher's exact test. The change points hot spots on chromosome 17 were

common on both IDC and ILC with p-values greater than 0.3 for all three methods. The copy number change on chromosome 8 at 35.5 Mb might be more likely for IDC than ILC. Based on the wavelet method, 18.0% of IDC had copy number changes at 35.5 Mb while only 7.7% of ILC had copy number changes (p=0.066).

Some of breast cancer CGH arrays have very high variance. In this case, the wavelet method may fail to detect change points. However, with the improvement of the "technology" which includes array-technology and better DNA extraction methods from archived samples and single cell, we expect that the noise level of the data continues to reduce. Moreover, as the marker density increases on the array, the resolution of the locations of change points becomes finer which also enhances the power for detecting the change points.

4.3 Summary

We applied the 2-scale wavelet product method to two real data sets. For comparison we also applied both the CBS and CGHseg method. All three methods gave comparable results for both the Coriel data and the breast cancer data. Only the wavelet method provides the estimation of adjusted p-values for each marker. The adjusted p-values provide some flexibility for scientists to call change points at their chosen significant level. The Coriel data suggested that the wavelet method has the least false positives while the CBS method has the most. All three methods detected all the change points but one singleton. The Coriel data also indicated that normality assumption might not be valid for the array CGH data.

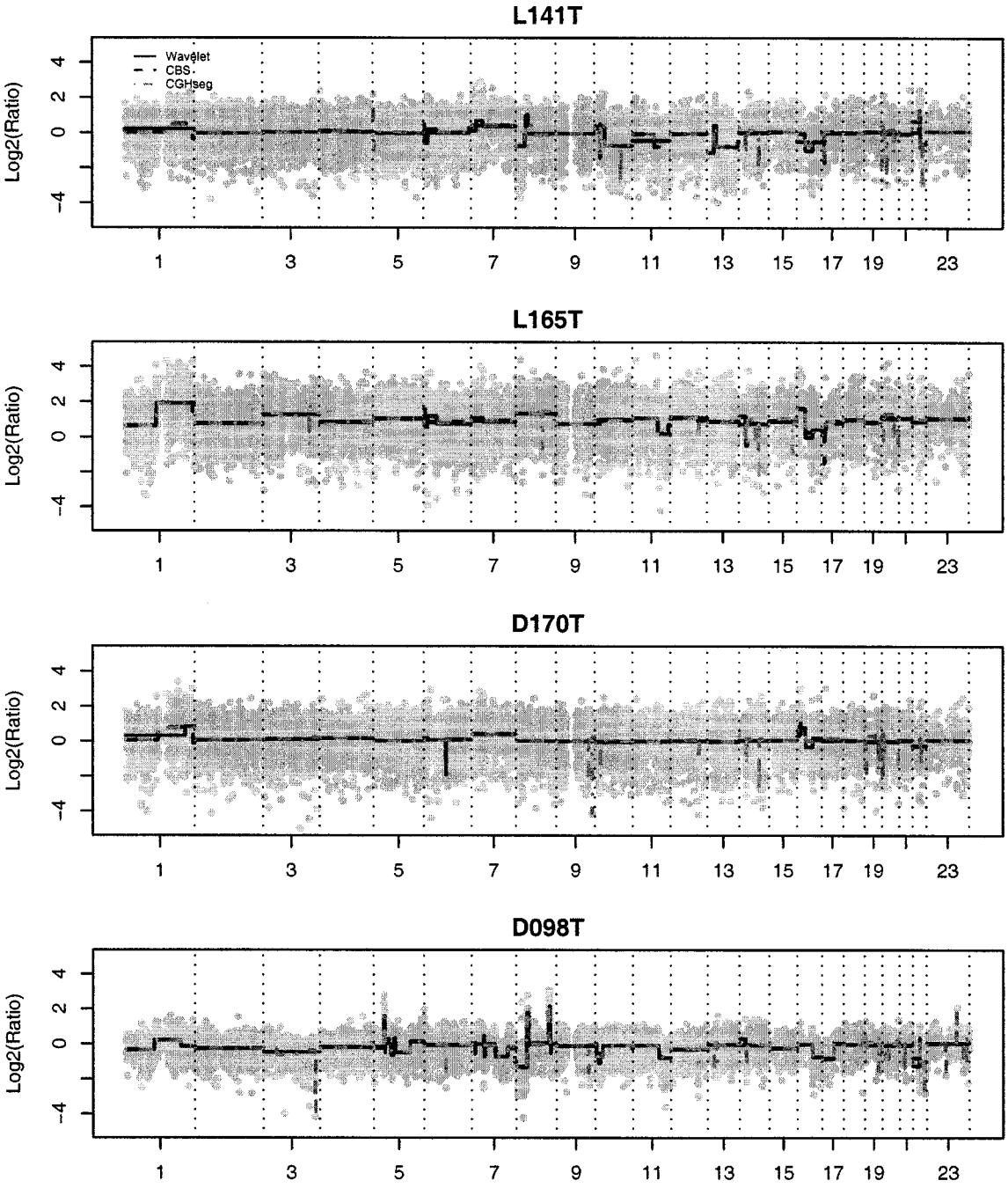


Figure 4.5: Segmented profiles using three methods for lobular tumors L141T and L165T, ductal tumors D170T and D098T. Red lines indicate wavelet method, blue lines indicate CBS method and green lines Picard method

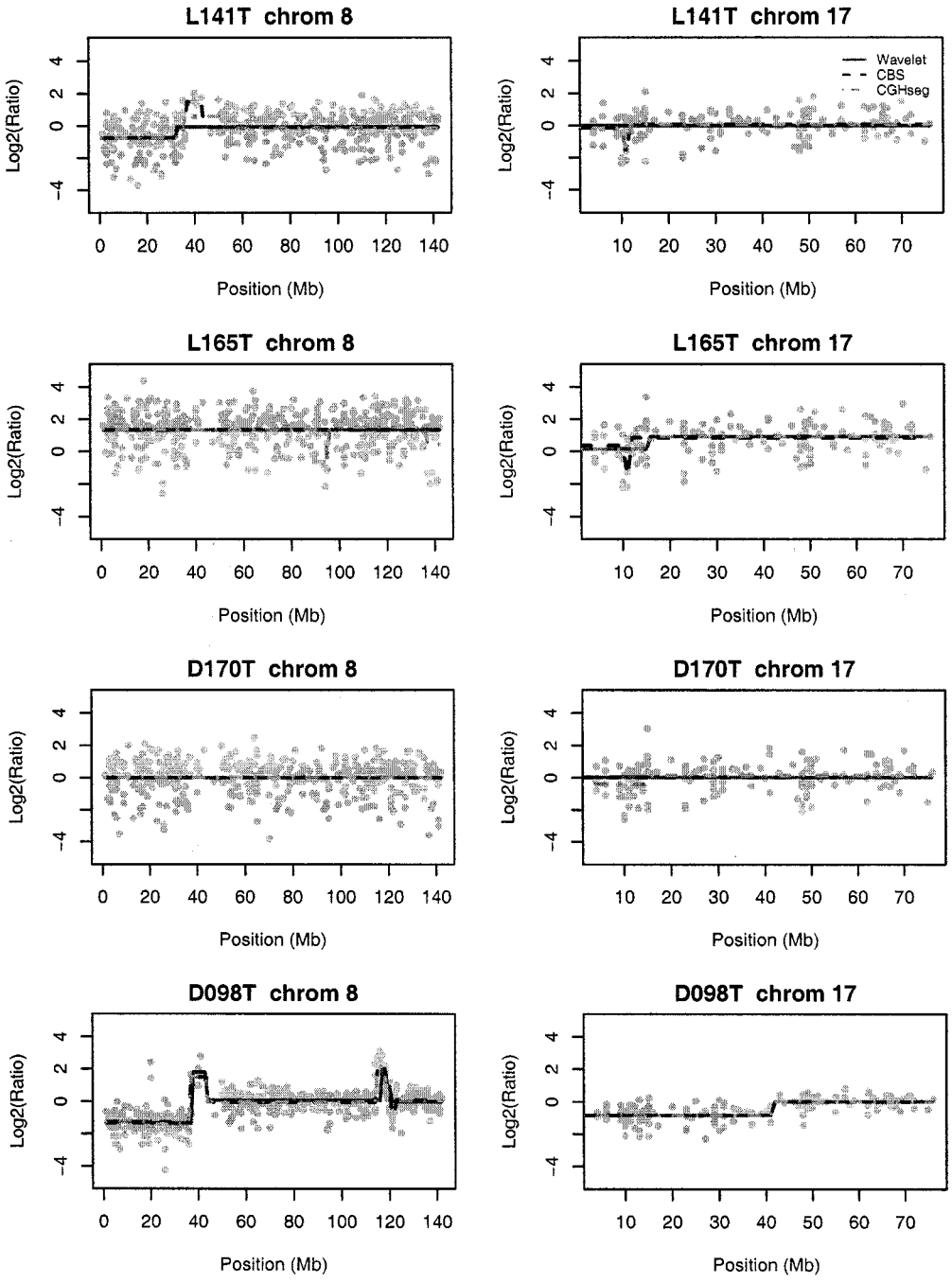


Figure 4.6: Segmented profiles using three methods for 8 tumors at chromosome 8. Red lines indicate wavelet method, blue lines indicate CBS method and green lines Picard method

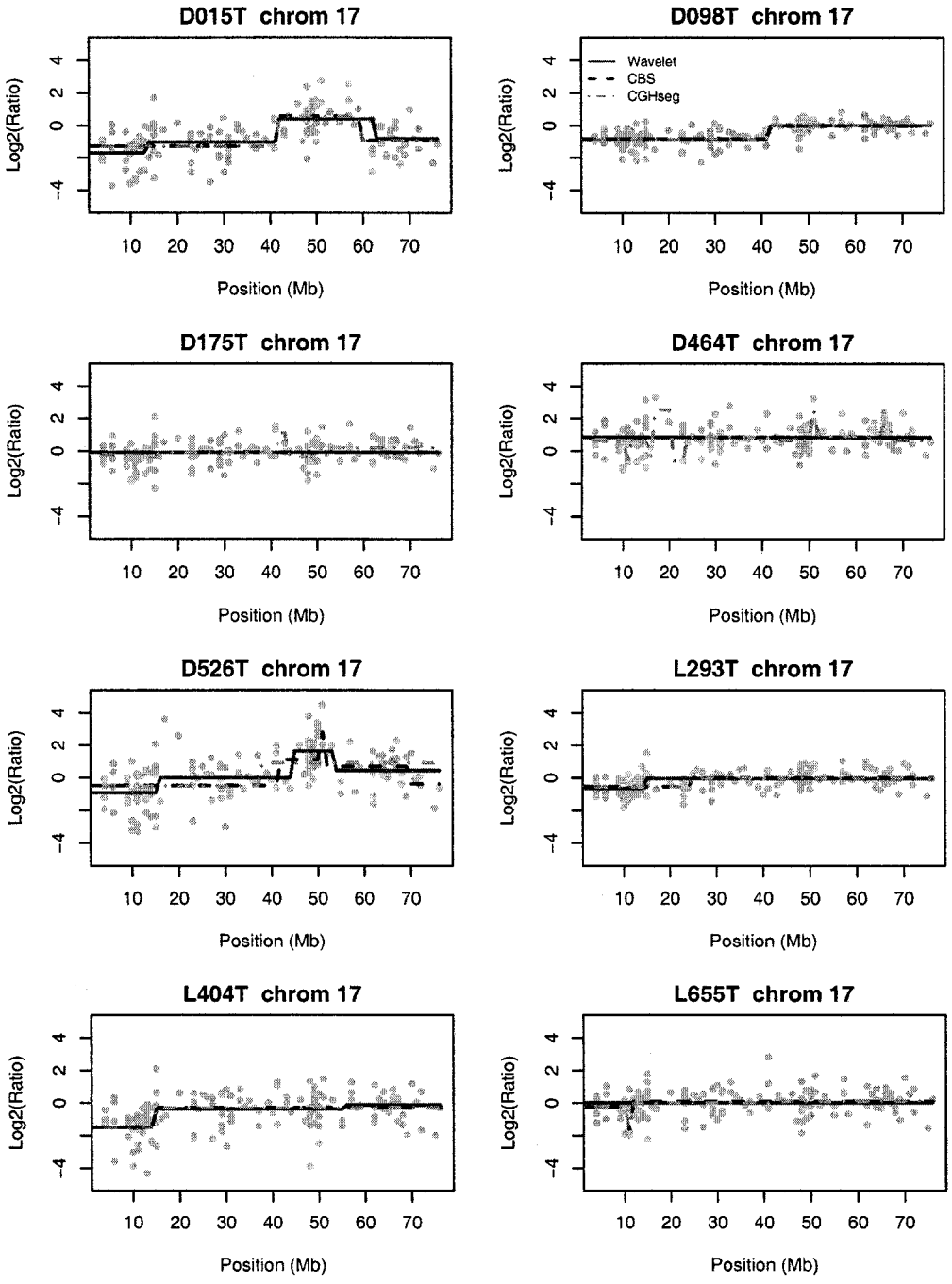


Figure 4.7: Segmented profiles using three methods for 6 tumors at chromosome 17. Red lines indicate wavelet method, blue lines indicate CBS method and green lines Picard method

Table 4.3: change points hot spots on chrome 8 and 17 detected by the wavelet, CBS and CGHseg methods. Significant level 0.01 for the wavelet and CBS methods. down=the number of tumors with mean decreased at change points, up=the number of tumors with mean increased at change points. p were calculated based on Fisher's exact test to compare two tumor subtypes.

				Wavelet			CBS			CGHseg		
SNP ID	Chrom.	Mb	Subtype	down	up	p	down	up	p	down	up	p
SNP_A-1518044	8p12	35.3	IDC	7	0	0.015	1	0	1.00	4	0	0.686
			ILC	0	0		0	0		2	0	
SNP_A-1517267	8p12	35.5	IDC	1	16	0.066	0	17	0.078	0	21	0.046
			ILC	0	6		0	7		0	9	
SNP_A-1507557	8p12	36.8	IDC	0	6	0.030	0	5	0.061	0	5	0.216
			ILC	0	0		0	0		0	1	
SNP_A-1514461	17P12	14.3	IDC	0	9	1.00	0	8	0.383	0	9	1.00
			ILC	0	8		0	4		0	8	
SNP_A-1508161	17P12	14.5	IDC	0	13	1.00	0	7	1.00	0	19	1.00
			ILC	0	11		0	7		0	17	
SNP_A-1514843	17q11.1	22.8	IDC	0	7	0.340	0	7	0.545	0	11	0.820
			ILC	0	3		0	4		0	11	

Table 4.4: Genes located around the change point hot spots on chromosomes 8 and 17

Gene	Chrom.	Mb	Full name
UNC5D	8p12	34.3-36.87	function unknown
-	17P12	14.08-14.79	function unknown, similar to ribosomal protein S18
LGALS9	17q11.1	22.98-23.00	lectin, galactoside-binding, soluble, galectin 9

Chapter 5

PERFORMANCE OF THE 2-SCALE PRODUCT WAVELET METHOD

In this chapter, we will evaluate finite sample performance of our proposed wavelet method for detecting change points using simulations. Specifically, the performance of the 2-scale product wavelet method was examined for the cases of ϵ being both normally and non-normally distributed. We also explored the performance of the wavelet method under a trend model. For all the simulations, the wavelet method was compared with the CBS method by Olshen et al. (2004) and the CGHseg method by Picard et al. (2005).

5.1 Measures of Performance

In order to assess the performance of our proposed wavelet method, we conducted a series of simulation studies. Again we compared the wavelet method with the CBS method and the penalized likelihood method implemented in CGHseg. The performance of each method was measured by the mean squared error (MSE), true positive rate (TPR), false discovery rate (FDR), number of estimated change points and number of exact detections. The mean squared error is defined as

$$MSE = n^{-1} \sum_{i=1}^{n} (\hat{f(i)} - f(i))^2.$$

The true positive rate is the proportion of true rejections among true change points. The FDR is the proportion of false rejections among total number of rejections. The number of exact detections is the number of cases such that the locations and number of all estimated change points are correct.

5.2 Simulation Under Normal Distribution

First we considered ϵ being normally distributed. The data were simulated according to the model in Equation (3.1) in Chapter 2. The performance of the wavelet method was evaluated under different underlying mean functions f and noise levels σ. For each scenario, a total of 500 datasets were simulated with each dataset having 500 markers. Since the ϵ's are normally distributed we could estimate the null distribution, the distribution of ϵ, by permuting scaled W_1 coefficients.

5.2.1 Simulation 1: Global Null

We first investigated the performance of the wavelet method under a global null situation where $f = 0$, i.e. no change point and Y are i.i.d normal with mean zero and $\sigma = 0.2$, or 0.4. For each method we estimated the number and locations of estimated change points at significant level $\alpha = 0.01$ for the wavelet and CBS methods. Table 5.1 shows the simulation results based on 500 simulated data sets. The type I error rate is defined as the probability of detecting one or more change points in the data set. Note that both wavelet and CBS methods had their type I error rate under control and their family wise error rates were 0.008 and 0.002 respectively, with the CBS method somewhat conservative. Two multiple comparison approaches, permuting W_1 and $\hat{\epsilon}$, gave very comparable results.

We observed that CGHseg tended to detect more false positives under the null. For example, the probability of having 2 false positives was 26% and the type I error rate, the probability of detecting at least one false positive, was 47.8%. Recall that the CGHseg method selects the most parsimonious model by balancing between maximizing a likelihood function and the number of change points R in the model. It selects R such that the second derivative of the likelihood, $\hat{\mathcal{L}}_R$, is smaller than a given threshold $S \times n$. We used the default $S = 0.75$. It is possible that by choosing a more stringent (larger) threshold value, we could reduce type I error rate

dramatically. For example for $S = 1$ and $S = 2$, the type I error rates are reduced to 0.275 and 0.012, respectively. Now the question is which S should be used. In the following simulations, we used $S = 2$ in addition to the default $S = 0.75$ to have a fair comparison among the three methods. It is worth noting that the performance of all three methods are not sensitive to the noise level σ under the null. The results were the same at $\sigma = 0.4$ and 0.2 for all three methods.

Table 5.1: Simulation results under the null hypothesis of no change points, $\alpha = 0.01$ and $\epsilon \sim N(0, \sigma^2)$.

	# of change points									
	$\sigma = 0.4$					$\sigma = 0.2$				
Method	0	1	2	3+	Type I error	0	1	2	3+	Type I error
Permute W_1	496	4	0	0	0.008	496	4	0	0	0.008
Permute $\hat{\epsilon}$	497	3	0	0	0.006	497	3	0	0	0.006
CBS	499	0	1	0	0.002	499	0	1	0	0.002
CGHseg($S = 0.75$)	261	16	130	93	0.478	261	16	130	93	0.478
CGHseg($S = 2$)	494	0	5	1	0.012	494	0	5	1	0.012

5.2.2 Simulation 2: Evenly-spaced change points

In the second simulation experiment we generated chromosome profiles with different noise levels, number of change points and aberration width to investigate how three methods performed in the presence of change points. The aberration regions were evenly spaced along the chromosome. Function $f(i)$ was set to be either 0 or 1 corresponding to no change or copy number gain. The number of change points (R) varied among 2, 4, and 6. The width of each aberration region (c) were set to be 20,

40 or 80 markers. To be specific, functions f for $R = 2, 4, 6$ can be written as

$$f_1(i) = 1_{\frac{n-c}{2} < i \leq \frac{n+c}{2}}(i),$$

$$f_2(i) = 1_{\frac{n-2c}{3} < i \leq \frac{n-2c}{3}+c}(i) + 1_{\frac{2(n-2c)}{3}+c < i \leq \frac{2(n-2c)}{3}+2c}(i),$$

$$f_3(i) = 1_{\frac{n-3c}{4} < i \leq \frac{n-3c}{4}+c}(i) + 1_{\frac{2(n-3c)}{4}+c < i \leq \frac{2(n-3c)}{4}+2c} + 1_{\frac{3(n-3c)}{4}+2c < i \leq \frac{3(n-3c)}{4}+3c}(i),$$

respectively, with $1_A(i)$ be an indicator function such that $1_A(i) = 1$ if $i \in A$, otherwise 0. For each scenario, 500 simulated datasets with each consisting of $n = 500$ markers were generated such that Y_is were independently and identically normally distributed with mean $f(i)$ and variance σ^2. The standard deviation σ was set to be either 0.2, 0.4 or 2/3 yielding SNR=5, 2.5 and 1.5, respectively; $\sigma = 0.4$ was chosen to mimic the SNR in the breast cancer data.

Table 5.2 and 5.3 display the simulation results at $\sigma = 0.2$ for three methods with significant level $\alpha = 0.01$ for the wavelet and CBS methods. Under such a low noise level, it is not surprising to see all three methods performed well. The CGHseg method had the perfect detection while the CBS had slightly more false positives than both the wavelet and the CGHseg methods. For the CGHseg method, the choice of the threshold S is not critical when the noise level is low. The results were the same for both default S and $S = 2$. We observed that the number of false positives increased with the number of change points for the CBS method. This is mainly because the CBS method uses the sequential testing approach which is not taken into account in the hypothesis testing. It is interesting to note that the CBS method tended to have more even number of false positives than odd number of false positives. This is because the CBS method tests square waves. Both the wavelet approaches gave very close results except at $R = 6$ where permuting $\hat{\epsilon}$ had slightly more false positives than permuting W_1. This is because the fitted residuals had slightly heavier tail than the true error when having more aberration regions. Figure 5.1 shows the plot for both fitted f and true f and histograms as well as the QQ-plot for both fitted residuals $\hat{\epsilon}$ and true ϵ.

Table 5.4, 5.5 and 5.6 display the simulation results at $\sigma = 0.4$ for three methods with significant level $\alpha = 0.01$ for the wavelet and CBS methods. Under this medium noise level, the wavelet and the CGHseg methods performed well in most scenarios and had similar FDR, TPR and number of exact detections. The wavelet method tended to underestimate slightly the number of change points when the aberration region is narrow, for example when $c = 20, R = 4$ and $\sigma = 0.4$, 2.2% of the times the wavelet method estimated the number of change points less than 4. This is a limitation common to all wavelet methods, not just our proposed wavelet method, because of its local property. Specifically, the optimal scale for detecting a narrow aberration region is often at finer scale, however at finer scale wavelet coefficients are dominated by noise which makes the wavelet coefficients less powerfull detecting change points. We also found that under the setting we examined, the performance of the wavelet method was very consistent regardless of the number of change points and the width of the aberration of region, provided that the adjacent change points are far enough, say 20 or more markers apart. This is a big improvement of our proposed wavelet method. Again as we observed in the low noise level case, permuting $\hat{\epsilon}$ had slightly more false positives than permuting W_1.

The CBS method had more false positives than all other methods as we observed in low noise level case when there existed change points, For example, at $\sigma = 0.4$, $c = 80$ and $R = 6$, 6.6% of the times the CBS method detected 8 or more change points while the wavelet and the CGHseg methods never detected 8 or more change points. FDR was 4.2% for the CBS method, comparing 1% and 0.8% for the wavelet method and the CGHseg method respectively. Again, the false positives increased with the number of change points. The CGHseg method with $S = 2$ performed the best with the perfect detection in all the scenarios we examined here. However when the number of aberrated clones is small relative to the total number of markers, with default ($S = 0.75$), the CGHseg method tended to overestimate the number of change points. This was somewhat expected as the profile is close to the global

null situation where we have observed a similar overestimation in the first simulation experiment (Table 5.1). For example when $c = 20, R = 2$ and $\sigma = 0.4$, the CGHseg method performed worst among three with many false positives. FDR equals 3.5% compared with 1.5% for the wavelet method. Also, 5% of the times the CGHseg method estimates 2 or more false positives compared with 0% for the wavelet method.

Table 5.7, 5.8 and 5.9 display the simulation results at $\sigma = 2/3$ for three methods with significant level $\alpha = 0.05$ for the wavelet and CBS methods. Under this very high noise level, the wavelet method did not perform as well as the CBS and CGHseg methods especially when the aberration region is narrow, such as $c = 20$. The wavelet method tended to have more number of false negatives at high noise level. The CGHseg with default S outperformed all the other methods. The CBS method had more number of false positives as we observed at low and medium noise levels.

5.2.3 Simulation 3: Trend Model

In the third simulation experiment we tested the performance of our method under a more challenging situation. The data sets were generated from a model created by Olshen et al. (2004). The model is as follows:

$$Y_i = f(i) + 0.25\sigma \sin(a\pi i) + \epsilon_i,$$

where ϵ_i are i.i.d. $N(0, \sigma^2)$, n=500, and the second term is a sinusoid trend component to make the simulated the data set more realistic and challenging as the periodic trends are often observed in array CGH data (Olshen et al. 2004). The noise parameter σ is set to be either 0.1 or 0.2, and the trend parameter a is chosen to be 0, 0.01 or 0.025 corresponding to no trend and local trend with long and short periods respectively. There are 7 segments with 6 change points along the chromosome. The means of log intensity ratios within segments are given by:

An example of a simulated data set using the trend model is given in Figure 5.2. The simulation results using three different trend parameters are given in Table 5.10.

Table 5.2: Simulation results for detecting evenly-spaced change points, $\sigma = 0.2$, $\alpha = 0.01$. R= # change points, c=# markers in an aberration region, MSE=mean squared error, FDR=#false rejection/# total rejections, TPR= #true rejections/R, Exact=# data sets with correct locations and number of estimated change points. CGHseg have exact same results for default and $S = 2$. c=20 and 40

R	Method	$\leq R-1$	R	$R+1$	$\geq R+2$	MSE	FDR	TPR	Exact
				c=20					
2	Permuting W_1	0	496	4	0	0	0.003	1	496
	Permuting $\hat{\epsilon}$	0	495	5	0	0	0.003	1	495
	CBS	0	483	5	12	0	0.015	1	483
	CGHseg	0	500	0	0	0	0	1	500
4	Permuting W_1	0	492	8	0	0.001	0.003	1	492
	Permuting $\hat{\epsilon}$	0	490	10	0	0.001	0.004	1	490
	CBS	0	477	7	16	0.001	0.013	1	477
	CGHseg	0	500	0	0	0.001	0	1	500
6	Permuting W_1	0	486	14	0	0.001	0.004	1	486
	Permuting $\hat{\epsilon}$	0	473	27	0	0.001	0.008	1	473
	CBS	0	465	11	24	0.001	0.016	1	465
	CGHseg	0	500	0	0	0.001	0	1	500
				c=40					
2	Permuting W_1	0	496	4	0	0	0.003	1	496
	Permuting $\hat{\epsilon}$	0	495	5	0	0.000	0.003	1	495
	CBS	0	484	0	16	0	0.016	1	484
	CGHseg	0	500	0	0	0	0	1	500
4	Permuting W_1	0	489	11	0	0.001	0.004	1	489
	Permuting $\hat{\epsilon}$	0	490	10	0	0.001	0.004	1	490
	CBS	0	472	11	17	0.001	0.016	1	472
	CGHseg	0	500	0	0	0.001	0	1	500
6	Permuting W_1	0	493	7	0	0.001	0.002	1	493
	Permuting $\hat{\epsilon}$	0	483	17	0	0.001	0.005	1	483
	CBS	0	467	8	25	0.001	0.015	1	467
	CGHseg	0	500	0	0	0.001	0	1	500

Table 5.3: Simulation results for detecting evenly-spaced change points, $\sigma = 0.2$, $\alpha = 0.01$. R= # change points, c=# markers in an aberration region, MSE=mean squared error, FDR=#false rejection/# total rejections, TPR= #true rejections/R, Exact=# data sets with correct locations and number of estimated change points. CGHseg have exact same results for default and $S = 2$. c=80

R	Method	$\leq R-1$	R	$R+1$	$\geq R+2$	MSE	FDR	TPR	Exact
				c=80					
2	Permuting W_1	0	483	17	0	0	0.011	1	483
	Permuting $\hat{\epsilon}$	0	478	23	0	0	0.015	1	478
	CBS	0	485	1	14	0	0.015	1	485
	CGHseg	0	500	0	0	0	0	1	500
4	Permuting W_1	0	484	16	0	0.001	0.006	1	484
	Permuting $\hat{\epsilon}$	0	490	10	0	0	0.004	1	490
	CBS	0	474	4	22	0.001	0.016	1	474
	CGHseg	0	500	0	0	0	0	1	500
6	Permuting W_1	0	490	10	0	0.001	0.003	1	490
	Permuting $\hat{\epsilon}$	0	480	20	0	0.001	0.006	1	480
	CBS	0	459	10	31	0.001	0.019	1	459
	CGHseg	0	500	0	0	0.001	0	1	500

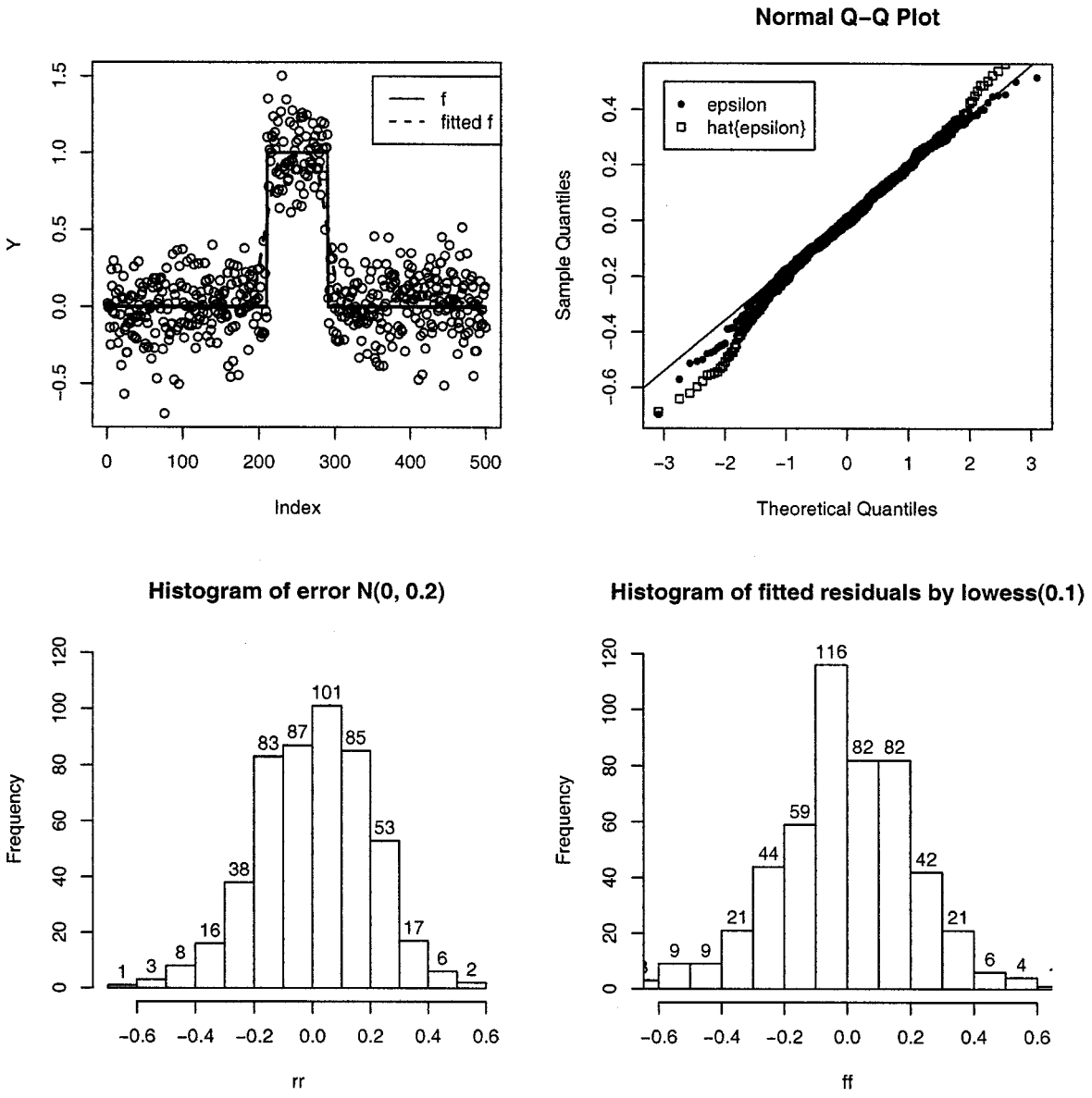


Figure 5.1: The plot for both fitted f and true f and histograms as well as the QQ-plot for both fitted residuals $\hat{\epsilon}$ and true ϵ.

Table 5.4: Simulation results for detecting evenly-spaced change points, $\sigma = 0.4$. R= # change points, MSE=mean squared error, FDR=#false rejection/# total rejections, TPR= #true rejections/R, Exact=# data sets with all the change points correctly estimated. If CGHseg have same results for default and $S = 2$, only one row listed. c=20

R	Method	$\leq R-1$	R	$R+1$	$\geq R+2$	MSE	FDR	TPR	Exact
				c=20					
2	Permuting W_1	3	491	6	0	0.002	0.015	0.988	484
	Permuting $\hat{\epsilon}$	5	488	7	0	0.003	0.022	0.980	475
	CBS	0	483	5	12	0.002	0.022	0.993	476
	CGHseg(default)	0	470	5	25	0.002	0.035	0.994	464
	CGHseg(S=2)	0	500	0	0	0.002	0.006	0.994	494
4	Permuting W_1	11	480	9	0	0.005	0.015	0.984	465
	Permuting $\hat{\epsilon}$	10	478	12	0	0.005	0.016	0.985	465
	CBS	0	474	8	18	0.005	0.022	0.992	462
	CGHseg	0	497	1	2	0.004	0.009	0.993	483
	CGHseg(S=2)	0	500	0	0	0.004	0.007	0.993	486
6	Permuting W_1	11	480	9	0	0.007	0.015	0.984	446
	Permuting $\hat{\epsilon}$	15	475	10	0	0.007	0.015	0.983	440
	CBS	0	466	11	23	0.006	0.025	0.989	439
	CGHseg	0	498	1	1	0.006	0.012	0.989	465
	CGHseg(S=2)	0	500	0	0	0.006	0.011	0.989	467

Table 5.5: Simulation results for detecting evenly-spaced change points, $\sigma = 0.4$. R= # change points, MSE=mean squared error, FDR=#false rejection/# total rejections, TPR= #true rejections/R, Exact=# data sets with all the change points correctly estimated. If CGHseg have same results for default and $S = 2$, only one row listed. $c = 40$

R	Method	$\leq R-1$	R	$R+1$	$\geq R+2$	MSE	FDR	TPR	Exact
				c=40					
2	Permuting W_1	0	498	2	0	0.002	0.012	0.989	487
	Permuting $\hat{\epsilon}$	0	495	5	0	0.002	0.013	0.990	485
	CBS	0	483	0	17	0.003	0.028	0.989	472
	CGHseg	0	498	0	2	0.002	0.013	0.989	487
	CGHseg(S=2)	0	500	0	0	0.002	0.011	0.989	489
4	Permuting W_1	0	491	8	1	0.004	0.015	0.989	473
	Permuting $\hat{\epsilon}$	0	483	17	0	0.004	0.015	0.993	470
	CBS	0	469	11	20	0.005	0.028	0.990	452
	CGHseg	0	500	0	0	0.004	0.010	0.990	482
6	Permuting W_1	0	490	10	0	0.006	0.010	0.993	470
	Permuting $\hat{\epsilon}$	0	485	15	0	0.006	0.010	0.994	470
	CBS	0	470	6	24	0.007	0.021	0.993	450
	CGHseg	0	500	0	0	0.006	0.007	0.993	479

Table 5.6: Simulation results for detecting evenly-spaced change points, $\sigma = 0.4$. R= # change points, MSE=mean squared error, FDR=#false rejection/# total rejections, TPR= #true rejections/R, Exact=# data sets with all the change points correctly estimated. If CGHseg have same results for default and $S = 2$, only one row listed. c=80

R	Method	$\leq R-1$	R	$R+1$	$\geq R+2$	MSE	FDR	TPR	Exact
				c=80					
2	Permuting W_1	0	495	5	0	0.002	0.011	0.992	488
	Permuting $\hat{\epsilon}$	0	495	5	0	0.002	0.021	0.983	480
	CBS	0	484	1	15	0.002	0.024	0.992	477
	CGHseg	0	499	0	1	0.002	0.01	0.991	491
	CGHseg(S=2)	0	500	0	0	0.002	0.009	0.991	492
4	Permuting W_1	0	488	11	1	0.004	0.013	0.992	472
	Permuting $\hat{\epsilon}$	0	490	10	0	0.004	0.011	0.993	475
	CBS	0	470	7	23	0.005	0.031	0.987	446
	CGHseg	0	500	0	0	0.004	0.009	0.992	483
6	Permuting W_1	0	492	8	0	0.006	0.01	0.992	471
	Permuting $\hat{\epsilon}$	0	490	10	0	0.006	0.010	0.993	470
	CBS	0	452	15	33	0.008	0.042	0.979	397
	CGHseg	0	500	0	0	0.006	0.008	0.992	478

Table 5.7: Simulation results for $\sigma = 2/3$ and c=20. R= # change points, MSE=mean squared error, FDR=#false rejection/# total rejections, TPR= #true rejections/R, Exact=# data sets with all the change points correctly estimated.

R	Method	$\leq R-1$	R	$R+1$	$\geq R+2$	MSE	FDR	TPR	Exact
				c=20					
2	Permuting W_1	55	113	27	5	0.017	0.229	0.708	92
	Permuting $\hat{\epsilon}$	56	111	28	5	0.017	0.224	0.710	90
	CBS	0	175	3	22	0.009	0.154	0.900	143
	CGHseg(default)	2	165	3	30	0.010	0.180	0.892	132
	CGHseg(S=2)	10	190	0	0	0.009	0.087	0.863	156
4	Permuting W_1	89	74	30	7	0.035	0.251	0.679	43
	Permuting $\hat{\epsilon}$	91	75	30	4	0.035	0.245	0.669	44
	CBS	0	160	7	33	0.017	0.164	0.895	104
	CGHseg	11	180	2	7	0.017	0.119	0.869	114
	CGHseg(S=2)	64	136	0	0	0.026	0.100	0.730	92
6	Permuting W_1	54	69	46	31	0.0582	0.2063	0.62	14
	Permuting $\hat{\epsilon}$	60	65	43	32	0.059	0.2087	0.61	13
	CBS	1	89	8	102	0.0274	0.1982	0.8875	60
	CGHseg	18	176	3	3	0.0257	0.1202	0.8533	88
	CGHseg(S=2)	130	70	0	0	0.0527	0.0738	0.5817	40

Table 5.8: Simulation results for $\sigma = 2/3$ and 40.

R	Method	$\leq R-1$	R	$R+1$	$\geq R+2$	MSE	FDR	TPR	Exact
				c=40					
2	Permuting W_1	11	139	43	7	0.012	0.198	0.863	115
	Permuting $\hat{\epsilon}$	12	140	40	8	0.012	0.198	0.860	117
	CBS	0	171	1	28	0.010	0.181	0.892	133
	CGHseg	0	186	3	11	0.009	0.138	0.895	147
	CGHseg(S=2)	0	200	0	0	0.007	0.105	0.895	160
4	Permuting W_1	22	101	58	19	0.022	0.191	0.866	67
	Permuting $\hat{\epsilon}$	23	102	56	19	0.022	0.188	0.866	68
	CBS	0	155	4	41	0.018	0.176	0.900	101
	CGHseg	0	199	0	1	0.014	0.100	0.902	135
	CGHseg(S=2)	5	195	0	0	0.016	0.099	0.890	132
6	Permuting W_1	23	99	51	27	0.033	0.187	0.857	52
	Permuting $\hat{\epsilon}$	28	93	51	28	0.034	0.188	0.851	50
	CBS	0	134	5	61	0.028	0.197	0.887	65
	CGHseg	0	196	3	1	0.023	0.118	0.884	101
	CGHseg(S=2)	22	178	0	0	0.032	0.107	0.832	94

Table 5.9: Simulation results for $\sigma = 2/3$ and c=80.

R	Method	$\leq R-1$	R	$R+1$	$\geq R+2$	MSE	FDR	TPR	Exact
				c=80					
2	Permuting W_1	13	154	27	6	0.017	0.165	0.858	124
	Permuting $\hat{\epsilon}$	18	150	26	6	0.020	0.165	0.840	122
	CBS	0	169	0	31	0.010	0.172	0.902	138
	CGHseg	0	195	0	5	0.008	0.118	0.895	155
	CGHseg(S=2)	0	200	0	0	0.007	0.105	0.895	160
4	Permuting W_1	23	118	44	15	0.029	0.171	0.871	78
	Permuting $\hat{\epsilon}$	25	118	41	16	0.031	0.178	0.861	75
	CBS	0	148	3	49	0.018	0.193	0.892	92
	CGHseg	0	199	0	1	0.014	0.104	0.897	128
	CGHseg(S=2)	0	200	0	0	0.014	0.102	0.897	129
6	Permuting W_1	33	94	51	22	0.039	0.181	0.850	43
	Permuting $\hat{\epsilon}$	36	95	50	19	0.041	0.172	0.849	44
	CBS	0	128	7	65	0.030	0.228	0.859	51
	CGHseg	0	199	0	1	0.022	0.112	0.889	94
	CGHseg(S=2)	6	194	0	0	0.026	0.108	0.873	93

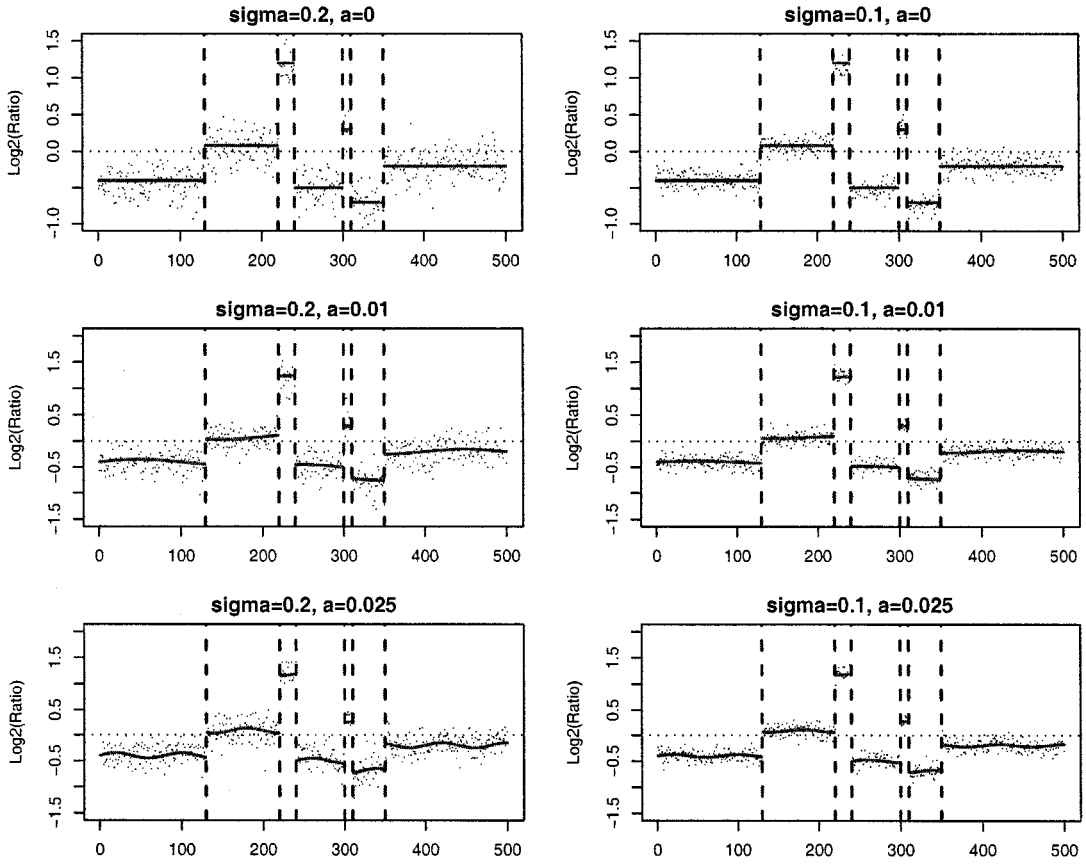


Figure 5.2: An example of a simulated data with $\sigma = 0.2$ and $a = 0$ for no trend (top panel) and $a = 0.025$ for long period trend (bottom panel). The red lines indicate the mean plus trend, i.e., $f(i) + 0.25\sigma \sin(a\pi i)$, the vertical blue lines indicate the locations of change points.

i	1-130	131-220	221-240	241-300	301-310	311-350	351-500
$f(i)$	-0.40	0.08	1.20	-0.50	0.30	-0.70	-0.20

We found that both the CGHseg (default) and the wavelet method outperformed the CBS method in no trend, long and short period trend models. When the noise level was high ($\sigma = 0.2$), the wavelet method had slightly larger FDR compared with the CGHseg method. Both the CGHseg and the wavelet methods were robust to local trend in the sense that MSE, FDR, TPR and the number of exact detections did not appear to change with the trend parameter a. The permuting W_1 approach seemed perform slightly better than the permuting $\hat{\epsilon}$ approach with larger TPR and the number of exact detections. The CBS method tended to overestimate the number of change points when there existed a local trend. The FDR increased with a and the number of exact change points detected decreased with a. We also found that the CGHseg method with $S = 2$ had much less power of detecting change point when the noise level was high ($\sigma = 0.2$) regardless of the values of the trend parameter a.

5.3 Simulation Under Non-normal Distribution

Recall that the null distribution of the test statistic is estimated using permutation method. When ϵ is normally distributed, we showed that permuting W_1 gives a good approximate estimation of the null distribution given that the number of change points is much less than the number of markers. However, when ϵ is not normally distributed, such as Cauchy, the test based on permutation of W_1 will be very anticonservative because the kurtosis of W_1 is only half of the kurtosis of the true error distribution. As a result the number of change points would be overestimated. So alternatively we estimate the null distribution by permuting the $\hat{\epsilon}$'s, the residuals from lowess smoothing. In this section we compared the performance of the wavelet method with the CBS and the CGHseg methods when ϵ was not normally distributed. We

Table 5.10: Simulation results under the trend model, $\alpha = 0.01$ for the wavelet and the CBS methods.

			number of change points								
σ	a	Method	4-	5	**6**	7	8+	MSE	FDR	TPR	#Exact
0.1	0	Permute W_1	0	0	490	10	0	0	0.003	1	490
		Permute $\hat{\epsilon}$	2	100	398	0	0	0.003	0.000	0.965	398
		CBS	0	0	472	11	17	0	0.012	1	472
		CGHseg	0	0	500	0	0	0	0	1	500
	0.01	Permute W_1	0	0	486	14	0	0	0.004	1	486
		Permute $\hat{\epsilon}$	2	107	391	0	0	0.003	0.000	0.963	391
		CBS	0	0	407	35	58	0	0.041	1	407
		CGHseg	0	0	500	0	0	0	0	1	500
	0.025	Permute W_1	0	0	487	13	0	0	0.004	1	487
		Permute $\hat{\epsilon}$	2	72	426	0	0	0.002	0.000	0.974	426
		CBS	0	0	394	22	84	0.001	0.051	1	394
		CGHseg	0	0	500	0	0	0	0	1	500
0.2	0	Permute W_1	0	0	436	57	7	0.001	0.023	0.997	428
		Permute $\hat{\epsilon}$	0	85	410	5	0	0.003	0.008	0.966	405
		CBS	0	0	450	34	16	0.001	0.023	0.995	439
		CGHseg	0	0	500	0	0	0.001	0.003	0.997	492
		CGHseg(S=2)	31	0	469	0	0	0.004	0.002	0.958	462
	0.01	Permute W_1	0	6	437	52	5	0.002	0.022	0.993	423
		Permute $\hat{\epsilon}$	5	150	343	2	0	0.005	0.008	0.940	328
		CBS	0	0	390	54	56	0.002	0.048	0.997	384
		CGHseg	0	0	500	0	0	0.002	0.004	0.996	487
		CGHseg(S=2)	136	0	364	0	0	0.015	0.003	0.816	355
	0.025	Permute W_1	0	7	432	54	7	0.002	0.022	0.994	422
		Permute $\hat{\epsilon}$	2	75	413	10	0	0.004	0.005	0.971	408
		CBS	0	0	376	42	82	0.003	0.06	0.995	364
		CGHseg	0	0	500	0	0	0.002	0.003	0.997	491
		CGHseg(S=2)	142	0	358	0	0	0.014	0.001	0.818	354

evaluated two approaches of estimating the null distribution– permuting W_1 and $\hat{\epsilon}$. We also explored the effect of window size in lowess smoothing. The window size for lowess smoothing is 0.10 if not otherwise specified. We started with the global null situation to examine the type I error rate and then followed by the power evaluation under an evenly-spaced change points model.

For the global null situation where $f = 0$, we generated the errors i.i.d from t distributions. Significant level $\alpha = 0.01$ was used for the wavelet and CBS methods and the default $S = 0.75$ was used for CGHseg. Table 5.11 shows the simulation results based on 500 simulated data sets. The CGHseg method greatly overestimated the number of change points and the type I error rate increased as the degrees of freedom decreased, where the tail gets heavier. The penalized likelihood function is based on the normal assumption, so when the assumption is not valid anymore, the CGHseg method performed very poorly. The CBS method performed consistently very well and the family-wise type I error rates were below or close to 0.01. The wavelet method had the correct type I error rates only if the null distribution was estimated by permuting the $\hat{\epsilon}$. Type I error rates were correct at both window size 0.1 and 0.05. As expected, the null distributions estimated by permuting W_1 was very anticonservative and the type I error rate increased as the degreed of freedom decreased. For t distribution at significant level 0.01, the simulation results showed that under the global null the type I error rate by permuting W_1 was less than 0.01 when degree freedom was greater than 3, but it became greater than 0.01 when degree freedom was less than 4 (Table 5.11).

We then investigated the power of our method in the presence of change points. We generated chromosome profiles with different degree of freedom, number of change points and aberration width. The aberration regions were evenly spaced along the chromosome. Function $f(i)$ was set to be either 0 or 1 corresponding to no change or copy number gains. The number of change points (R) varies among 2, 4, and 6. The width of each aberration region (c) increases from 20, 40 to 80 markers. For each

Table 5.11: Simulation results under global null for $\epsilon \sim t$, $\alpha = 0.01$

		# of change points						
df	Method	0	1	2	3	4	5+	Type I error
3	Permute W_1	463	36	1	0	0	0	0.074
	Permute $\hat{\epsilon}(0.10)$	498	2	0	0	0	0	0.004
	Permute $\hat{\epsilon}(0.05)$	498	1	1	0	0	0	0.004
	CBS	494	0	6	0	0	0	0.012
	CGHseg	6	1	75	1	109	308	0.988
2	Permute W_1	439	56	5	0	0	0	0.122
	Permute $\hat{\epsilon}(0.10)$	496	4	0	0	0	0	0.008
	Permute $\hat{\epsilon}(0.05)$	494	5	1	0	0	0	0.012
	CBS	498	0	2	0	0	0	0.004
	CGHseg	2	0	35	1	82	380	0.996
1	Permute W_1	306	182	11	1	0	0	0.388
	Permute $\hat{\epsilon}$	493	6	1	0	0	0	0.014
	Permute $\hat{\epsilon}(0.05)$	497	2	1	0	0	0	0.006
	CBS	495	0	5	0	0	0	0.010
	CGHseg	0	0	10	0	38	452	1.000

scenario, 500 simulated datasets with a sample of $n = 500$ markers were generated such that Y_is are independently t distributed with df=3. Table 5.12, 5.13 and 5.14 show the simulation results for $c = 20, 40, 80$, respectively. Since the CGHseg is based on the penalized likelihood function with the assumption of normality, it is not surprising to see the method didn't perform well: it tended to overestimate the number of change points under non-normal distribution. Both the wavelet and CBS method were robust to non-normal distribution as we expected.

The wavelet method was very comparable to the CBS method under all the settings we examined except for $c = 80$ and $R = 2$. As we expected, permuting W_1 tended to overestimate the number of change points compared to permuting $\hat{\epsilon}$. We also found that smaller window sizes for lowess smoothing, say 0.05, was slightly more conservative and less powerful for detecting the change points. And there was a trade off between FDR and TPR, so two window sizes, 0.10 and 0.05, gave very comparable values in terms of the number of exact. Based on the settings we examined here, window size 0.10 was a reasonable choice.

5.4 Summary

In this chapter, we evaluated the performance of our 2-scale product wavelet method and compared it with two existing competing methods–the CBS method and the CGHseg method. Our simulation models cover the global null, evenly-spaced change points and unevenly-spaced change points with or without trend under both normal and t distributions. Although the simulation results showed that no one method outperformed the other methods in all the settings we examined, the wavelet method performed well in most of settings in the sense that it has the right type I error rates under global null and is robust to local trend and the non-normal error distribution. However, the wavelet method has low power of detecting change points when the noise level is very high ($\sigma = 2/3$). The CBS method tends to overestimate the number of change points when there exists multiple change points or trend. Since

Table 5.12: Simulation results for t_3, n=500, c=20, $\alpha = 0.01$.

R	Method	≤R-2	R-1	**R**	R+1	≥R+2	MSE	FDR	TPR	#Exact
		c=20								
2	Permute W_1	22	8	432	37	1	0.094	0.036	0.94	427
	Permute $\hat{\epsilon}(0.10)$	24	17	451	7	1	0.113	0.022	0.924	446
	Permute $\hat{\epsilon}(0.05)$	29	25	441	4	1	0.135	0.018	0.908	437
	CBS	40	0	449	5	6	0.114	0.014	0.915	445
	CGHseg	0	0	67	6	427	0.745	0.556	0.998	67
4	Permute W_1	27	7	433	32	1	0.168	0.035	0.941	419
	Permute $\hat{\epsilon}(0.10)$	28	14	444	13	1	0.181	0.028	0.936	430
	Permute $\hat{\epsilon}(0.05)$	36	19	441	3	1	0.217	0.027	0.915	430
	CBS	44	0	441	9	6	0.215	0.014	0.908	428
	CGHseg	0	0	121	6	373	0.69	0.338	0.995	116
6	Permute W_1	30	4	433	30	3	0.222	0.024	0.949	423
	Permute $\hat{\epsilon}(0.10)$	26	7	434	29	4	0.213	0.028	0.954	423
	Permute $\hat{\epsilon}(0.05)$	37	17	436	9	1	0.281	0.023	0.930	428
	CBS	48	0	438	10	4	0.319	0.007	0.906	432
	CGHseg	0	0	170	15	315	0.658	0.216	0.996	166

Table 5.13: Simulation results for t_3, n=500, c=40, $\alpha = 0.01$.

R	Method	≤R-2	R-1	**R**	R+1	≥R+2	MSE	FDR	Power	#Exact
		c=40								
2	Permute W_1	8	3	446	42	1	0.075	0.037	0.973	439
	Permute $\hat{\epsilon}(0.10)$	10	7	474	8	1	0.094	0.016	0.964	466
	Permute $\hat{\epsilon}(0.05)$	10	9	478	2	1	0.101	0.014	0.961	470
	CBS	13	0	479	3	5	0.085	0.016	0.965	470
	CGHseg	0	0	106	6	388	0.677	0.481	0.994	105
4	Permute W_1	12	4	456	27	1	0.145	0.021	0.971	442
	Permute $\hat{\epsilon}(0.10)$	12	0	473	14	1	0.136	0.018	0.972	458
	Permute $\hat{\epsilon}(0.05)$	15	8	469	7	1	0.178	0.013	0.963	456
	CBS	16	0	471	3	10	0.169	0.015	0.961	457
	CGHseg	0	0	191	11	298	0.61	0.255	0.995	185
6	Permute W_1	12	3	462	22	1	0.191	0.015	0.973	441
	Permute $\hat{\epsilon}(0.10)$	12	2	467	18	1	0.182	0.015	0.976	446
	Permute $\hat{\epsilon}(0.05)$	16	5	470	8	1	0.230	0.013	0.967	452
	CBS	25	0	458	11	6	0.3	0.011	0.948	444
	CGHseg	0	0	231	14	255	0.601	0.169	0.993	222

Table 5.14: Simulation results for t_3, n=500, c=80, $\alpha = 0.01$.

R	Method	≤R-2	R-1	**R**	R+1	≥R+2	MSE	FDR	Power	#Exact
						c=80				
2	Permute W_1	5	7	448	38	2	0.111	0.033	0.977	443
	Permute $\hat{\epsilon}(0.10)$	8	7	475	9	1	0.128	0.012	0.972	470
	Permute $\hat{\epsilon}(0.05)$	13	4	480	2	1	0.144	0.007	0.965	475
	CBS	5	0	492	1	2	0.07	0.009	0.984	486
	CGHseg	0	0	168	4	328	0.614	0.388	0.995	167
4	Permute W_1	10	1	458	29	2	0.176	0.021	0.974	442
	Permute $\hat{\epsilon}(0.10)$	9	5	473	12	1	0.181	0.016	0.974	456
	Permute $\hat{\epsilon}(0.05)$	15	4	474	6	1	0.234	0.012	0.964	458
	CBS	9	0	478	5	8	0.169	0.017	0.973	459
	CGHseg	0	0	245	8	247	0.566	0.205	0.993	236
6	Permute W_1	10	5	455	29	1	0.222	0.015	0.976	438
	Permute $\hat{\epsilon}(0.10)$	9	3	465	22	1	0.2	0.015	0.978	447
	Permute $\hat{\epsilon}(0.05)$	14	5	472	8	1	0.261	0.009	0.970	455
	CBS	15	0	453	16	16	0.293	0.021	0.962	428
	CGHseg	0	0	258	16	226	0.573	0.145	0.996	253

the CBS method is also a distribution-free method, it performs well when the error distribution is not normally distributed. Since the CGHseg method is a parametric method, it performs very poorly when the error distribution is not normal. In addition the CGHseg method tends to overestimate the number of change points under global null or a setting close to global null. However, when the error distribution is indeed normally distributed, the CGHseg outperforms both the wavelet and CBS methods in both trend and no trend settings especially when the noise level is high.

Chapter 6

SUMMARY AND FUTURE RESEARCH

6.1 Summary

In this dissertation, we proposed a novel non-parametric approach for detecting change points using wavelet transform. Because wavelet transform has good local-frequency localization, it has advantage over other smoothing methods, such as Fourier transformation and kernel smoothing in detecting sharp changes (Donoho et al. 1994, Wang 1995). The maximum of 2-scale wavelet products across scales was proposed as a novel test statistic for detecting change points. This new test statistic was motivated by combining information across scales to improve power. The justification of using 2-scale product was made by comparing 2-scale product with 2-scale sum and single-level wavelet coefficients. We then developed two non-parametric approaches for estimating the null distribution, including permuting wavelet coefficients at finest scale W_1 and permuting $\hat{\epsilon}$. Individual p-value for each marker locus was estimated using step-down maxT permutation algorithm. However, due to autocorrelations between adjacent wavelet coefficients, a direct application of multiple testing procedures detects not only the true change point but also several marker loci that are around the true change point, depending on the width of μ_r and the scales used in the test statistics. To avoid this, we proposed to test locations at which local maxima occur. The estimated change points were then shown to be consistent for a step function f and the Haar wavelet.

Finally, we applied the 2-scale wavelet product method to two real data sets and performed an comprehensive simulation study to investigate the performance of the wavelet method compared with two existing competing methods–the CBS method

and the CGHseg method. The CBS method is a sequential testing approach which detects change points iteratively using a two-sample t-test statistic. The CGHseg is a model selection approach which estimates the number of change points by maximizing a data-driven penalized likelihood function. Among these methods, only the wavelet method provides the estimation of adjusted p-values for each marker locus. The adjusted p-values provide some flexibility for investigators to call change points at their chosen significant level. All three methods gave comparable results for the two real data sets, the Coriel data and the breast cancer data. The Coriel data suggested that the wavelet method has the least false positives while the CBS method has the most. Although the simulation results showed that no one method outperformed the other methods in all the settings we examined, the wavelet method performed well in most of settings in the sense that it has the right type I error rates under global null and is robust to local trend and the non-normal error distribution. However, the wavelet method has low power of detecting change points when the noise level is very high. In other words, the wavelet method may fail to detect change points for very noisy data. However, with the improvement of the "technology" which includes array-technology and better DNA extraction methods from archived samples and single cell, we expect that the noise level of the data continues to reduce. Moreover, as the marker density increases on the array, the resolution of the locations of change points becomes finer which also enhances the power for detecting the change points. The CBS method tends to overestimate the number of change points when there exists multiple change points or trend. Since the CBS method is also a non-parametric method, it performs well when the error distribution is not normally distributed. Since the likelihood function in the CGHseg method is based on normal assumption, it performs very poorly when the error distribution is not normal. In addition the CGHseg method tends to overestimate the number of change points under global null or a setting close to global null. However, when the error distribution is indeed normally distributed, the CGHseg outperforms both the wavelet and CBS methods in both trend and no

trend settings especially when the noise level is high.

The number of false positives by the wavelet method is dependent on how the local maximum is defined. Mathematically, a real-valued function f defined on the real line have a local maximum at the point x^*, if there exists some $c > 0$, such that $f(x^*) \geq f(x)$ when $|x - x^*| \leq c$. However, for a discrete f, it is not trivial to find local maximum as it is dependent on c. Our proposed wavelet method detect local maximum by taking wavelet transform at certain level j and searching for zero crossing from positive to negative. The higher the level j, the fewer local maxima as the transformed data get smoother with j increasing. Therefore a smaller j tend to give more false positives if clones around the true change points happen to be local maxima. On the other hand, a larger j might miss the change points at narrow regions as the local maximum at change point might be smoothed out. Our simulations and Coriel data application suggest level 4 is a fairly reasonable choice for searching local maximum. Alternatively we could use hill climbing algorithm, a local search technique, to find local maximum at x^* for $c = 1$ as traversing through the clones. However, this method tends to give more false positives when data is very noisy.

6.2 Future Research

There are many possible directions that our work on the 2-scale wavelet product can be further explored. First, the method is computational intensive in the sense that the p-values are estimated based on Monte Carlo simulation which needs thousands of resampling to get accurate estimate of tail probability. To save computation time, one may adopt sequential testing strategy to speed up the algorithm by stopping resampling earlier if the early replication indicate a very large or small p-value. For a single test, methods have been developed in an attempt to estimate Monte Carlo p-value using sequential testing, such as Besag et al. (1991) and the truncated sequential probability ratio test boundary which has been implemented in MChtest R package by Fay et al. (2007). Venkatraman and Olshen (2007) used this strategy, which

significantly reduced the computation time. However it will be challenging to apply the sequential testing directly to our proposed method as it involves estimating of multiple p-values simultaneously and control for multiple testing. If we look at one test at a time, the stopping boundaries are very likely different for these tests. So it will be an interesting research problem in the future.

It is believed that tumorigenesis is a result of interaction among group of DNA aberrations that may spread on several different chromosomes. Therefore finding how these DNA aberrations interact with each other is an important step in understanding of the genetic network in tumor progression. Such genetic interaction are sometimes called oncogenic pathways. Desper et al. (1999, 2000) and Beerenwinkel et al. (2004) developed an tree model where both internal nodes and leaves are aberration events and the aberration events are temporally ordered. These tree models are special cases of graphical models which in general allows multiple genetic aberrations lead to one subsequent aberration. For array CGH data after the segmentation, we could infer an oncogenic network, a directed acyclic graph (DAG), among these aberrated locus via graphical modeling approach by controlling the overall error rates of false edges or correlations.

The segmentation analysis can perhaps be considered as data pre-processing step. The next step is to evaluate how these segmentation procedures impact the down-stream analyses. Segmentation of copy number changes using the proposed or other similar methods results in stepwise constant function. However, this stepwise constant function does not lend itself for classification of whether a segment is truly aberrant or not, because it is common that some segments have values just above or below zero. The means of segments don't fall at exact values of no change or one copy loss/gain and this is partly due to possible contamination of normal cells and/or heterogeneity of tumor cells. How it influences the downstream analyses is not clear. Focusing on one of the primary analyses–comparison of copy number changes between two groups, we will perform simulation studies to compare following four approaches in terms of

both type I error rates and power. These are: (1) raw data; (2) segment the data using the CBS, proposed or other methods and classify the clones into aberrant or not; (3) smoothing the data without classifying the clones (e.g. wavelet-based method, Hsu et al. 2006; Huang et al. 2007); (4) Bayesian methods that build in the states of no change, gains, losses into modeling. Finally, we will apply our method and these existing methods to the Gene Copy Number Changes and Breast Cancer Survival Study to explore the merits of various methods in real data.

Chapter 7

LOCALLY WEIGHTED TRANSMISSION/DISEQUILIBRIUM TEST (TDT)

High-dimensional single nucleotide polymorphism (SNP) data have become increasingly available due to the advancement of high throughput genotyping technologies. These data enable researchers unprecedented capabilities for localizing regions that may be associated with genetic diseases. An often-used strategy for searching for disease-causing genes is to first perform association analyses using genome-wide microsatellite or SNP markers to identify a candidate marker that may be either the disease susceptible locus or in linkage disequilibrium with the disease locus. As a follow-up, even denser SNP markers than those in genome-wide association studies are sometimes genotyped around the candidate markers so that the location of the disease gene can be further confirmed or refined.

The family-based study design is often chosen in studying genetic association with disease risk because it is robust against spurious association caused by the varying disease prevalence and marker allele frequencies in subpopulations (population stratification). In this chapter, we will study the use of dense markers in enhancing the power to detect disease associated markers. For a single marker, the transmission/disequilibrium test (TDT) (Spielman et al. 1993) is a popular approach in assessing the linkage and linkage disequilibrium between a marker locus and disease loci. The TDT has been extended to multiple tightly linked markers (e.g. Zhao et al. 2000) by constructing haplotypes statistically to account for local dependency in the presence of phase ambiguity.

As an alternative to haplotype-based approaches, we proposed an approach that

weights the contribution of multiple SNPs according to their association with the locus of interest. This approach does not require determination of haplotypes. The idea is similar to kernel smoothing in nonparametric regression methods, where the kernel function is like a sliding window and markers that fall in the window all contribute to the test statistic but with differential weights. The weight here is determined by the distance and correlation of the markers to the locus of interest.

We begin the chapter by describing our proposed locally weighted TDT method. In Section 7.2, we apply the method to the simulated GAW14 data. Then we compare the proposed method with the popular TDT method through simulation studies in Section 7.3. The chapter ends with some concluding remarks and possible extensions in Section 7.4.

7.1 Methods

Consider K case-parent trios in which each individual is genotyped with the same M autosomal markers at $\{t_1, \ldots, t_M\}$. Denote Φ_k the disease status of the kth offspring for $k = 1, \ldots, K$. Let $H(t)$ and $h(t)$ be the two alleles at marker locus t. For simplicity, we use h to denote the rare allele among the affected offspring. This, however, is neither necessary nor consequential. For the kth trio, the transmission status $Y_k(t)$ for paternal alleles at locus t can be described as:

$$Y_k(t) = \begin{cases} 1 & \text{H transmitted and h not-transmitted} \\ -1 & \text{h transmitted and H not-transmitted} \\ 0 & \text{otherwise} \end{cases}$$

Similarly, one can define the maternal transmission status $X_k(t)$. Assuming that there is only one disease locus at t_0 in the region framed by these M markers, the expectation of the transmission status (Liang et al. 2001) is

$$\begin{aligned} \mu(t, t_0) &= E\{Y(t)|\Phi_k = 1\} = E\{X(t)|\Phi_k = 1\} \\ &= (1 - 2\theta_{t,t_0}) E\{Y(t_0)|\Phi = 1\} d(t, t_0) \end{aligned}$$

where $d(t,t_0) = \Pr\{H(t)|H(t_0)\} - \Pr\{H(t)|h(t_0)\}$, a measure for linkage disequilibrium and θ is the recombination fraction. We further assume that there is no imprinting in this data set, that is, $E\{X(t)\} = E\{Y(t)\}$. Denote $C = E\{Y(t_0)|\Phi = 1\}$. One can see that the value of C is determined by the penetrance function and the allele frequencies of disease locus t_0. Under the assumptions of initial complete LD, random mating, and constant $\Pr\{H(t_0)\}$ over time, $d(t,t_0)$ can be expressed as $d(t,t_0) = (1-\theta_{t,t_0})^N \Pr\{h(t)|h(t_0)\}$ (Devlin and Risch 1995). Here N is the number of generations since the introduction of a disease-causing mutation at location t_0. The parameters of interest in the mean function $\mu(t,t_0)$ are C for penetrance, N for the number of generations, and t_0 the location of disease locus. Since $Y(t)$ and $X(t)$ are potentially correlated over M markers, Liang et al. (2001) proposed a generalized estimating equation approach to estimate these parameters. An appealing feature of this approach is that the derived parameter estimates remain valid as long as $\mu(t)$ is correctly specified. Liang et al. (2001) also proposed to test the null hypothesis of no linkage or LD to the region framed by the observed M markers by testing $C = 0$. The test statistic is based on a Wald-type statistic, that is, $\hat{C}^2/var(\hat{C})$, requiring a simultaneous estimation of (t_0, N, C) under the assumption that there is a disease locus in the region. However, this approach has several limitations: (1) t_0 is unidentifiable under the null hypothesis; (2) there is a lack of robustness if the assumption of constant $\Pr\{H(t_0)\}$ over time is not met; and (3) in testing $C = 0$, one would still need to estimate all parameters.

With this consideration we propose to derive a score test statistic for testing $C = 0$ at locus t_0, that is, t_0 is not a disease locus. Based on the Equation 10 in Liang et al. (2001), a test statistic can be derived as

$$T_1 = \sum_{k=1}^{K} \left\{ \frac{\partial}{\partial C}\underline{\mu}(t_0) Cov^{-1}(\underline{Y}_k)\underline{Y}_k + \frac{\partial}{\partial C}\underline{\mu}(t_0) Cov^{-1}(\underline{X}_k)\underline{X}_k \right\},$$

where $\underline{\mu}(t_0) = \{\mu(t_1,t_0), \ldots, \mu(t_M,t_0)\}^T$, $\underline{X}_k = \{X_k(t_1), \ldots, X_k(t_M)\}^T$, $\underline{Y}_k = \{Y_k(t_1), \ldots, Y_k(t_M)\}^T$, and superscript T indicates the transpose. Under the independence

working assumption among M marker loci, the test statistic can be further simplified to

$$T_2 = \sum_{k=1}^{K}\sum_{m=1}^{M}(1-2\theta_{t_m,t_0})d(t_m,t_0)\{X_k(t_m)+Y_k(t_m)\}. \quad (7.1)$$

One could insert $(1-\theta_{t_m,t_0})^N Pr\{h(t_m)|h(t_0)\}$ for $d(t_m,t_0)$, but it would require a good estimation of N as well as the probability of $h(t_m)$ conditional on $h(t_0)$. Instead of estimating $d(t_m,t_0)$ under a population genetic model, which is often unverifiable, an empirical estimate can be used to quantify the concordance between the two marker loci in the affected offspring. Devlin and Risch (1995) provided a comparison of various measures for estimating the LD. Upon a close examination the weight in Equation (7.1) essentially determines how close marker locus t_m is to locus t_0. In other words, if marker locus t_m is closer to locus t_0, it is expected that the transmission status at t_m would contribute more information to the test statistic at t_0. It then seems logical that one should estimate directly the concordance of the transmission status at locus t_m and at locus t_0. Since both $X(t)$ and $Y(t)$ take discrete values, a natural measure for concordance is the kappa statistic, which is defined as the ratio of the difference between the probabilities of expected and observed disagreements to the probability of expected disagreement. Here, the disagreement between the two marker loci would be the probability of one marker locus transmitting the rare allele, h, whereas the other marker has transmitted the common allele, H. Specifically , let $Z_k(t)$ take value -1 if $X_k(t)+Y_k(t)$ is negative, i.e. either both parents transmitting h allele but not H, or one parent transmitting h allele but not H and the other parent is non-informative. Similarly, $Z_k(t)$ takes value 1 if $X_k(t)+Y_k(t)$ is positive. Then one can form a 2×2 table for $Z(t_m)$ and $Z(t_0)$ at loci t_m and t_0 as follows

		$Z(t_0)$		
		1	-1	
$Z(t_m)$	1	a	b	$a+b$
	-1	c	d	$c+d$
		$a+c$	$b+d$	$n=a+b+c+d$

Define $Obs = (b+c)/n$, $Exp = (a+b)(b+d)/n^2 + (a+c)(c+d)/n^2$. Then

$$kappa = \frac{Exp - Obs}{1 - Exp}.$$

A nice feature of kappa is that the proportion of agreements is calculated after excluding chance agreement. The value of kappa statistic ranges from -1 (negative complete linkage disequilibrium) to 1 (positive complete linkage disequilibrium). Clearly, each term in the sum of Equation (7.1) remains unchanged if the allele designation, H versus h, is switched.

It is easy to generalize the test statistic T_2 in a couple of ways. For example, rather than summing over the total M markers in the test statistic, one can also use the markers within a prespecified neighborhood of t_0. In addition, the test statistic T_2 can be extended to accommodate multiple affected siblings. The following statistic describes these extensions:

$$T = \sum_{k=1}^{K} \sum_{i=1}^{n_k} \sum_{m \in \mathcal{B}} (1 - 2\theta_{t_m,t_0}) d(t_m, t_0) \{X_{ki}(t_m) + Y_{ki}(t_m)\}, \tag{7.2}$$

where n_k is the number of affected in the kth family and $\mathcal{B}$ is a pre-specified neighborhood around marker locus t_0. We name test statistic T as the *locally weighted TDT*. The choice of the size of a neighborhood depends on many factors such as the nature of the disease mutation and population under study and the marker density. An examination of inter-marker linkage disequilibrium may help determine the window size. By the central limit theorem, $K^{-1/2}T$ is asymptotically normal with a variance that can be empirically estimated by

$$K^{-1} \sum_{k=1}^{K} \left[\sum_{i=1}^{n_k} \sum_{m \in \mathcal{B}} (1 - 2\theta_{t_m,t_0}) d(t_m, t_0) \{X_{ki}(t_m) + Y_{ki}(t_m)\} \right]^2.$$

To account for the multiple comparisons in the tests, one may combine test statistics of all the markers by taking the maximum and determine its critical values by a simulation-based procedure in the sense that the transmission status for each affected offspring are randomly assigned for a large number of times.

7.2 Analysis of GAW14 data

We illustrate the locally weighted TDT approach on a simulated data from Genetic Analysis Workshop 14 (GAW14, http://www.gaworkshop.org/). We begin with a brief introduction of the GAW14 data and followed by the analysis result compared to TDT test. The detailed description of simulation approach and the genetic model can be found in a report by Greenberg et al. (2005).

7.2.1 Introduction of GAW14 data

The phenotype in this simulated data set is a binary disease, Kofendrerd Personality Disorder (KPD), which is a hypothetical psychiatric syndrome characterized by an overwhelming concern with the meaning of the patient's inner emotions and world view and at the same time subsuming the emotions of others into the self. Nosology for KPD falls into three different groups: (1) "communally-shared emotions" symptoms such as joining/founding cults and fear or discomfort with strangers; (2) behavior-related symptoms such as fascination with automobiles and aversion to walking; (3) anxiety-related symptoms such as morbid anger/fear/terror concerning rain/snow and reluctance to wear clothing appropriate for subjective temperature. All three or combination thereof have been used for diagnosis of KPD. The condition is thought to be genetic in origin, possibly exacerbated by prevailing social conditions. The underlying disease model contained four major genes and two modifier genes. The four major genes interacted with each other to produce three different phenotypes, which were themselves heterogeneous.

In this study we analyzed the data from the Aipotu population with a high prevalence of KPD. The cases were classified as anyone with "notable clusters" of symptoms from any of the three groups as KPD. The families in this data set were ascertained when at least two siblings could be classified under any of the diagnostic groups or any combination. The data that we analyzed here consisted of all affected offspring and their parents from the first replicate of Aipotu study. There were a total of 100 nuclear families with 283 affected offsprings. We had no knowledge of the "answers" at the time when we performed the following analyses.

7.2.2 Results

We first performed a single point linkage analysis using the microsatellite markers genotyped on the affected sib pairs. The microsatellite markers were on average about 7.5cM apart. We found that the LOD scores for marker D3S0124 and D3S0127 on chromosome 3 were 4.51 and 3.06, respectively. Both exceeded the cut-off threshold of lod score 3 for IBD testing. Marker D3S0124 was even beyond 3.6, a critical value suggested by Lander and Kruglyak (1995) for genome-wide significance. Based on these results, we subsequently purchased 7 packets of basic SNP markers in this region flanked by microsatellite markers D3S0123 and D3S0127. This covers all available SNP markers for the telomere end of chromosome 3. Excluding the microsatellite markers, there were a total of 134 SNP markers covering about 35cM in genetic distance.

We applied our proposed test statistics to these 134 SNP markers using all affected-parent trios. We used 2 markers on each side as a pre-determined neighborhood within which the marker contributions to the test statistic are considered. Figure 7.1 shows the χ^2 values for the TDT (left panel) and locally weighted TDT (right panel). The lower two plots are the enlarged plots for the 10 markers toward the telomere, some of which showed significant associations with the disease occurrence. The critical values corresponding to level 0.05, indicated by the horizontal lines in the plots, were

obtained from the permutation procedure described in Section 7.1. They were 11.2 and 11.0 for the TDT and locally weighted TDT test statistics, respectively. SNP marker B03T3056 was the only marker that exceeded the threshold for the TDT. Using the locally weighted test statistic both markers B03T3056 and B03T3057 showed significant associations with the disease occurrence in the affected offspring. We have also analyzed the data using larger size windows up to all markers. Although the peak at marker B03T3056 remains significant for all window sizes, the magnitude of the peak decreases with increasing window size. Further examination of the pairwise linkage disequilibrium (LD) using Haploview (Mark Daly's lab, Whitehead Institute for Biomedical Research, Barrett et al. 2005) indicated an overall weak inter-marker LD with exception for marker B03T3056 and B03T3057. The LD measure D' between these two markers is 0.60 and the 95% confidence interval is (0.53,0.67).

To study whether SNP B03T3056 and B03T3057 partly explain the linkage peaks at microsatellite markers D3S0124 and D3S0127, we then included SNP B03T3056 and B03T3057 separately as covariate in the single-point linkage analysis using the same affected sib pairs as in the initial linkage analysis scan (Table 7.1) (Houwing-Duistermaat et al. 2005). The overall LOD score for microsatellite marker D3S0127 and SNP B03T3056 was increased 1.10 compared to the LOD score for the microsatellite marker only ($p = 0.02$). But the increase in the overall LOD score was fairly minimal when SNP B03T3057 was considered. For microsatellite marker D3S0124, only a moderate improvement was observed in the overall LOD scores after including the SNPs. Based on these results, we postulate that SNP B03T3056 only partially explains the linkage signal at microsatellite markers D3S0124 and D3S0127 and other unknown genes may still be present in the region. The truth is that one of the disease loci, D2, is located between B03T3067 and C04R0282 at the end of the chromosome 3, so C04R0282 should not be linked to B03T3067 or the disease. There is LD in the region between SNP B03T3056 and B03T3068 (Greenberg et al. 2005). Therefore our conclusion is confirmed by the underlying truth.

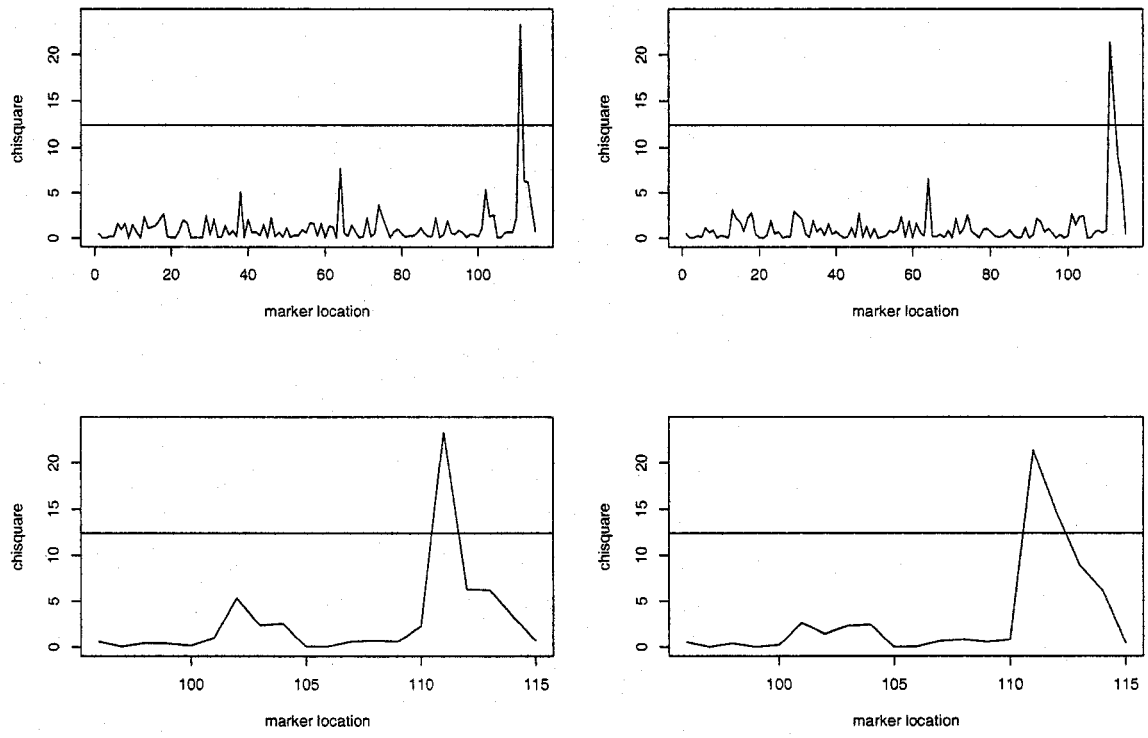


Figure 7.1: χ^2 for TDT (left) and smoothed TDT (right).

Table 7.1: Single-point LOD scores using affected sibpairs model:

	Microsatellite Marker				
	D3S0123	D3S0124	D3S0125	D3S0126	D3S0127
no SNP	1.65	4.51	0.41	1.97	3.06
B03T3056	2.25	5.15	0.43	2.15	4.16
B03T3057	2.46	4.64	0.58	2.46	3.17

7.3 Simulation Studies

To further evaluate the performance of our proposed locally weighted TDT method, we conducted a series of simulations to compare with the TDT test. For the locally weighted TDT, we chose three different window sizes, size=1, 3 or all markers on each side. We accounted for the multiple comparisons in all tests by determining the appropriate critical values using the following randomization procedure: 1) simulate the data set by randomly assigning parental alleles to each affected offspring with equal probability; 2) calculate the test statistics for all the markers; 3) take the maximum of the test statistics. Repeat the above steps for a large number of times, say 5000. Then the critical value at significant level α is $(1-\alpha)$ quantile.

7.3.1 Simulation Settings

In the simulations, we considered the affected-trio family design and the families were ascertained through the affected offspring. A simulation program by Hudson (2002), ms, were used to generate a random sample of 2000 single nucleotide polymorphism (SNP) haplotypes with 17 SNPs from a population. The ms program generates samples using the standard coalescent model in which the random genealogy of the sample is generated first and then mutations are randomly placed on the genealogy. The coalescent model assumes the marker sites are infinite, thus multiple-hits and back mutations do not occur. Then for each data set, we randomly selected 800 haplotypes out of 2000 haplotypes pool and randomly paired haplotypes to form 400 genotypes. Random mating were assumed in generating father and mother pair, then all four parental haplotype alleles had equal probability transmitted to their offspring. Hence we ended up with total of 200 trio families.

The disease status was simulated based on the penetrance functions. We considered two models: single-locus model and two-locus model. For the single-locus model, the penetrance functions were $P(\Phi = 1|h/h) = P(\Phi = 1|h/H) = a_1$ and

$P(\Phi = 1|H/H) = a_2$ where h denotes the high risk dominant allele. The average allele frequency for high risk allel h is 0.15. For the two-locus model, there are four haplotypes: HH, Hh, hH, hh. The high risk haplotype was hH and $P(\Phi = 1|hH) = a_1$, all other haplotypes were considered low risk with the same risk, $P(\Phi = 1|other) = a_2$. The Haldane mapping function was used to connect the recombination fraction and the map distance measured in centimorgans, $\theta = (1 - e^{-0.02|t-t_0|})/2$. After the disease status was simulated for each subject, trio families with affected offspring were ascertained. The actual number of affected-trio families may vary from one simulated data set to another. The mean number of families is 98 with $sd = 6.7$.

For each data set, both the TDT and locally weighted TDT were calculated. A total of 500 data sets were simulated.

7.3.2 Simulation Results

We first verified that all the test statistics had the correct nominal type I error rates. We assumed that all genotypes had the same disease risk and the disease prevalence was 0.3. At significant level 0.05, the type I error rates were 0.044, 0.050, 0.047 and 0.046 for TDT and locally weighted TDT at window size = all, 3, 1, respectively. We can see that the type I error rates were at nominal level for all the test statistics.

In Table 7.2, we summarized the power for all four test statistics under single-locus and two-locus models. When the disease loci were among the markers, the TDT test had the highest power among all four test statistics regardless of whether it is one- or two-locus models. In this case, the power of locally weighted TDT decreased as the window size increases. The test statistic with one marker on each side has the power very close to the TDT. We also examined the power when the disease loci were not among the SNPs though with high LD. The power was defined as the probability of detecting any of the loci adjacent to the unobserved disease locus. The locally weighted TDT had the higher power than the TDT regardless of its window size.

To explore why the locally weighted TDT test was less powerfull than the TDT

Table 7.2: Power comparison between TDT and locally weighted TDT under single-locus and two-locus models. T(All), T(3) and T(1) are for locally weighted TDT with window size equal to all, 3 and 1 markers.

	Single-locus model				Two-locus model			
	TDT	T(All)	T(3)	T(1)	TDT	T(All)	T(3)	T(1)
Disease locus observed								
$a_1 = 0.9, a_2 = 0.1$	0.998	0.972	0.980	0.994	1.000	0.976	0.988	0.998
$a_1 = 0.5, a_2 = 0.1$	0.704	0.532	0.580	0.642	0.738	0.588	0.634	0.702
Disease locus not observed								
$a_1 = 0.9, a_2 = 0.1$	0.324	0.388	0.422	0.432	0.320	0.366	0.368	0.404
$a_1 = 0.5, a_2 = 0.1$	0.078	0.128	0.120	0.132	0.112	0.158	0.140	0.128

test in our simulation settings, we calculated the LD coefficients for each pair of markers in one simulated data set using Haploview program. The LD plot is shown in Figure 7.2. The darkness of colors increases with LD values, where darker color indicates a larger LD. The number within each square indicates the LD*100 between 2 markers. The missing number means a complete LD (i.e., LD=1). The adjactent markers with very strong LD are divided into blocks. A total of 5 blocks are found among 17 markers. The disease marker, marker 10, is in block 3 which includes marker 11. In the two-loci model, the disease markers are marker 10 and 12 which are not in the same block. We can see that the overall LD in this simulated data set is not very strong. In particualr, the LDs between the disease marker 10 and some neighboring markers are not strong, for example, the LD is 0.66 for marker 10 and 9 and 0.56 for marker 10 and 13. The slight loss of power is attributed to the overall weak LD among the markers especially for the large window size, such as window size equal to all markers.

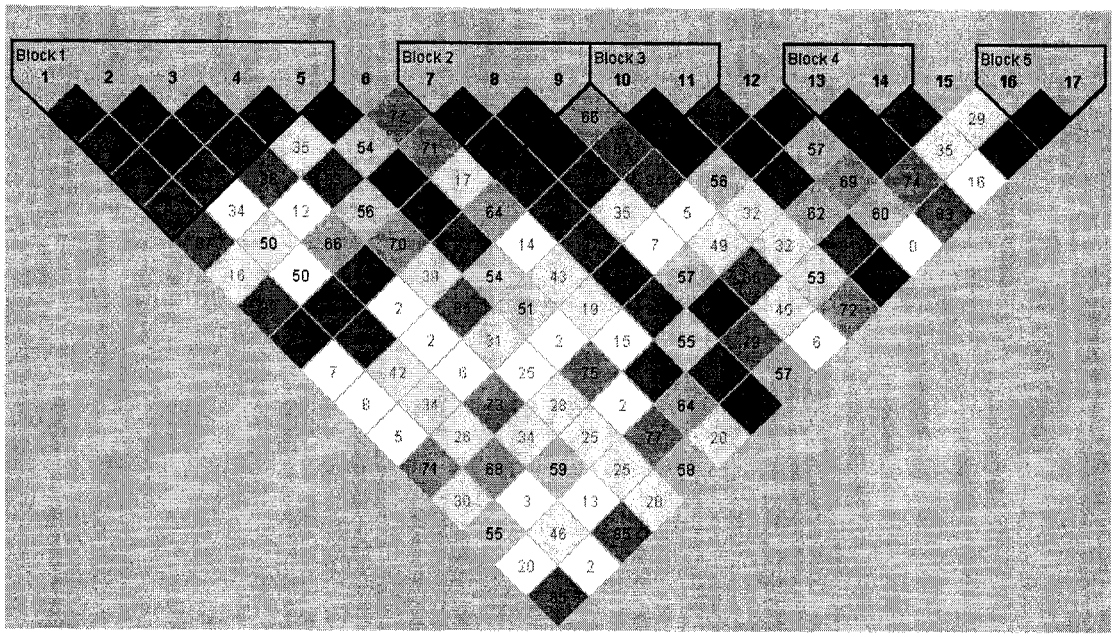


Figure 7.2: The LD plot for each pair of markers. The number within each square indicates the LD*100 between 2 markers. The missing number means a complete LD (i.e., LD=1). The marker IDs are listed at the top of the plot. The darkness of color increases with LD values, where darker color indicates a larger LD. The adjactent markers with very strong LD are divided into blocks.

7.4 Conclusions

We proposed a method that accounts for the local dependencies among adjacent markers. We applied it to the GAW14 data set and showed that the proposed test statistics yield a smoothed signal between marker SNP B03T3056 and SNP B03T3057. The proposed method did not show much more power than the conventional TDT, in part due to an overall weak inter-marker linkage disequilibrium in this SNP data set. We also evaluated the performance of the proposed method in simulation studies. Under the settings we examined, the proposed method did not show more power than the TDT when the disease loci were among the markers. However, the proposed method had slightly more power than TDT when the disease loci were not among the markers. Further work on the performance of the proposed method under a wide range of scenarios will be needed. The choice of window size in the locally weighted test statistic depends on the nature of the disease mutation and population under study as well as marker density. One possible choice is to first examine an overall LD in the region and use it as guidance for determining the window size. A strong LD suggests a wide window size and vice versa. Another possible choice is to calculate the locally weighted test statistics for a few different window sizes and combine them into one test statistic by taking the maximum. The appropriate critical threshold value needs to be adjusted for such a combinational test statistic. Here we are testing the null hypothesis $C = 0$. An alternative may be to construct confidence bands for $\widehat{C}$, turning the testing problem into an estimation one. The region for which the confidence bands do not include 0 is likely an indication for a disease locus. An advantage of such a approach is that it provides a confidence interval for which the disease locus might reside.

BIBLIOGRAPHY

[1] F. Abramovich, T. C. Bailey, and T. Sapatinas. Wavelet analysis and its statistical applications. *The statistician*, 49:1–29, 2000.

[2] A. Antoniadis and I. Gijbels. Detecting abrupt changes by wavelet methods. *Nonparametric Statistics*, 14:7–29, 2002.

[3] P. Bao and L. Zhang. Noise reduction for magnetic resonance images via adaptive multiscale products thresholding. *IEEE Trans. Medical Imaging*, 22:1089–1099, 2003.

[4] J.C. Barrett, B. Fry, J. Maller, and M.J. Daly. Haploview: analysis and visualization of ld and haplotype maps. *Bioinformatics*, 21:263–265, 2005.

[5] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a pratical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300, 1995.

[6] Charles K. Chui. *An Introduction to Wavelets.* Academic Press, 1992.

[7] W. S. Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of American Statistical Assocication*, 74, 1979.

[8] R.R. Coifman and D.L. Donoho. Translation-invariant de-noising. In A. Antoniadis and G. Oppenheim, editors, *Wavelets and Statisitcs*, pages 125–50. New York: Springer-Verlag, 1995.

[9] Ingrid Daubechies. *Ten Lectures on Wavelets.* Society for Industrial and Applied Mathematics, 1992.

[10] B. Devlin and N. Risch. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics*, 29:311–322, 1995.

[11] D. L. Donoho and I.M. Johnstone. Ideal spatial adaption by wavelet shrinkage. *Biometrika*, 81, 1994.

[12] D. L. Donoho and I.M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal American Statistical Association*, 90, 1995.

[13] S. Dudoit, J. P. Shaffer, and J. C. Boldrick. Multiple hypothesis testing in microarray experiments. *Statist. Sci.*, 18:71–103, 2003.

[14] P. H. C. Eilers and Rene X. de Menezes. Quantile smoothing of array cgh data. *Bioinformatics*, 21:1146–1153, 2005.

[15] D.A. Engler, G. Mohapatra, D.N. Louis, and R.A. Betensky. A pseudolikelihood approach for simultaneous analysis of array comparative genomic hybridizations (acgh). *Biostatistics*, 7:399–421, 2006.

[16] J. Fridlyand, A.M. Snijders, D. Pinkel, D.G. Albertson, and A.N. Jain. Hidden markov models approach to the analysis of array cgh data. *Journal of multivariate analysis*, 90:132–153, 2004.

[17] Yongchao Ge, Sandrine Dudoit, and Terence P. Speed. Resampling-based multiple testing for microarray data analysis. *Test*, 12:1–77, 2003.

[18] R. Gencay, F. Selcuk, and B. Whitcher. *An Introduction to Wavelets and Other Filtering Methods in Finance and Economics.* Academic Press, 2002.

[19] David A. Greenberg, Junying Zhang, Dvora Shmulewitz, Lisa J Strug, Regina Zimmerman, Veena Singh, and Sudhir Marathe. Construction of the model for the genetic analysis workshop 14 simulated data: genotype-phenotype relationships, gene interaction, linkage, association, disequilibrium, and ascertainment effects for a complex phenotype. *BMC Genetics*, 6(Suppl 1):S3, 2005.

[20] A. Grossmann and J. Morlet. Decomposition of hardy functions into square integrable wavelets of constant shape. *SIAM Journal on Mathematical Analysis*, 15, 1984.

[21] S. Guha, Y. Li, and D. Neuberg. Bayesian hidder markov modeling of arrasy cgh data. *Harvard University Biostatistics Working Paper Series*, 23:http://www.bepress.com/harvardbiostat/paper24, 2006.

[22] J.J. Houwing-Duistermaat, H. W. Uh, H. Putter, and L. Hsu. Modeling the effect of an associated single-nucleotide polymorphism in linkage studies. *BMC Genetics*, 6(Suppl 1):S46, 2005.

[23] L. Hsu, S. G. Self, D. Grove, T. Randolph, K. Wang, J.J. Delrow, L. Loo, and P. Porter. Denoising array-based comparative genomic hybridization data using wavelets. *Biostatistics*, 6:211–226, 2005.

[24] H. Huang, S. Chin, and C. Ginestier et al. A recurrent chromosome breakpoint in breast cancer at the nrg1/neuregulin 1/heregulin gene). *Cancer Research*, 64:6840–6844, 2004.

[25] J. Huang, A. Gusnanto, K. O'Sullivan, J. Staaf, A. Borg, and P. Pawitan. Robust smoothing segmentation approach for array cgh data analysis. *Bioinformatics*, 23:2463–2469, 2007.

[26] R.R. Hudson. Generating samples under a wright-fisher neutral model of genetic variation. *Bioinformatics*, 18, 2002.

[27] P. Hupe, N. Stransky, J. Thiery, F. Radvanyi, and E. Barillot. Analysis of array cgh data: from signal ratio to gain and loss of dna regions. *Bioinformatics*, 20:34133422, 2004.

[28] K. Jong, E. Marchiori, and A. van der Vaart. Chromosomal breakpoint detection in array comparative genomic hybridization data. *Applications of evolutionary computing: Evolutionary computation and bioinformatics*, 2611:54–65, 2003.

[29] W.R. Lai, M.D. Johnson, R. Kucherlapati, and P.J. Park. Comparative analysis of algorithms for identifying amplifications and deletions in array cgh data. *Bioinformatics*, 21:3763–3770, 2005.

[30] E.S. Lander and L. Kruglyak. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nature Genet.*, 11:241–247, 1995.

[31] K.Y. Liang, F.C. Hsu, T.H. Beaty, and K.C. Barnes. Multipoint linkage disequilibrium mapping approach based on the case-parent trio design. *Am. J. Hum. Genet.*, 68:937–50, 2001.

[32] S. Mallat. Theory for multiresolution signal decomposition: the wavelet representation. *IEEE Trans. on PAMI*, 11:674–692, 1989.

[33] S. Mallat and W.L. Hwang. Singularity detection and processing with wavelets. *IEEE Trans. Inform. Theory*, 38:617–643, 1992.

[34] S. Mallat and S. Zhong. Characterization of signals from multiscale edges. *IEEE Trans. Pattern Anal. Machine Intell.*, 14:710–732, 1992.

[35] Stephane Mallat. *A Wavelet Tour of Signal Processing.* Academic Press, 1998.

[36] Yves Meyer. *Wavelets and Operators.* Cambridge University Press, 1992.

[37] Yves Meyer. *Wavelets: Algorithms & Applications.* Society for Industrial and Applied Mathematics, 1993.

[38] K.S. Miller. *Multidimensional gaussian distributions.* New York: Wiley, 1964.

[39] R. Todd Ogden. *Essential Wavelets for Statistical Applications and Data Analysis.* Birkhauser, 1997.

[40] R.T. Ogdena and J.D. Lynch. Bayesian analysis of change-point models. Lect. Notes Statist, 1999.

[41] A. B. Olshen and E. S. Venkatraman. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5(4):557–572, 2004.

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

[43] F. Picard, S. Robin, M. Lavielle, C. Vaisse, and J. J. Daudin. A statistical approach for array cgh data analysis. *BMC Bioinformatics*, 6:27–41, 2005.

[44] D. Pinkel and D.G. Albertson. Array comparative genomic hybridization and its applications in cancer. *Nat. Genet.*, 37:S11–S17, 2005.

[45] D. Pinkel, R. Segraves, D. Sudar, S. Clark, I. Poole, D. Kowbel, C. Collins, W.L. Kuo, C. Chen, Y. Zhai, S.H. Dairkee, B.M. Ljung, J.W. Gray, and D.G. Albertson. High resolution analysis of dna copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.*, 20:207211, 1998.

[46] M. Raimondo and Nader Tajvidi. A peaks over threshold model for change-point detection by wavelets. *Statistica Sinica*, 14:395–412, 2004.

[47] A. Rosenfeld. A nonlinear edge detection technique. *Proc. IEEE*, 3:814–816, 1970.

[48] B.M. Sadler and A. Swami. Analysis of multiscale products for step detection and estimation. *IEEE Trans. Inform. Theory*, 45:1043–1051, 1999.

[49] S. Sardy, D.B. Percival, A.G. Bruce, H-Y. Gao, and W. Stuetzle. Wavelet denoising for unequally spaced data. *Statistics and Computing*, 9:65–75, 1999.

[50] A. Sen and M. Srivastava. On tests for detecting a change in mean. *Annals of Statistics*, 3:98–108, 1975.

[51] Y. Sheng. *Wavelet Transform.* CRC Press, 2000. In: The transforms and applications handbook ed. by A. D. Poularikas.

[52] A.M. Snijders, N. Nowak, R. Segraves, S. Blackwood, N. Brown, J. Conroy, and G. Hamilton *et al.* Assembly of microarrays for genome-wide measurement of DNA copy numbers. *Nat. Genet.*, 29:263264, 2001.

[53] R.S. Spielman, R.E. McGinnis, and W.J. Ewens. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). *Am. J. Hum. Genet.*, 52:506–16, 1993.

[54] R.L. Stallings, P. Nair, J.M. Maris, D. Catchpoole, M. McDermott, A. O'Meara, and F. Breatnach. High-resolution analysis of chromosomal breakpoints and genomic instability identifies ptprd as a candidate tumor suppressor gene in neuroblastoma. *Cancer Research*, 66:3673–3680, 2006.

[55] M. G. Tadesse, J. G. Ibrahim, M. Vannucci, and R. Gentleman. Wavelet thresholding with bayesian false discovery rate control. *Biometrics*, 61, 2005.

[56] R. Tibshirani and P. Wang. Spatial smoothing and hot spot detection for cgh data using the fused lasso. *Biostatistics*, 8:1–12, 2007.

[57] Joris A. Veltman, Eric F. P. M. Schoenmakers, Bert H. Eussen, Irene Janssen, and et al. High-throughput analysis of subtelomeric chromosome rearrangements by use of array-based comparative genomic hybridization. *Am. J. Hum. Genet.*, 70, 2002.

[58] E. S. Venkatraman and A. B. Olshen. A faster circular binary segmentation algorithm for the analysis of array cgh data. *Bioinformatics*, 23:657–663, 2007.

[59] Brani Vidakovic. *Statisitcal Modeling by Wavelets.* Wiley, 1999.

[60] M.P. Wand and M.C. Jones. *Kernel Smoothing.* Chapman, 1995.

[61] P. Wang, Y. Kim, J. Pollack, B. Narasimhan, and R. Tibshirani. A method for calling gains and losses in array cgh data. *Biostatistics*, 6:45–58, 2005.

[62] Y. Wang. Jump and sharp cusp detection by wavelets. *Biometrika*, 82:385–397, 1995.

[63] C.C. Wen, Y.J. WU, Y.H. Huang, W.C. Chen, S.C. Liu, S.S. Jiang, J.L. Juang, C.Y. Lin, W.T. Fang, C.A. Hsiung, and I.S. Chang. A bayes regression approach to array-cgh data. *Statistical Applications in Genetics and Molecular Biology*, 5:Iss.1, 2006.

[64] P.H. Westfall and S.S. Young. *Resampling-based multiple testing: Examples and methods for p-value adjustment.* John Wiley & Sons, 1993.

[65] H. Willenbrock and J. Fridlyand. A comparison study: applying segmentation to array cgh data for downstream analyses. *Bioinformatics*, 21:4084–4091, 2005.

[66] Randy K. Young. *Wavelet Theory and its Applications.* Kluwer Academic, 1993.

[67] Nancy R. Zhang and David O. Siegmund. A modified bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63:22–32, 2007.

[68] H. Zhao, S. Zhang, K.R. Merikangas, M. Trixler, D.B. Wildenauer, F. Sun, and K. Kidd. Transmission/disequilibrium tests using multiple tightly linked markers. *Am. J. Hum. Genet.*, 67:936–46, 2000.

VITA

Xuesong Yu was born in Jiangxi, China on December 4, 1970. She received her BS in Biology from Nanjing University in 1992 and MS in Forestry Genetics from Beijing Forestry University in 1998. After spending two years in the Ph.D. program of Forest Resources at the University of Washington, Seattle, she joined the Ph.D. program in Biostatistics at the University of Washington. In 2002, she received an MS in Biostatistics from the University of Washington.