DiNAMIC - A Method for Assessing the Statistical Signficance of DNA Copy Number Aberrations

by
Vonn Walter

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill
2010

Approved by:

Advisor: Dr. Andrew Nobel

Advisor: Dr. Fred Wright

Reader: Dr. Fei Zou

Reader: Dr. Wei Sun

Reader: Dr. D. Neil Hayes

Abstract

VONN WALTER: DiNAMIC - A Method for Assessing the Statistical Signficance of DNA Copy Number Aberrations.
(Under the direction of Dr. Andrew Nobel and Dr. Fred Wright.)

DNA copy number gains and losses are commonly found in tumor tissue, and some of these aberrations play a role in tumor genesis and development. Although high resolution DNA copy number data can be obtained using array-based techniques, no single method is widely used to distinguish between recurrent and sporadic copy number aberrations. Here we introduce Discovering Copy Number Aberrations Manifested In Cancer (DiNAMIC), a novel method for assessing the statistical significance of recurrent copy number aberrations. DiNAMIC uses two resampling schemes - a permutation method and a bootstrap procedure - both of which largely preserve the correlation structure found in the underlying DNA copy number data. It is important to maintain as much of the correlation structure as possible when resampling, and we believe this may yield additional power to detect recurrent aberrations. Extensive simulation studies show that DiNAMIC controls false positive discoveries in a variety of realistic scenarios. We use DiNAMIC to analyze two publicly available tumor datasets, and our results show that DiNAMIC detects multiple loci that have biological relevance. Although DiNAMIC provides methods for detecting CNAs, loci that exhibit aberrant copy number do not always lie in genes related to the tumor phenotype. Therefore we introduce methods for computing confidence intervals around CNAs. Copy number datasets often contain data obtained from subjects with different subtypes of a given tumor type, and the tumor subtypes can harbor distinct genetic mutations. Because groups of subjects may have similar copy number profiles, we present methods for determining which subjects contribute to a given CNA. Some studies collect both DNA copy number and clinical data from subjects, but no methods for jointly analyzing both data types are widely used. We describe a preliminary testing procedure for comparing the locations of CNAs and loci whose copy number values are highly associated with a given clinical variable.

Acknowledgements

I'd like to begin by acknowledging Fred Wright and Andrew Nobel for the guidance they provided while I was working on this thesis. The breadth and depth of their knowledge never ceases to amaze me, and I feel fortunate to have had the opportunity to work with them. I am also extremely grateful to Fred for providing financial support.

Thanks to Fei Zou, Wei Sun, and Neil Hayes for serving on my thesis committee. I appreciate their willingness to take time out from their busy schedules to help with this work.

Thanks to my professors, classmates, and friends in the Biostatistics department. In particular, I'd like to extend my deep appreciation to Kunthel By, Dave Kessler, Seunggeun Lee, Ryan May, and John Schwarz.

Finally, thanks to Natalie and Erin for being the two people I most want to see at the beginning and the end of each day.

Table of Contents

List of Tables

List of Figures

LIST OF ABBREVIATIONS

Abbreviations

- CNA - copy number aberration
- LOH - loss of heterozygosity
- SNP - single nucleotide polymorphism
- aCGH - array comparative genomic hybridization
- MLE - maximum likelihood estimate

Chapter 1

Introduction to DiNAMIC

Background

DNA copy number aberrations (CNAs) are commonly found in tumor tissue, and can range from losses (deletions) of one or both copies of chromosomal regions to gains of numerous additional copies (amplifications). The size of these aberrations can range from entire chromosome arms to less than 100 kb (Myllykangas and Knuutila, 2006). A variety of platforms are used to detect CNAs, and provide quantitative signals that reflect the underlying discrete copy number (Coe et al. 2007, Davies et al. 2005, Zhao et al. 2004). Much of the statistical effort in analyzing CNAs has focused on discerning copy number at each location within individual tumors (Olshen et al. 2004, Venkatraman and Olshen 2007, Hupe et al. 2004), and in handling the potential contamination of normal tissue in tumor samples (Sun et al. 2009).

In contrast to heritable copy number variation, CNAs are the result of genomic instability in somatic tumor tissue (Albertson et al. 2003). From the earliest days of modern cancer genetics, it was recognized that such instability could unmask or promote the effects of tumor suppressors and oncogenes (Knudsen 1971, Stratchan and Read 1999). However, surveys of a number of tumor types (e.g. Miller et al. 2003) demonstrate that sporadic gains and losses can also occur throughout the genome, likely representing generic genomic instability, with little effect on tumor progression. The phenomenon of *recurrent* CNAs, which affect the same region in multiple tumors, is of great interest, as such CNAs may highlight genes or regions that are directly involved in tumor progression. Past studies have detected recurrent CNAs in a wide range of tumor types, and extensive catalogs of these discoveries can be found in the Mitelman Database (Mitelman et al. 2010) and the Genetic Alterations in Cancer (GAC) database (Jackson et al. 2006).

Despite the apparent successes in the field, there is no clear basis for a general approach for sensitive detection of recurrent CNAs, as many regions important for tumor progression may affect only a minority of tumors. The task of distinguishing between sporadic and recurrent CNAs is thus largely a statistical issue. The instability-selection model provides a statistical framework specific to loss of heterozygosity data (Newton et al. 1998), but even for this specific data type difficulties remain in assessing significance over multiple markers (Sterrett and Wright 2007). The problem of assessing significance for general copy-number data has received relatively little attention until recently (Shah 2008). Few of the existing methods (reviewed below) provide an explicit description of the null hypothesis being tested, or fully acknowledge the inherent correlation structure of copy-number data. For these reasons, it has been difficult to place the techniques in a traditional statistical framework, or to understand error rates on a genome-wide scale. The purpose of this thesis is to introduce an explicit testing scheme for recurrent CNAs that preserves correlations inherent to the data.

Before proceeding to our testing framework, we review the current methods for copy number calling/segmentation, which can serve as a useful intermediary to the detection of recurrent CNAs. Numerous technologies are available to measure DNA copy number, ranging from array comparative genomic hybridization (Coe et al. 2007, Davies et al. 2005) at tens of thousands of probes, to high density SNP platforms (up to 1 million probes or more, Zhao et al. 2004). Reviews of the technologies are provided elsewhere (Davies et al. 2005, Zhao et al. 2004), but a common feature is that a quantitative signal is extracted at each probe that reflects underlying copy number, with additional noise and potentially probe-specific bias inherent to the platform.

Regional losses and gains within a single tumor typically cover contiguous sets of numerous probes (Myllykangas and Knuutila 2006), and so segmentation approaches (Olshen et al. 2004, Venkatraman and Olshen 2007, Hupe et al. 2004) are popular as a means to estimate the underlying copy number state at each position per tumor. Here we distinguish between *discrete* segmentation, where the copy number is constrained to the non-negative integers, and *continuous* segmentation, where the segmented values need not be integers. Examples of continuous segmentation include methods that essentially average over quantitative probe values within a genomic region determined by the algorithm to be a copy number segment. Regardless of the segmentation procedure, technical artifacts and differences in probe characteristics can lead to probe-specific bias, potentially reducing the accuracy of segmentation. A number of authors have established the presence of probe bias. Marioni et al. (2007) showed that aCGH data exhibits serial autocorrelation, a phenomenon they termed "genomic waves," and Komura et al. (2006) noted a correlation between apparent DNA copy number and GC content. For sufficient sample sizes, the approach introduced in Chapter 3 can be used to correct the bias by comparing intensities of individual probes using data from surrounding probes (via segmentation), without the need to model or otherwise consider the sequence context.

We also clarify that we are interested in somatic copy number changes in tumors, rather than heritable copy number variants (CNVs). As the resolution of typing technologies increases, it is possible that CNVs, which are rarely larger than 1 Mb (Itsara et al. 2009) and thus considerably shorter than the aberrations found in solid tumors (Albertson et al. 2003), can be mistaken for recurrent CNAs. The distinction can be clarified by comparisons of matched tumor and normal tissue. Researchers using tumor-only datasets should be alert to the possible presence of common copy number polymorphisms when interpreting the results of our method (Redon et al. 2006).

Over a dozen software packages for analyzing DNA copy number data are discussed by Rueda and Diaz-Uriate (2008), Baross et al. (2007), and Shah (2008). We focus here on the approaches that attempt to identify recurrent copy-number changes, highlighting the input formats and a few relevant similarities and differences.

- STAC (Diskin et al. 2006) and CGHregions (van de Wiel and van Wieringen 2007) require discrete segmented input data, i.e. categorical values such as aberrant/normal, gain/normal/loss, or some numerical equivalent.

- GISTIC (Beroukhim et al. 2007) requires continuous segmented input data, such as one might obtain from a segmentation program such as GLAD (Hupe et al. 2004) or DNAcopy (Olshen et al. 2004, Venkatraman and Olshen 2007).
- KC-SMART (Klijn et al. 2008) and MSA (Guttman et al. 2007) accept *continuous* input data, such as $\log_2$ intensity ratios, although MSA performs discrete segmentation internally and then makes multiple calls to the STAC algorithm.
- GISTIC, KC-SMART, STAC, and MSA assess the statistical significance of the most striking marker or region using permutation-based null distributions, while adjusting for multiple comparisons. However, the resulting output differs among the methods. GISTIC produces false discovery rate (FDR) q-values for the ’significant’ regions. STAC and MSA control the family-wise error rate (FWER) by using the max-T procedure of Westfall and Young (1993), while KC-SMART controls FWER by using a Bonferroni adjustment.
- GISTIC and KC-SMART analyze genome-wide data, whereas STAC and MSA analyze data at the level of the chromosome or chromosome arm.

Here we introduce DiNAMIC (Discovering Copy Number Aberrations Manifested In Cancer), a new procedure to map recurrent CNAs and assess their statistical significance. DiNAMIC can be applied in the analysis of data from individual chromosomes or genome-wide. The input can consist of segmented data, either discrete or continuous. Alternately, quantitative probe measurements may be used directly, although the reader is advised to read the material on probe bias in Chapter 2 before analyzing individual probe-level data. DiNAMIC is computationally fast, statistically robust, and requires no specialized software. We believe that DiNAMIC is a valuable addition to the methods available to search for recurrent CNAs.

Data Format and Definitions

We assume that the available numerical data is contained in an $n \times m$ matrix X. Each row of X corresponds to DNA copy number or LOH data obtained from one subject at m markers, while each column of X corresponds to data at a single marker for n subjects. Thus $x_{i,j}$ represents DNA copy number or LOH data for subject i at marker j. Because LOH data can be viewed as a special case of copy number data, all subsequent discussions of copy number data will also refer to LOH data.

Markers at which multiple subjects exhibit high or low copy number are of interest, because these are potentially sites of recurrent copy number aberration. Thus it is natural to examine local summary statistics for each marker. We define S_i to be the sum of the entries in the i^{th} column of X, leading to the local summary statistics $S_1, S_2, \ldots, S_m$.

In addition to the local summary statistics, we also want a global summary statistic $T(X)$ for the entire data matrix that is sensitive to the presence of copy number gains and losses. We will restrict our attention to

$$T(X) = \max(S_1, S_2, \ldots, S_m),$$

and, if appropriate,

$$T(X) = \min(S_1, S_2, \ldots, S_m),$$

where S_i represents the i^{th} column sum. These choices focus attention on the markers that are most likely to be important, and they will be used when we assess the statistical significance of CNAs.

Permutation, Cyclic Shift, and Assessing Statistical Significance

Random variation in DNA copy number will be found in both normal and tumor samples, so it important to have methods for determining whether a given dataset contains statistically significant CNAs. This can be done if we have a distribution for $T(X)$ under the null hypothesis that no CNAs are present. Because we make no assumptions about the entries in X - i.e. they may be discrete segmented, continuous segmented, or continuous values - it is not possible to find or estimate parametric distributions of $T(X)$. For this reason we consider a permutation null distribution for $T(X)$, an approach that is also taken by GISTIC, STAC, MSA, and KC-SMART.

A variety of permutation schemes are possible. Because entries in different rows come from different subjects, permuting entries across rows should be avoided. KC-SMART randomly permutes the DNA copy number values within a given row. GISTIC's null distribution is based on a convolution of histograms, which is equivalent to randomly permuting entries in a given row. On the other hand, STAC and MSA perform random rearrangements of aberrant regions within a given row. Such permutations maintain the serial structure of the aberrant regions but not the serial structure of the normal regions.

DNA copy number data is inherently correlated, and both discrete segemented and continuous segmented copy number data can be very highly correlated. Information is lost if we ignore the existence of this correlation, so within each row it is desirable to maintain as much of the serial structure as possible under permutation. This provides motivation for DiNAMIC's permutation scheme.

Let $X_{i\cdot} = x_{i,1}x_{i,2}\ldots x_{i,m}$ be the i^{th} row of X, which corresponds to the data from the i^{th} subject. If $1 \leq k \leq m$, we define a *cyclic shift* of $X_{i\cdot}$ of index k to be

$$\sigma_k(X_{i\cdot}) = x_{i,k}x_{i,k+1}\ldots x_{i,m}x_{i,1}\ldots x_{i,k-1}.$$

More generally, a *cyclic shift* $\sigma(X)$ of X is found by applying cyclic shifts σ_k to each row of X, where the shift index k can vary from one row to the next. This yields a total of m^n distinct cyclic shifts.

Biological motivation for using cyclic shifts can be found by considering DNA copy number on circular bacterial chromosomes. Here the serial structure of the copy number data from a given row is completely preserved under cyclic shifts. Thus if the observed copy numbers for each row mimic a circular stationary process, the correlation structure is not changed by performing cyclic shifts.

Although human chromosomes are linear, not circular, the cyclic shift $\sigma_k(X_{i\cdot})$ maintains the serial structure between the markers, except at the breakpoint $x_{i,k-1}$. In an $n \times m$ matrix the number of markers m is much larger than the total number of breakpoints n, so it follows that the difference between linear chromosomes and circular chromosomes is negligible. Therefore under stationarity we again conclude that when performing cyclic shifts there is no appreciable alteration in the correlation structure. We discuss this topic in greater depth in Chapter 8.

We now describe our method for assessing the statistical signficance of $T(X) = \max(S_1, S_2, \ldots, S_m)$ using cyclic shifts.

1. Perform N random cyclic shifts $\sigma^1(X), \sigma^2(X), \ldots, \sigma^N(X)$.

2. Compute $T(\sigma^i(X))$ for $i = 1, 2, \ldots, N$.

3. $p(T(X)) = \dfrac{\sum_{i=1}^{N} I(T(\sigma^i(X)) \geq T(X))}{N}$.

When $T(X) = \min(S_1, S_2, \ldots, S_m)$ we obtain $p(T(X))$ by reversing the inequality in step 3. This definition yields a p-value for $T(X)$ that is easy to interpret and adjusted for multiple comparisons. Moreover, it allows us to assess (a) 'high only' significance, as one would do for LOH data, or (b) 'high and low' significance, which would be of interest for copy number data.

In tumor samples markers can exhibit recurrent high or low copy number because of somatic mutations that provide a growth advantage. However, certain copy number variants (CNVs) are known to be common in populations, and these could also lead to markers that exhibit statistically significant high or low copy number. Although GISTIC automatically compares its discoveries with a database of known CNVs, currently DiNAMIC does not have this capability. As a result, the user is advised to compare DiNAMIC's discoveries with a CNV database such as the Database of Genomic Variants (Iafrate et al. 2004).

Chapter 2

Methods for DiNAMIC

As noted in the introduction, variations in probe hybridization affinity can lead to probe-specific bias for DNA copy number measurements. Here we present a simple bias-correction procedure that does not require information about probe length or nucleotide content. We then introduce "peeling," DiNAMIC's sequential technique to assess the significance of multiple markers, while accounting for previously discovered markers. However, we begin this chapter with a brief discussion of the null hypothesis and some working assumptions regarding stationarity.

Null Hypothesis, Working Assumptions, and Stationarity

Suppose X is a random matrix with iid rows $X_{i\cdot}$. Our null hypothesis is that the (multivariate) distributions of the $X_{i\cdot}$ are finite dimensional distributions of a stationary process. Stationarity of the mean structure of the $X_{i\cdot}$ implies that the expected values of the column sums of X are the same, and thus there are no recurrent CNAs. Chapter 8 contains a more complete discussion of the relevant distributions of $T(X)$, covariance structures, and the effect of cyclic shifts on covariance structures. For now, however, we note that the formula $\hat{p}(T(X) > t) = \dfrac{1}{N}\sum_{i=1}^{N} I(T(\sigma^i(X)) > t)$ is likely to give a good approximation to the true probability when the covariance structure of the $X_{i\cdot}$ is stationary and the number of markers is large.

In practice we have a fixed data matrix X, not a realization of a random matrix. Our null hypothesis is that no recurrent CNAs are present, and thus we expect to see only random variation in the column sums of X. We may no longer make assumptions about the covariance structure of the rows or the column sums of X. Thus we make a working assumption that the empirical correlation structure of the columns of X mimics that of a stationary process.

Probe Bias in DNA Copy Number Data

Probe-specific variations in hybridization affinity can lead to corresponding variations in array intensity. These in turn can result in biased estimates of DNA copy number, and hence markers may appear to harbor recurrent CNAs even though their underlying copy number values are normal. Thus probe bias can lead to a situation where the null hypothesis is violated, not because of the presence of underlying recurrent CNAs, but rather due to technical artifacts.

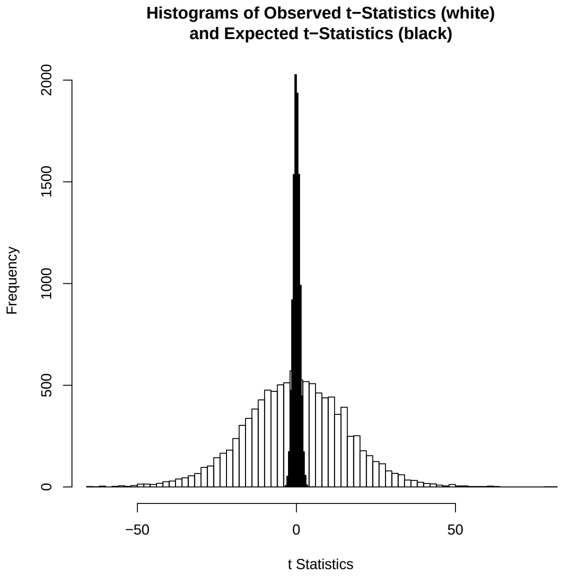

Figure 2.1: Histograms of Observed (White) and Expected (Black) t-Statistics Based on t-Tests of Columns of $\mathrm{Resid}(Z) = Z - \mathrm{Seg}(Z)$, where Z is the chromosome 2 data from Kotliarov et al. (2006)

To get some sense of the potential magnitude of the bias, suppose Z is the chromosome 2 data from the glioma dataset of Kotliarov et al. (2006), and let $\mathrm{Seg}(Z)$ be a continuous segmented version of Z obtained using DNAcopy. Segmentation algorithms use the existing data to model the true underlying copy number as a piece-wise constant function, so the column means of $\mathrm{Resid}(Z) = Z - \mathrm{Seg}(Z)$ have expected value zero, and any variation should reflect random error. However, the histograms in Figure 2.1 show that the t-statistics obtained by performing t-tests on each column of $\mathrm{Resid}(Z)$ are very different from the expected t-statistics.

Probe bias can lead to matrices with statistically significant column sums, even in the absence of recurrent CNAs. Failure to correct for this bias can result in increased type I error. Nevertheless, none of the currently available methods for analyzing DNA copy number data appear to have addressed this issue. One possible method for obtaining a bias-corrected version of the data is to perform continuous segmentation as a preprocessing step, and then analyze the segmented data. (GISTIC takes this approach, although Beroukhim et al. (2007) make no mention of probe bias.) Unfortunately, simulations show that probe bias may still be present in segmented data.

We use DiNAMIC to analyze simulated 50×2000 data matrices that include probe bias. First we create 50×2000 matrices Y that contain null copy number values. The specific scheme for creating the matices Y is discussed in detail in the Simulation Studies section in Chapter 3. Here we simply note that $Y = 2 + ((G1 - G2) * S) + N$, where $G1$ and $G2$ are simulated with the instability-selection model that is discussed in Chapter 8. Next we use the chromosome 2 data from Kotliarov et al. (2006) to create a 50×2000 matrix W that contains probe bias. In particular, W is a 50×2000 submatrix of the matrix $\mathrm{Resid}(Z)$ described earlier.

The segmented matrix Seg(X) is obtained by applying DNAcopy to $X = Y + W$. We obtain an observed type I error of .2368 when we follow the procedure for analyzing null datasets at the $\alpha = .05$ level, as outlined in the section on Simulation Studies of Chapter 3.

Because of this problem we recommend the following procedure for removing probe bias in $n \times m$ matrices X containing continuous copy number data as a pre-processing step in DiNAMIC.

1. Segment X to get Seg(X), and then compute Resid(X) = X − Seg(X).
2. Let $\mathbf{d} = (d_j)_{j=1}^m$, where $d_j = \overline{\text{Resid}(X)}_{\cdot j}$, the mean of the entries in the j^{th} column of Resid(X).
3. Let D be an $n \times m$ matrix with each row equal to $\mathbf{d}$.
4. Define $\tilde{X} = X - D$.
5. Segment $\tilde{X}$ to get Seg($\tilde{X}$).

The vector $\mathbf{d}$ is an estimate of the probe bias in X, and this estimated bias is removed when we compute $\tilde{X} = X - D$. However, it is not appropriate to use DiNAMIC to analyze $\tilde{X}$, for reasons that we now describe. First note that the column sums of $\tilde{X}$ and Seg(X) are identical, by construction. Although the rows of Seg(X) are piecewise constant, the rows of $\tilde{X}$ are not. Instead, the entries of $\tilde{X}$ contain noise, and thus extra variability, when compared to the entries of Seg(X). As a result, additional variability is also seen in the column sums of cyclic shifts of $\tilde{X}$, but not in the column sums of $\tilde{X}$. Thus values of $T(\sigma(\tilde{X}))$ tend to be more extreme than those of $T(\tilde{X})$, which leads to conservative behavior by DiNAMIC. Performing the final segmentation to obtain Seg($\tilde{X}$) appears to solve this problem.

We now discuss simulations that illustrate the effectiveness of our bias correction procedure. Let $X = Y + W$ be the simulated copy number matrices with bias that were defined earlier. Then let X_k be the matrix obtained by performing k iterations of the bias correction procedure. For example, $X_0 = X$, and $X_1 = \text{Seg}(\tilde{X})$. We may then view

$$MSE_k = \frac{1}{mn} \sum_{i=1}^{n} \sum_{j=1}^{m} (Y_{ij} - (X_k)_{ij})^2$$

as a measure of goodness of fit. When we average over 50 simulated matrices X, Figure 2.2 shows that $k = 1$ has a better fit than $k = 0$. However, repeating (1) - (5) does not yield improvements in fit.

Peeling

Natrajan et al. (2006) obtain genome-wide copy number data based on tumor samples taken from patients with Wilms' tumors. By analyzing this data in conjuction with data on tumor relapse, these authors conclude that copy number gains in chr1q are associated with increased risk of tumor relapse. In addition, they note correlations between gains in chr1q and losses in both chr16 and chr1p. Based on these results, it is likely that that the dataset contains CNAs at multiple loci.

Let X represent the matrix of $\log_2$ copy numbers from Natrajan et al. (2006). The column sums of X after segmentation and bias-correction are plotted in Figure 2.3. We performed 1000 random cyclic shifts of X, and after each cyclic shift the maximum and minimum column

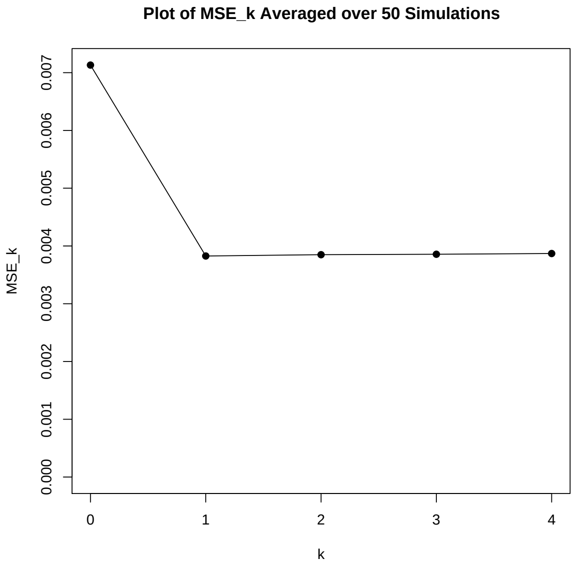

Figure 2.2: Plot of Mean Values of MSE_k to Illustrate Goodness of Fit Based on the Peeling Procedure

sum were recorded. The .975 quantile of the maxima and the .025 quantile of the minima are represented by horizontal green lines in the figure. The horizontal red line represents the mean of the column sums.

Although marker 196, the location of the maximum column sum of X, appears to be highly significant, comparison to the given quantiles shows that there are a number of other markers that also appear to be significant. Therefore it would be useful to extend DiNAMIC so that the significance of multiple markers can be assessed. The GISTIC procedure of Beroukhim et al. (2007) does this, and the significance of a 'new' region is assessed conditional on having found the previously most significant region - a process termed *peeling*. GISTIC's peeling algorithm is based on the q-values associated with genomic regions. Because DiNAMIC computes the p-value of the most aberrant marker, our peeling method, which is described below, is different from that of GISTIC.

For a given data matrix X there are three components of our peeling procedure. First, we find the most significant marker, call it k. Then we find all of the entries in X that contribute to the significance of marker k. Finally, we multiply these entries of X by a scaling factor τ to create a new data matrix $\hat{X}$ in which marker k has been nullified. At this point we can apply the ideas discussed earlier to assess the significance of the markers in $\hat{X}$ conditional having found marker k in X.

We begin by indicating how marker k is chosen. If X contains LOH data, the most significant marker k is the one corresponding to the maximum column sum. The situation is slightly more complicated if X contains copy number data. First, we use N random cyclic shifts to find the p-values of the minimum and maximum column sums. If these p-values are distinct, k is the marker corresponding to the column with the smallest p-value. When these p-values are equal, k is chosen to be marker whose column sum is farthest from the median

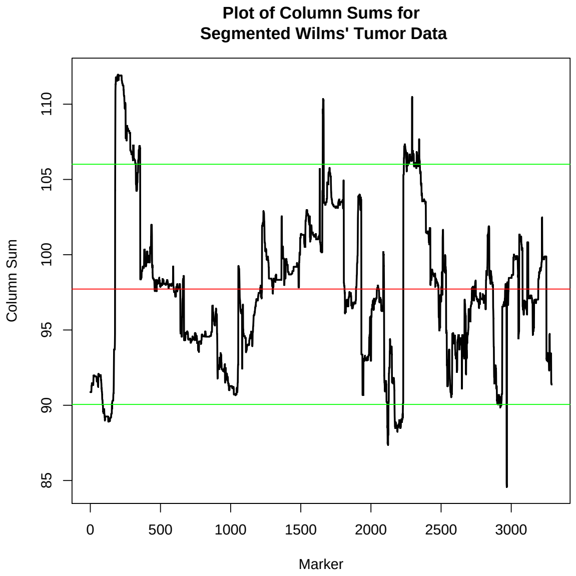

Figure 2.3: Plot of Column Sums of the Wilms' Tumor Data of Natrajan et al. (2006), together with Horizontal Lines Representing Quantiles of the Maximum and Minimum Column Sums Under Repeated Cyclic Shifts

of all column sums.

Next we describe how to find the matrix entries that contribute to the significance of marker k. For convenience, we assume that the k^{th} column sum is maximal. We write $\overline{x}_{i\cdot}$ and $\overline{x}_{\cdot j}$ for the means of the i^{th} row and j^{th} column of X, respectively, and $\overline{x}_{\cdot\cdot}$ for the grand mean of X.

1. Find the largest interval $[a, b]$ containing k such that the column means $\overline{x}_{\cdot j} > \overline{x}_{\cdot\cdot}$ for all $j \in [a, b]$.

2. If necessary, reduce $[a, b]$ so that the interval contains only markers from the same chromosome arm as marker k.

3. Let $I = \{i : x_{ik} > \overline{x}_{i\cdot}\}$ be the set of rows such that the entry x_{ik} exceeds the mean of the i^{th} row.

4. For each $i \in I$ find the maximal interval $[a_i, b_i]$ such that (i) $[a_i, b_i] \subseteq [a, b]$, (ii) $k \in [a_i, b_i]$, (iii) $x_{ij} > \overline{x}_{i\cdot}$ for all $j \in [a_i, b_i]$.

We say that $\{x_{ij} : i \in I, j \in [a_i, b_i]\}$ is the set of all matrix entries that contribute to the significance of marker k.

We now provide an example to illustrate the first two steps of the peeling procedure. The top plot in Figure 2.4 shows the column sums for a 50×100 simulated data matrix X, as well as a horizontal line representing the mean of the column sums. Based on the column sums, the most significant marker is $k = 37$. The bottom plot of Figure 2.4 shows a heat map of a 50×100 binary matrix Y. An entry y_{ij} of Y is 1 if the corresponding entry x_{ij} was found in

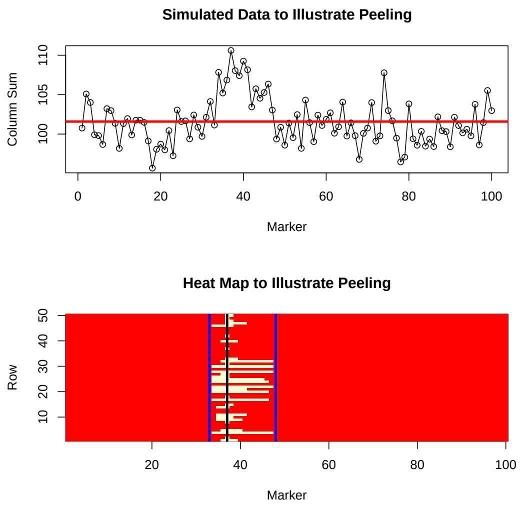

Figure 2.4: An Illustration of the Peeling Procedure in a Simulated 50 × 100 Matrix X. A Plot of the Column Sums of X (top), and a Heat Map of Binary Matrix Y Whose Entries are Defined by the Peeling Procedure (bottom)

steps (1) - (4); these are represented in the figure by white blocks. Otherwise $y_{ij} = 0$, which correspond to the red blocks in the figure. The vertical blue lines represent the interval from steps (1) and (2). We refer to this as the *peak interval*.

We conclude by showing how to compute τ, the scaling factor, and $\hat{X}$, the new data matrix.

5. Find a constant τ such that

$$n\overline{x}_{\cdot\cdot} = \sum_{i=1}^{n} x_{ik} I(x_{ik} \le \overline{x}_{i\cdot}) + \sum_{i=1}^{n} \overline{x}_{i\cdot} I(x_{ik} > \overline{x}_{i\cdot}) + \tau \sum_{i=1}^{n} (x_{ik} - \overline{x}_{i\cdot}) I(x_{ik} > \overline{x}_{i\cdot}).$$

6. Define

$$\hat{x}_{ij} = \begin{cases} \tau x_{ij} & \text{if } i \in I \text{ and } j \in [a_i, b_i]; \\ x_{ij} & \text{otherwise.} \end{cases}$$

7. Let $\hat{X}$ be an $n \times m$ matrix whose entries are $\hat{x}_{ij}$.

By construction, the mean of the k^{th} column of $\hat{X}$ is $\overline{x}_{\cdot\cdot}$. Thus marker k is null in the new dataset $\hat{X}$, as are neighboring markers. Figure 2.5 shows the column sums for the Wilms' tumor data that appeared in Figure 2.3 before peeling (black) along with the column sums after peeling (green). Before peeling column 196 yielded the most significant column sum, but after peeling the column sums around this marker are no longer significant.

Now that $\hat{X}$ has been constructed we can assess the statistical significance of the most aberrant marker in $\hat{X}$ conditional on having found marker k in the original data matrix X. This is done by using the ideas presented earlier, but now applied to $X = \hat{X}$. Because we are

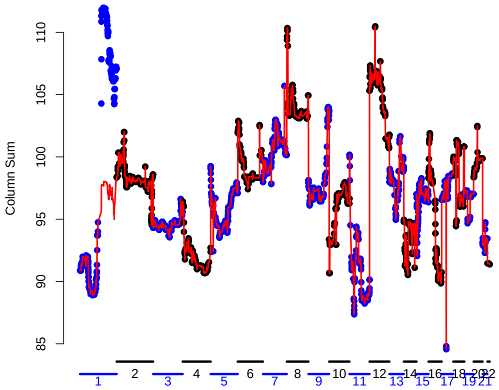

Figure 2.5: Plot of the Column Sums of the Wilms' Tumor Data of Natrajan et al. (2006) Before and After Peeling Marker 196

interested in assessing the statistical significance of $T(\hat{X})$, we must use cyclic shifts of $\hat{X}$, our current observed data. Using a null distribution based on values of $T(\sigma(\hat{X}))$ provides more power to detect aberrant markers in $\hat{X}$ than a null distribution based on values of $T(\sigma(X))$, because the extreme values that contributed to the signficance of marker k in X have been nullified. Although GISTIC performs peeling, it uses the same null distribution to assess the significance of all peeled regions, an approach that may yield conservative results.

Quick Look and Detailed Look

DiNAMIC provides two methods for assessing the statistical significance of CNAs present in a data matrix X: Quick Look and Detailed Look. Both options start by using N random cyclic shifts to simulate a null distribution for $T(X)$. In Quick Look this null distribution is used to obtain the p-value of the most significant marker k. DiNAMIC then performs the peeling algorithm to find the peak interval around k and the new data matrix $\hat{X}$. The output consists of the genomic locations of k, the endpoints of the peak interval, and $p(k)$. Then X is set equal to $\hat{X}$, and the process is repeated R times using *the original* distribution of $T(X)$. Because we use the distribution of $T(X) = \max(S_1, \ldots, S_m)$ and $T(X) = \min(S_1, \ldots, S_m)$ from the original matrix X to assess the significance of markers found in peeled versions of X, the resulting p-values are adjusted for multiple comparisons.

In contrast, Detailed Look starts by using the null distribution of $T(X)$ to find the most significant marker k and its p-value $p(k)$. It then performs the peeling algorithm to find the peak interval around k and the new data matrix $\hat{X}$. Finally, it produces the genomic locations of k, the endpoints of the peak interval, as well as $p(k)$ as output. X is set equal to $\hat{X}$, and then the process is repeated R times, including the simulation of the null distribution. Detailed Look is much more computationally intensive than Quick Look, because a new null distribution for $T(X)$ is created after each peeling. However, we believe that Detailed Look provides more accurate p-values and may have additional power to detect CNAs missed by

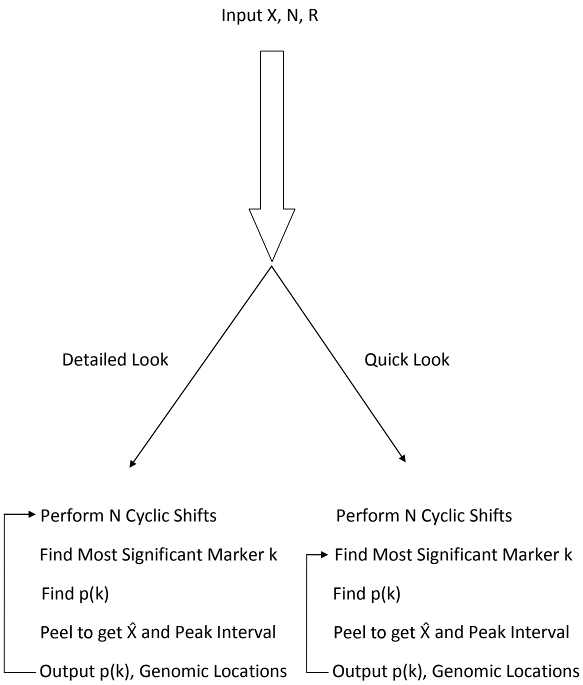
Input X, N, R
Detailed Look
Quick Look
Perform N Cyclic Shifts
Find Most Significant Marker k
Find p(k)
Peel to get X̂ and Peak Interval
Output p(k), Genomic Locations
Perform N Cyclic Shifts
Find Most Significant Marker k
Find p(k)
Peel to get X̂ and Peak Interval
Output p(k), Genomic Locations

Figure 2.6: Flow Chart Illustrating Quick Look and Detailed Look

Quick Look.

Chapter 3

Results of Data Analysis with DiNAMIC

In this chapter we describe the real and simulated datasets that were analyzed in order to assess DiNAMIC's performance and statistical properties. We use DiNAMIC to analyze a number of different datasets that were simulated under the null hypothesis that no recurrent CNAs are present, and the results of these analyses are discussed. In addition, we present the results of simulations that were performed under alternative hypotheses to measure power and test peeling accuracy. We conclude this chapter with a discussion of the loci found when DiNAMIC was used to analyze two publicly available tumor datasets.

Simulation Studies and Statistical Properties

A variety of simulated null datasets were created and subsequently analyzed with DiNAMIC in order to study its behavior under the null hypothesis that no recurrent CNAs are present. The instability-selection model of Newton *et al.* (1998) is used to simulate certain copy number matrices. Briefly, the instability-selection model produces matrices in which the entries in a given row are simulated according to a binary Markov chain. The reader is referred to Chapter 8 for a more complete description of this model. Different marker spacing and correlation schemes were considered in an effort to show that DiNAMIC is robust to the type of deviation from stationarity that can be found in real datasets. The simulated datasets include:

- 50×2000 matrices $X = 2 + ((G1 - G2) * S) + N$, where $*$ denotes element-wise multiplication. Here $G1$ and $G2$ are independently generated under the instability-selection model with equally spaced markers on the interval (0, 1), $\omega = \delta = .05$, $\lambda = 50$, and starting locations $x_1 = .15$ and $x_2 = .6$, respectively. Note that $2 + G1 - G2$ represents a matrix of idealized copy number data in which the entries in a given row correspond to a Markov chain with three states (copy number = 1, 2, or 3). The transition probabilities for the Markov chain are derived from the instability-selection model. Element-wise multiplication by the matrix S is used to simulate adjustments in observed copy number arising from normal tissue contamination of tumor samples. The degree of normal contamination should be constant for a given sample, so each row of S is constant. It follows that all columns of S are identical, and we simulate these entries by taking a random sample of size n from a $Uniform(.7, .9)$ distribution. The entries of the matrix N are iid normal with mean 0 and variance .25. The variance is chosen to be .25 so that random variation in the entries of N can occasionally result in one-unit copy number changes in X.

Null Simulation Model	Type I Error
Copy Number Data	.0449
Segmented Copy Number Data	.0408
Serially Correlated Normal	.0414
Clumped Copy Number Data (25%)	.0487
Clumped Copy Number Data (50%)	.0494
Clumped Copy Number Data (75%)	.0469
Clumped Copy Number Data (100%)	.0420

Table 3.1: DiNAMIC's Observed Type I Error for Different Copy Number Data Simulation Procedures Under the Null Hypothesis

- Continuous segmented versions of $X = 2 + ((G1 - G2) * S) + N$, where the segmentation was performed by DNAcopy.
- A variant of $X = 2 + ((G1 - G2) * S) + N$ in which the matrices $G1$ and $G2$ are generated using a common set of unequally spaced markers. A fraction of the markers, which ranges from 25% to 100%, are contained in one of eight equally spaced clumps of size .025. The remaining markers are uniformly distributed on the remaining intervals in (0, 1).
- 97×3288 matrices X, where the entries of each row $X_{i\cdot}$ are serially correlated normal random variables with mean 0 and variance 1. In order to make the data structure as realistic as possible, the correlations of adjacent entries in $X_{i\cdot}$ are fixed to equal the sample correlations between the corresponding columns of the Wilms' tumor dataset of Natrajan et al. (2006).

Table 3.1 gives the observed type I error under each of the above null simulation scenarios. In each case the observed type I error was computed as follows.

1. Create a data matrix X^l using the appropriate simulation scheme.
2. Compute $\hat{p}(T(X^l))$ using $N = 1000$ cyclic shifts of X^l.
3. Determine whether $T(X^l)$ is significant at the $\alpha = .05$ level.

Steps (1) - (3) are repeated 10,000 times, and the observed type I error is defined to be the proportion of $T(X^l)$ that are significant at the $\alpha = .05$ level. We consider the more extreme of the maximum and minimum column sums, so $T(X^l)$ is significant if $\hat{p}(T(X^l)) < .025$.

The values of the observed type I error given in Table 3.1 suggest that DiNAMIC is slightly conservative, which seems reasonable in light of the effect of the cyclic shift procedure on the underlying correlation of the markers. Markers on either side of a breakpoint will be essentially independent, and hence they are more likely to exhibit greater variability than neighboring markers in the original data. As a result, the distribution of the maximum column sum after cyclic shift should yield larger values than the corresponding distribution for the original data, and similarly for the minimum column sums. Because the values in Table 3.1 are quite close to .05, any difference in the distributions appears to be very minor.

Although we have considered realistic correlation structures so far, additional simulations demonstrate that type I error can be inflated in situations where a sufficiently large fraction of the markers are much more highly correlated than others. As an example, we simulate 50×2001 matrices X in which the entries of each row $X_{i\cdot}$ of X are serially correlated normal with mean 0 and variance 1. The correlations between the first 1000 pairs of neighboring markers in $X_{i\cdot}$ is defined to be .9999, whereas the correlations between the second 1000 pairs of markers in $X_{i\cdot}$ is set to 0. A total of 1000 data matrices are simulated, and an observed type I error of .411 is found using the procedure described above. The likely cause of the observed anticonservative behavior is that the empirical correlation structure of such datasets does not mimic that of a stationary process. Based on our investigation of publicly available datasets and our knowledge of marker spacing in arrays that yield DNA copy number data, it appears to be highly unlikely that such extreme correlation structures will be found in real datasets.

Power Simulations and Peeling Accuracy

Earlier we noted that type I error is preserved when we analyze a variety of null datasets using DiNAMIC. Now we discuss power simulations based on datasets that are simulated under the alternative hypothesis. Following the notation introduce earlier, we begin by simulating 50×2000 matrices $Y = 2 + ((G1 - G2) * S) + N$. Initially we consider the case when one of $G1$ and $G2$ is simulated under the alternative hypothesis $\omega > \delta$ and the other is simulated under the null hypothesis $\omega = \delta$. We then create $X = Y + W$, where W is the 50×2000 matrix containing probe bias that was defined in the section on Probe Bias in Chapter 2. Finally, we apply DiNAMIC to $\text{Seg}(\tilde{X})$, where $\tilde{X}$ is the bias-corrected version of X.

DiNAMIC detects the most aberrant locus in a given dataset, even if multiple CNAs are present. For this reason we believe that simulations based on a single alternative loci demonstrate DiNAMIC's power to detect CNAs. Simulating $G1$ under the alternative hypothesis corresponds to the situation where we have a single gain locus, while simulating $G2$ under the alternative means that we have a single loss locus. Power values are shown in Figure 3.1 when $.10 \leq \omega \leq .35$. By symmetry, we expect to have equal power to detect gains and losses, and this conclusion is supported by the power curves in Figure 3.1. Larger values of $\omega - \delta$ lead to greater power, as expected.

Next we present a variant of the above scenario in which both $G1$ and $G2$ are simulated under the alternative hypothesis. Simulations with one gain and one loss locus use $Y = 2 + ((G1 - G2) * S) + N$, whereas simulations with two gain loci use $Y = 2 + ((G1 + G2) * S) + N$. The preceeding discussion shows that DiNAMIC has equal power to detect gains and losses, so simulations based on two loss loci should yield similar results to those obtained from two gain loci. Thus we do not consider this scenario.

To find both alternative loci we must use DiNAMIC to find the first alternative locus, peel it, then find the second alternative locus. Because of the randomness associated with the instability-selection model, the two most significant markers need not occur exactly at the alternative loci. However, the most significant marker should be close to the true location if $\omega - \delta$ is large. We choose $\omega = .35$ and $\delta = .05$. We are interested in whether peeling the most significant marker affects the location of the next most significant marker, especially when two different loci (representing gains and losses, respectively) are in the same region. This situation is among the most challenging, as the presence of gain and loss loci in the same region can "cancel" each other and appear as normal copy number. The presence of multiple loci

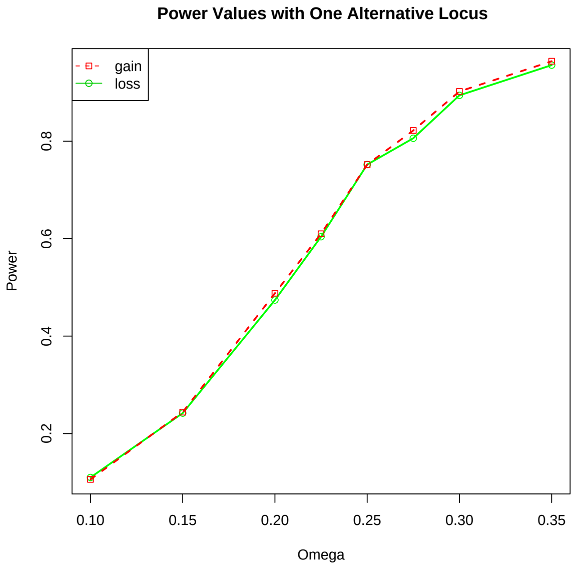

Figure 3.1: Power Curves for DiNAMIC's Cyclic Shift Procedure

of the same type (e.g. gain/gain) presents challenges as well, with an intermediate location potentially appearing as the most significant.

Let p_1 and p_2 represent the locations of the first two peeled markers, and suppose t_1 and t_2 are the locations of the true alternative loci. We measure the accuracy of the peeled markers by computing

$$SS = min\{(p_1 - t_1)^2 + (p_2 - t_2)^2, (p_1 - t_2)^2 + (p_2 - t_1)^2\}.$$

If t_1 and t_2 are sufficiently close, then peeling the most significant marker could affect the next most significant marker. This would cause the accuracy of the second peeling to decrease, and in turn lead to large values of SS. However, the effect of the first peeling on the second should decrease as the distance between t_1 and t_2 increases. Thus SS should decrease as the distance between t_1 and t_2 increases, and beyond a certain distance threshold we expect to see only random variation in SS. In addition, peeling p_1 should have more of an effect on the location of p_2 if both alternative loci are gains, but less of an effect if one alternative locus is a gain while the other is a loss.

Various distances between t_1 and t_2 are considered, where distance is measured by the number of markers between t_1 and t_2. SS is then computed for each simulated matrix Seg$(\tilde{X})$. Because SS is sensitive to differences between p_i and t_j caused by the random nature of the instability-selection model, we compute a trimmed mean of the SS values, denoted $TMSS$, where we trim 20% of both the highest and lowest observations. Table 3.2 shows the $TMSS$ values for various spacings between the alternative loci. The results are as expected. For the gain/loss scenario, once the markers are sufficiently far apart (greater than 100 markers in the simulations), the TMSS drops dramatically as the true locations can be nearly discerned. For the gain/gain scenario, the locations need to be at a greater distance (200 markers or more) for detecting both loci, as the peeling procedure for one locus can effectively nullify the

Distance	$TMSS$ for gain/loss	$TMSS$ for gain/gain
100	21144.5	98097.4
125	140.2	57980.5
150	156.9	50404.1
200	83.5	84.3
300	58.3	64.7
400	94.7	84.5

Table 3.2: Values of TMSS to Illustrate the Accuracy of DiNAMIC's Peeling Procedure When There Are Two Alternative Loci

resolution for the detecting the second locus (Table 3.2).

Gain Marker	DiNAMIC	GISTIC
1q21		X
1q23	X	
2p16	X	
6p25	X	X
6q24	X	
7q21		X
7q34	X	
8p23	X	X
8q24	X	
9q34	X	X
11p15		X
12p13	X	X
12q12	X	
13q31	X	
15q11	X	
16p13	X	
18q11	X	
18q22	X	
20p11	X	

Loss Marker	DiNAMIC	GISTIC
1p31	X	
1p21		X
3q13	X	
4p15	X	
4q31	X	
4q32		X
5p15	X	
5q11	X	
9p21		X
10p15	X	X
10q11	X	
11p13	X	X
11q22	X	
11q23		X
13q21		X
14q21	X	X
15q12	X	X
16q23		X
16q24	X	
17q12	X	X
18q11		X
19p12		X
21q21	X	X
22q12		X
22q13	X	

Table 3.3: Markers in the Glioma Dataset of Natrajan et al. (2006) Discovered by DiNAMIC's Detailed Look and GISTIC

Application to Real Datasets

As we saw above, the dataset of Natrajan et al. (2006) contains a number of copy number gain and loss loci that are potentially statistically significant. Using both DiNAMIC's Detailed Look and GISTIC, we analyze a segmented version of this dataset that was obtained by applying the bias correction scheme. Because no normal reference set is available, the thresholds for amplification and deletion, which are required input parameters for GISTIC, are set to the default values of $\pm .1$. Table 3.3 shows all markers that are peeled by DiNAMIC with a p-value less than .05, as well as all regions that are found by GISTIC to have q-values less than .05. Under the null hypothesis the FWER and the FDR are identical, so when $\alpha = .05$ it is not appropriate to use GISTIC's default q-value threshold of .25. GISTIC automatically analyzes gains and losses separately; for DiNAMIC we call a peeled marker a gain (loss) if its column sum was maximal (minimal).

Natrajan et al. (2006) note that the most common copy number gains are found in 1q, 8, and 12, with focal gains located at 1q22-25, 8p21-12, and 12p13. Both DiNAMIC and GISTIC detect markers corresponding to these gains. In addition, both methods detect markers in 9q34, the site of the *SET* oncogene. An analysis of different Wilms' tumor samples conducted by Carlson et al. (1998) noted an overabundance of SET protein. Natrajan et al. (2006) state that gains at 13q31 and 16p13 are associated with tumor relapse, and both gain loci are found by DiNAMIC but not by GISTIC. DiNAMIC's detection of 7q34 and 8q24 is noteworthy because the oncogenes *BRAF* and c-Myc lie in these regions, respectively, neither of which was detected by GISTIC.

Losses at 10p15 and 11p13 are found by Natrajan et al. (2006) in a number of subjects; these are the sites of *WT1* and *WT2*, genes known to be associated with Wilms' tumor. Both loci are detected by DiNAMIC and GISTIC. The same authors conclude that loss of 21q22 is associated with tumor relapse; both methods detect the nearby locus 21q21. Although the loss sites that the two methods detect on 1p, 11q, 16q, and 22q are not identical, the differences appear to be minor. Using linkage analysis, Rahman et al. (1996) discovered *FWT1/WT4*, a familial Wilms' tumor gene located on 17q12. This site is also detected by both methods. The gene *PDCD6* is located on 5p15, a site that is found by DiNAMIC but not GISTIC. Because *PDCD6* is know to be associated with programmed cell death, detection of this locus may have biological relevance.

GISTIC and DiNAMIC's Detailed Look are also used to analyze the glioma dataset of Kotliarov et al. (2006). As above, GISTIC's amplification and deletion thresholds are set to the default values of $\pm.1$, because no normal reference set was available. In addition, the equality of the FDR and the FWER under the null hypothesis implies that GISTIC's q-value threshold should be set equal to $\alpha = .05$. With these settings, GISTIC finds 47 significant gain regions and 20 significant loss regions. Using DiNAMIC, over 100 loci for gains and losses are found to be significant at the $\alpha = .05$ level.

Table A1 in the Appendix provides a list of all of the significant regions found by GISTIC, as well as a list of the 100 most significant loci detected by DiNAMIC. Instead of providing a complete discussion of these results, we highlight some of the similarities and differences. Of the 67 significant regions detected by GISTIC, 53 lie in cytobands that are also detected by DiNAMIC. Eight of the remaining 14 regions that GISTIC classifies as significant lie in peak intervals around sites detected by DiNAMIC, so the overall differences in these locations is not large. Kotliarov et al. (2006) provide lists of markers that exhibited homozygous deletions, heterozygous deletions, or greater than 5-fold amplifications in at least 10% of all samples in their Supplementary Tables S2 - S4. GISTIC detects no significant regions on chr1. However, on chr1 DiNAMIC finds four significant sites that lie in cytobands for markers that also appeared in Tables S2 - S4 of Kotliarov et al. (2006).

Chapter 4

Bootstrap Methods for Analyzing Copy Number Aberrations

In the section on Simulation Studies of Chapter 3 we noted that DiNAMIC can exhibit inflated type I error under extreme marker correlation structures. Although the exact reason for this phenomenon is not clear, the likely cause is the lack of stationarity in the correlation structure. This motivates our interest in developing an alternate resampling scheme that preserves the correlation structure of the markers.

The Centered Bootstrap Procedure

We briefly recall the notation from the Data Format and Definitions section of Chapter 1. The numerical data is contained in an $n \times m$ matrix X. The ith row $X_{i\cdot}$ of X corresponds to the data from the i^{th} subject at m markers, and the j^{th} column $X_{\cdot j}$ represents the copy number values for the n subjects at marker j. The column sums $S_1, \ldots, S_m$ of X form local summary statistics for the copy number data at each marker, and markers that exhibit recurrent high or low copy numbers will yield large or small column sums, respectively. The global summary statistics $T(X) = \max(S_1, \ldots, S_m)$ and $T(X) = \min(S_1, \ldots, S_m)$ focus attention on the markers that are the most likely sites of recurrent copy number aberrations.

DiNAMIC assesses the statistical significance of the observed value of $T(X)$ under the null hypothesis that no recurrent CNAs are present. An approximation to the null distribution of $T(X)$ is obtained by computing $T(\sigma(X))$ for a large number of independent cyclic shifts σ. We now introduce a bootstrap approach for creating an approximation for the null distribution of $T(X)$. As is shown in Proposition 4.1, the bootstrap approach to resampling preserves the underlying correlation structure of the data.

The following procedure is used to assess the statistical significance of the observed value of $T(X) = \max(S_1, \ldots, S_m)$ in a matrix X whose expected column sum is zero.

1. Let $\mathbf{d} = (d_j)_{j=1}^m$, where $d_j = \overline{x}_{\cdot j}$ is the mean of the entries in the j^{th} column of X,

2. Let D be an $n \times m$ matrix with each row equal to $\mathbf{d}$,

3. Define $Y = X - D$,

4. For $b = 1, \ldots, B$, form the $n \times m$ matrix $Y^{b,*}$ by taking a random bootstrap sample of the rows of Y,

5. Compute $T(Y^{1,*}), \ldots, T(Y^{B,*})$,

6. $p(T(X)) = \dfrac{1}{B}\displaystyle\sum_{b=1}^{B} I(T(X) > T(Y^{b,*}))$.

Reversing the inequality in (6) allows us to compute $p(T(X))$ for $T(X) = \min(S_1, \ldots, S_m)$.

The matrix Y is created in order to center the column sums of X at zero, and because of this we refer to the method as the *centered bootstrap procedure*. Although it may seem unusual at first, this approach is similar to what is done in other bootstrap hypothesis testing scenarios. For example, when a bootstrap t-test is performed to determine if an observed sample mean $\hat{\mu}_{\text{obs}}$ is different from zero, the significance of the observed statistic $t_{\text{obs}} = \dfrac{\hat{\mu}_{\text{obs}}}{\hat{\sigma}_{\text{obs}}}$ is assessed using the empirical distribution of statistics $t_b = \dfrac{\hat{\mu}_{\text{obs}} - \hat{\mu}_b}{\hat{\sigma}_b}$, where $\hat{\mu}_b$ and $\hat{\sigma}_b$ are obtained from the b^{th} bootstrap sample of the data. The expected value of t_b is zero, so the resulting empirical distribution of the t_b is appropriate for assessing the significance of t_{obs}.

If the rows $X_{i\cdot}$ of X are iid with multivariate distribution F, then by combining the results of Proposition 4.1 and Proposition 4.2 we see that the centered bootstrap procedure preserves the underlying correlation structure of the column sums of X. In practice the data matrix X is fixed, not random, but we expect the empirical correlations $\text{Corr}(X_{\cdot j}, X_{\cdot j'})$ and $\text{Corr}(Y^{b,*}_{\cdot j}, Y^{b,*}_{\cdot j'})$ to be similar for any columns $1 \leq j, j' \leq m$.

Proposition 4.1. *Let X be an $n \times m$ matrix whose rows $X_{i\cdot}$ are iid with mean $\mu = (\mu_1, \ldots, mu_m)$, define $\mathbf{d} = (d_1, \ldots, d_m)$ to be the vector of column means of X, and let Y be an $n \times m$ matrix with rows $Y_{i\cdot} = X_{i\cdot} - \mathbf{d}$. Define Y^* to be an $n \times m$ matrix whose rows are obtained by taking a random bootstrap sample of the rows of Y. Then $\text{Cov}\Big(\displaystyle\sum_{i=1}^{n} Y^*_{ij}, \sum_{i=1}^{n} Y^*_{ij'}\Big) = (n-1)\text{Cov}(X_{ij}, X_{ij'})$ for $1 \leq j, j' \leq m$.*

Proof. By definition,

$$\text{Cov}\Big(\sum_{i=1}^{n} Y^*_{ij}, \sum_{i=1}^{n} Y^*_{ij'}\Big) = E\Big(\sum_{i=1}^{n} Y^*_{ij} \sum_{i=1}^{n} Y^*_{ij'}\Big) - E\Big(\sum_{i=1}^{n} Y^*_{ij}\Big) E\Big(\sum_{i=1}^{n} Y^*_{ij'}\Big).$$

Now

$$E\Big(\sum_{i=1}^{n} Y^*_{ij}\Big) = E\Big(E\Big(\sum_{i=1}^{n} Y^*_{ij}|Y\Big)\Big) = nE\Big(E\Big(Y^*_{ij}|Y\Big)\Big) = nE\Big(\frac{1}{n}\sum_{i=1}^{n} Y_{ij}\Big) = E\Big(\sum_{i=1}^{n} Y_{ij}\Big) = 0,$$

and similarly $E\Big(\displaystyle\sum_{i=1}^{n} Y^*_{ij'}\Big) = 0$. Next note that $\displaystyle\sum_{i=1}^{n} Y^*_{ij} \sum_{i=1}^{n} Y^*_{ij'} = \sum_{i=1}^{n} Y^*_{ij} Y^*_{ij'} + \sum_{\substack{i=1 \\ i \neq i'}}^{n} \sum_{i'=1}^{n} Y^*_{ij} Y^*_{i'j'}$. Because Y^*_{ij} and $Y^*_{i'j'}$ are independent when $i \neq i'$, the above computations show that $E\Big(\displaystyle\sum_{\substack{i=1 \\ i \neq i'}}^{n} \sum_{i'=1}^{n} Y^*_{ij} Y^*_{i'j'}\Big) = 0$. Thus $\text{Cov}\Big(\displaystyle\sum_{i=1}^{n} Y^*_{ij}, \sum_{i=1}^{n} Y^*_{ij'}\Big)$ reduces to $E\Big(\displaystyle\sum_{i=1}^{n} Y^*_{ij} Y^*_{ij'}\Big)$. However this equals $nE\Big(E\Big(Y^*_{ij}Y^*_{ij'}|Y\Big)\Big) = nE\Big(\frac{1}{n}\sum_{i=1}^{n} Y_{ij}Y_{ij'}\Big)$, which equals $E\Big(\sum_{i=1}^{n} Y_{ij}Y_{ij'}\Big)$. By definition of the entries in Y, this in turn may be rewritten as $nE\Big(\Big(X_{ij} - d_j\Big)\Big(X_{ij'} - d_{j'}\Big)\Big) = nE\Big[\Big(X_{ij} - \frac{1}{n}\sum_{i=1}^{n} X_{ij}\Big)\Big(X_{ij'} - \frac{1}{n}\sum_{i=1}^{n} X_{ij'}\Big)\Big]$. Because of the independence of elements in distinct rows, this simplifies to $(n-1)E(X_{ij}X_{ij'}) - (n-1)\mu_j\mu_{j'} = (n-1)\mathrm{Cov}(X_{ij}, X_{ij'})$. □

In the following proposition we restrict our attention the entries in a single column of the matrices X and Y from the centered bootstrap procedure.

Proposition 4.2. *Let* $X = \{X_1, \ldots, X_n\}$ *be iid with variance* σ^2*, and define* $Y = \{X_1 - \overline{X}, \ldots, X_n - \overline{X}\}$ *to be a centered version of* X*. If* $Y^* = \{Y^*_1, \ldots, Y^*_n\}$ *is a random bootstrap sample of* Y*, then* $\mathrm{Var}\Big(\sum_{i=1}^{n} Y^*_i\Big) = (n-1)\sigma^2$.

Proof. Write $\mathrm{Var}\Big(\sum_{i=1}^{n} Y^*_i\Big) = E\Big(\mathrm{Var}\Big(\sum_{i=1}^{n} Y^*_i|Y\Big)\Big) + \mathrm{Var}\Big(E\Big(\sum_{i=1}^{n} Y^*_i|Y\Big)\Big)$. Since $E(Y^*_i|Y) = E\Big(\frac{1}{n}\sum_{j=1}^{n} Y_j\Big) = 0$ for all i, it follows that $\mathrm{Var}\Big(E\Big(\sum_{i=1}^{n} Y^*_i|Y\Big)\Big) = 0$. Therefore the independence of the Y^*_i conditional on Y allows us to reduce $\mathrm{Var}\Big(\sum_{i=1}^{n} Y^*_i\Big)$ to $\sum_{i=1}^{n} E\Big(\mathrm{Var}\Big(Y^*_i|Y\Big)\Big)$. We view Y as a population when taking bootstrap samples, so $\mathrm{Var}\Big(Y^*_i|Y\Big) = s^2_Y = s^2_X$, the sample variance of X. Since $E(s^2_X) = \frac{n-1}{n}\sigma^2$, the result follows. □

Simulation Studies

Here the centered bootstrap procedure is used to analyze copy number matrices simulated under both the null and alternative hypotheses. The 50×2000 simulated matrices have the form $X = ((G1 - G2) * S) + N$, where the notation is the same as in the section on Simulation Studies in Chapter 3. We consider the null hypothesis $\omega = \delta$, and also the case when one of $G1$ and $G2$ is simulated under the alternative hypothesis $\omega > \delta$ and the other is simulated under the null hypothesis $\omega = \delta$.

Simulating $G1$ under the alternative corresponds to the situation where we have a single gain locus, while simulating $G2$ under the alternative means we have a single loss locus. Power values are shown in Figure 4.1 when $\delta = .05$ and $.05 \leq \omega \leq .3$ for both the centered bootstrap procedure (Boot in the figure) and cyclic shift (CS in the figure). For each value of ω we compute $p(T(X))$ using the cyclic shift procedure with $N = 1000$ random cyclic shifts and the centered bootstrap procedure with $B = 1000$ random bootstrap samples. The process is repeated for 1000 matrices X, and the proportion of p-values significant at the $\alpha = .05$ level is recorded. Like cyclic shift, the centered bootstrap procedure has equal power to detect gains and losses of the same magnitude. Moreover, larger values of $\omega - \delta$ lead to greater power. However, the centered bootstrap procedure is noticably less powerful than cyclic shift.

This decrease in power is also apparent when analyzing real datasets. For example, when we use the centered bootstrap procedure to analyze the Wilms' tumor dataset of Natrajan et al. (2006), a total of 26 loci are significant at the $\alpha = .05$ level based on $B = 1000$ random

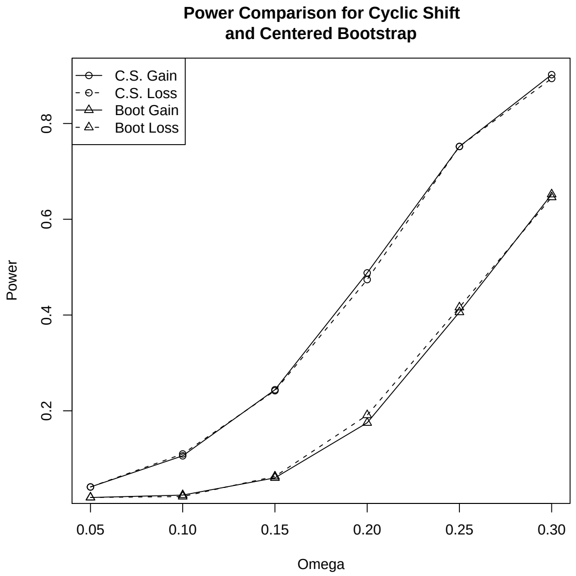

Figure 4.1: Power Curves for the Centered Bootstrap and Cyclic Shift Procedures

bootstrap samples. These loci are listed in Table 4.1, which appears at the end of this chapter. On the other hand, the cyclic shift procedure identifies 32 significant loci at the $\alpha = .05$ level using $N = 1000$ random cyclic shifts.

The Quantile-Adjusted Bootstrap Procedure

The conservative behavior exhibited by the cented bootstrap procedure implies that the values of $T(Y^{b,*})$ tend to be more extreme than those of $T(X)$. By assumption the expected column sums of X are zero, and the same is true for the expected column sums of $Y^{b,*}$. Proposition 4.2 shows that we cannot attribute the conservative behavior to increased variance. However, our next result offers a possible explanation, because it shows that in certain situtations the column sums of $Y^{b,*}$ can have larger kurtosis than the column sums of the original data matrix.

Proposition 4.3. *Let X, Y, and Y^* be as in Proposition 4.2. In addition, suppose the $X_i \in X$ have kurtosis κ. If $W = \sum_{i=1}^{n} Y_i^*$, then the kurtosis of W is $\kappa\Big[\frac{n^2-3n+3}{n^2(n-1)} + \frac{3(n-1)}{n^2\sigma^4}\Big] + \frac{6}{n}$.*

Proof. By definition, the kurtosis of W is $\dfrac{E((W-\mu_W)^4)}{E((W-\mu_W)^2)^2} - 3$. However, $\mu_W = E(W) = E(E(W|Y))$, and $E(W|Y) = \sum_{i=1}^{n} E(Y_i^*|Y) = n\overline{Y} = 0$ by construction. Thus the kurtosis of W reduces to $\dfrac{E(W^4)}{E(W^2)^2} - 3$. For convenience, we examine the numerator and denominator separately, starting with the denominator.

If we expand, we see that $W^2 = \Big(\sum_{i=1}^{n} Y_i^*\Big)^2 = \sum_{i=1}^{n}(Y_i^*)^2 + \sum_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n} Y_i^* Y_j$. Thus $E(W^2|Y) =$

$\sum_{i=1}^{n} E((Y_i^*)^2|Y) + \sum_{i=1}^{n}\sum_{\substack{j=1\\j\neq i}}^{n} E(Y_i^* Y_j^*|Y)$. However, Y_i^* and Y_j^* are independent conditional on Y, so $E(Y_i^* Y_j^*|Y) = E(Y_i^*|Y)E(Y_j^*|Y) = 0$, as noted above. Thus $E(W^2|Y) = \sum_{i=1}^{n} E((Y_i^*)^2|Y) = nm_2$, where $m_2 = \frac{1}{n}\sum_{i=1}^{n}\Big(Y_i - \overline{Y}\Big)^2$. It follows that $E(W^2) = nE(m_2) = (n-1)\sigma^2$.

Now consider W^4. If we expand we get sums of terms like $(Y_i^*)^4, (Y_i^*)^3 Y_j^*$, $(Y_i^*)^2(Y_j^*)^2, (Y_i^*)^2 Y_j^*, Y_k^*$, and $Y_i^* Y_j^* Y_k^* Y_l^*$. Because the Y_i^* and Y_j^* are independent conditional on Y, all of $E((Y_i^*)^3 Y_j^*|Y), E((Y_i^*)^2 Y_j^* Y_k^*|Y)$, and $E(Y_i^* Y_j^* Y_k^* Y_l^*|Y)$ are zero. Thus the expectation of W^4 reduces to the sum of expectations of terms of the form $(Y_i^*)^4$ and $(Y_i^*)^2(Y_j^*)^2$. As above, $E((Y_i^*)^4|Y) = m_4$, where $m_4 = \frac{1}{n}\sum_{i=1}^{n}\Big(Y_i - \overline{Y}\Big)^4$. In addition, $E((Y_i^*)^2(Y_j^*)^2|Y) = E((Y_i^*)^2|Y)E((Y_j^*)^2|Y) = m_2^2$. Now we need to compute the coefficients of $(Y_i^*)^4$ and $(Y_i^*)^2(Y_j^*)^2$ in the expansion of W^4 and count the number of times the various terms arise.

Since $(Y_1^* + \cdots + Y_n^*)^4 = \sum_{k_1,\ldots,k_n} \binom{4}{k_1,\ldots,k_n}(Y_1^*)^{k_1}\cdots(Y_n^*)^{k_n}$, the coefficients of $(Y_i^*)^4$ and $(Y_i^*)^2(Y_j^*)^2$ can be found by computing the appropriate multinomial coefficient. For example, the multinomial coefficient of $(Y_1^*)^4$ is $\binom{4}{4,0,\ldots,0} = 1$, and similarly for any other $(Y_i^*)^4$. On the other hand, the coefficient of $(Y_1^*)^2(Y_2^*)^2$ is $\binom{4}{2,2,0,\ldots,0} = 6$, and similarly for the other $(Y_i^*)^2(Y_j^*)^2$.

Next we count the number of times the terms arise. The number of terms of the form $(Y_i^*)^4$ can be found by computing the number of ways to partition 4 into non-negative summands $k_1,\ldots,k_n$, one of which is 4. There are $\binom{n}{1} = n$ ways to choose the non-zero term 4, and the remaining $n-1$ terms are automatically zero. Similarly, there are $\binom{n}{2} = \frac{n(n-1)}{2}$ ways to partition 4 into non-negative summands $k_1,\ldots,k_n$, two of which equal 2.

Combining the above results we see that the kurtosis of W is $\kappa(W) = \frac{E(W^4)}{E(W^2)^2} - 3 = \frac{nE(m_4) + 3(n)(n-1)E(m_2^2)}{(n-1)^2\sigma^4} - 3$. Joanes and Gill (1998) note that $E(m_4) = \frac{(n-1)(n^2-3n+3)}{n^3}\mu_4 + \frac{3(n-1)(2n-3)}{n^3}\sigma^4 = q_1(n)\mu_4 + q_2(n)\sigma^4$, where μ_4 is the fourth central moment. Next we write $E(m_2^2) = Var(m_2) + E(m_2)^2$. Since $m_2 = \frac{n-1}{n}s^2$, it follows that $Var(m_2) = \Big(\frac{n-1}{n}\Big)^2\Big(\frac{2\sigma^4}{n-1} + \frac{\kappa}{n}\Big)$. If we write $q_3 = \frac{(n-1)^2}{n^2}$, then $\kappa(W)$ may be written as $\frac{n[q_1(n)\mu_4 + q_2(n)\sigma^4] + 3(n)(n-1)[Var(m_2) + q_3(n)\sigma^4]}{(n-1)^2\sigma^4} - 3$. Simplifying this expression gives the desired form. □

Assume the entries in X are normally distributed with mean 0 and variance σ^2. The column sums of X are normally distributed with mean 0, variance $n\sigma^2$, and kurtosis equal to zero. However, Propositions 4.2 and 4.3 show that the column sums of $Y^{b,*}$ have mean 0, variance $(n-1)\sigma^2$, and kurtosis equal to $\frac{6}{n}$. The positive kurtosis implies that the distribution of the column sums of $Y^{b,*}$ has more weight in the tails than the distribution of the column sums of X. Therefore the values of $T(Y^{b,*})$ are more extreme than those of $T(X)$, which may explain the conservative behavior of the centered bootstrap procedure.

We now present a modified version of the centered bootstrap procedure that attempts to adjust for the increased kurtosis found in the column sums of the bootstrap matrices. The *quantile-adjusted bootstrap procedure* is based on a t-distribution with $n+4$ degrees of freedom, which is known to have kurtosis equal to $\frac{6}{n}$. In practice, other symmetric distributions with mean 0 and the correct kurtosis could be chosen. Although we have not investigated this problem to date, study of the higher moments of the column sums before and after bootstrap resampling may yield additional insight into whether quantile adjustment based on a t-distribution with $n+4$ degrees of freedom is optimal.

The following procedure assesses the statistical significance of the observed value of $T(X) = \max(S_1, \ldots, S_m)$ in a matrix X whose expected column sum is zero.

1. Let $\mathbf{c} = (c_j)_{j=1}^m$, where $c_j = \overline{x}_{\cdot j}$ is the mean of the entries in the j^{th} column of X,

2. Let C be an $n \times m$ matrix with each row equal to $\mathbf{c}$,

3. Define $Y = X - C$,

4. For $b = 1, \ldots, B$, form the $n \times m$ matrix $Y^{b,*}$ by taking a random bootstrap sample of the rows of Y,

5. For each $b \in 1, \ldots, B$ do the following:

   (a) Compute $S_1^{b,*}, \ldots, S_m^{b,*}$, the column sums of $Y^{b,*}$.

   (b) Let σ^b be the standard deviation of $S_1^{b,*}, \ldots, S_m^{b,*}$.

   (c) Let $\mathbf{t}^b = (t_1^b, \ldots, t_m^b)$, where $t_i^b = \sqrt{\frac{n+4}{(n+4)-2}}\left(\frac{1}{\sigma^b}\right)S_i^{b,*}$. By construction the entries in $\mathbf{t}^b$ have variance $\frac{n+4}{n+4-2}$, which is appropriate for t random variable with $n+4$ degrees of fredom.

   (d) Let $\mathbf{z}^b = (z_1^b, \ldots, z_m^b)$, where $z_i = \Phi^{-1}(F_{n+4}(t_i))$, where F_{n+4} is the cumulative distribution function for a t distribution with $n+4$ degrees of freedom, and Φ is the cumulative distribution function for a standard normal distribution. This quantile adjustment is used to account for the increased kurtosis caused by bootstrap resampling.

   (e) Let $\tilde{\mathbf{z}}^b = (\tilde{z}_1^b, \ldots, \tilde{z}_m^b)$, where $\tilde{z}_i = \sqrt{\frac{n}{n-1}}\left(\sigma_b\right)z_i$. The entries of $\tilde{\mathbf{z}}^b$ should be normally distributed with approximately the same variance as the column sums of X.

6. $p(T(X)) = \frac{1}{B}\sum_{b=1}^{B} I(T(X) > \max(\tilde{\mathbf{z}}^b))$.

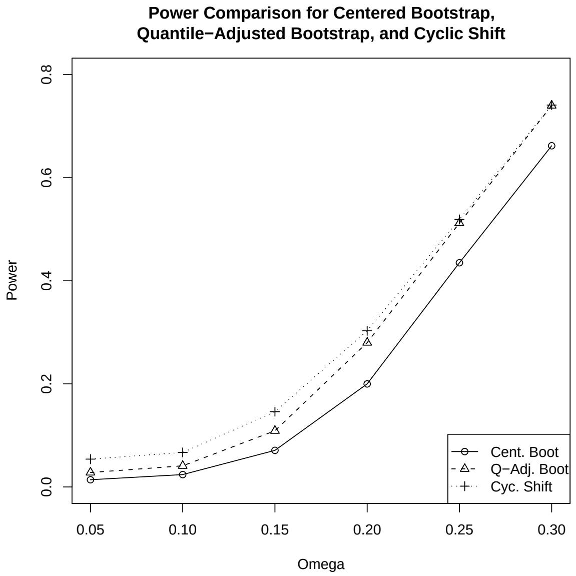

Figure 4.2: Power Curves for the Centered Bootstrap, Quantile-Adjusted Bootstrap, and Cyclic Shift Procedures

The simulations described in the discussion of the centered bootstrap procedure were repeated using the quantile-adjusted bootstrap procedure. The centered bootstrap procedure exhibited equal power to detect gains and losses under the alternative hypothesis, and by construction the quantile-adjusted bootstrap should behave similarly. Thus we restrict our attention to gains when simulating the alternative hypothesis. As we see from Figure 4.2, the quantile-adjusted bootstrap procedure is conservative under the null hypothesis, but not as conservative as the centered bootstrap procedure. Moreover, even though the quantile-adjusted bootstrap procedure is less powerful than the cyclic shift procedure, the difference in power decreases as omega increases. This suggests that if a CNA has a sufficiently large effect size, it will be classified as statistically significant by both the quantile-adjusted bootstrap and cyclic shift procedures.

We now apply the centered bootstrap and quantile-adjusted bootstrap procedures to the Wilms' tumor dataset of Natrajan et al. (2006). For the sake of comparison, we also include the results produced by the cyclic shift procedure. As we see from Table 4.1, a total of 27 loci were detected by the quantile-adjusted bootstrap procedure (Q.B.), which is one more than the number found by the centered bootstrap procedure (C.B.), but five less than the number detected by cyclic shift (C.S.). Thus when analyzing real datasets we obtain results similar to those suggested by the power simulations.

Gain	C.S.	C.B.	Q.B.	Loss	C.S.	C.B.	Q.B.
1q23	X	X	X	1p31	X	X	X
2p16	X	X	X	3q13	X		
6p25	X	X	X	4p15	X	X	X
6q24	X	X	X	4q31	X	X	X
7q34	X	X	X	5p15	X	X	X
8p23	X	X	X	5q11	X		
8q24	X	X	X	10p15	X	X	X
9q34	X	X	X	10q11	X		
12p13	X	X	X	11p13	X	X	X
12q12	X	X	X	11q22	X	X	X
13q31	X	X	X	14q21	X	X	X
15q11	X			15q12	X	X	X
16p13	X	X	X	16q24	X	X	X
18q11	X		X	17q12	X	X	X
18q22	X			21q21	X	X	X
20p11	X	X	X	22q13	X	X	X

Table 4.1: Locations of Significant Markers in the Glioma Dataset of Natrajan et al. (2006), as Determined by the Cyclic Shift (C.S.), Centered Bootstrap (C.B.), and Quantile-Adjusted Bootstrap (Q.B.) Procedures

Chapter 5

Confidence Intervals for Aberrant Markers

By construction, the cyclic shift, centered bootstrap, and quantile-adjusted bootstrap procedures identify and assess the statistical significance of recurrent CNAs. Although we expect the most aberrant marker k in a given dataset to lie near a relevant gene, it is possible that marker k does not lie inside a gene, or lies inside a gene that does not contribute to the tumor phenotype. Thus we would like to have methods for identifying genomic regions around aberrant markers that potentially harbor relevant genes. The peeling procedure identifies a peak region around each aberrant marker, but there is nothing statistically meaningful about the peak region. This motivates our interest in developing methods for finding confidence intervals or sets around aberrant markers.

Conceptual motivation for the construction of confindence intervals in the context of genetic aberrations can be found in the instability-selection model of Newton et al. (1998). These authors assume the existence of a true tumor suppressor gene locus x_s. Under the instability-selection model, a cell is more likely to be found in tumor tissue if at least one of its ancestors exhibited LOH at x_s than if none of its ancestors exhibited LOH at x_s. We take an analogous approach for copy number aberrations. If k_{true} is the locus of a gene that contributes to the tumor phenotype, then a cell is more likely to be found in tumor tissue if at least one of its ancestors had a copy number aberration at k_{true} than if none of its ancestors had a copy number aberration at k_{true}. We assume that the initiating copy number aberration at k_{true} - gain or loss - corresponds to the function of the gene containing the marker k_{true} - oncogene or tumor suppressor, respectively.

It is unlikely that all tumor samples in a given dataset contain cells that are ancestors of cells that had a CNA at the same locus k_{true}. However, if a marker k is highly aberrant, a subset of the samples may contain cells that are ancestors of cells that had a common underlying CNA. Thus it is reasonable to view a confidence interval around an aberrant marker k in the traditional setting as the estimate of an unknown parameter, which in this case is the location of the common underlying CNA.

A number of researchers have investigated methods for computing confidence intervals for quantitative trait loci (QTLs). Lander and Botstein (1989) proposed confidence intervals based on a one unit change from the maximum LOD score, and this approach was refined by Mangin et al. (1994) for QTLs with a small effect size. The bootstrap approach of Visscher et al. (1996) for constructing confidence intervals around QTLs provides the basis for our first approach.

A Bootstrap Method for Computing Confidence Intervals

As noted above, Visscher et al. (1996) present a bootstrap method for computing confidence intervals for QTLs. Briefly, for a given $n \times m$ data matrix X of allelotypes and vector $\mathbf{y} = (y_1, \ldots, y_n)^T$ of phenotypes, the authors create matrices $X^{b,*}$ and vectors $\mathbf{y}^{b,*}$ by taking the same bootstrap sample of the rows of X and the entries of $\mathbf{y}$. For $b = 1, \ldots, B$, the LOD scores for each dataset $X^{b,*}$ and vector $\mathbf{y}^{b,*}$ are computed, and the location z^b of the maximum LOD score is recorded. If c and d are the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles of $\{z^b\}_{b=1}^B$, then $[c, d]$ is a confidence interval at level $1 - \alpha$ for the location of the true quantitative trait locus for the original data X and $\mathbf{y}$.

Although we are interested in computing confidence intervals for underlying CNAs, not QTLs, the basic idea behind the method of Visscher et al. (1996) translates readily to our situation. Let X be an $n \times m$ matrix of copy number data, and assume the most aberrant marker k corresponds to the maximum column sum. The bootstrap method for computing a confidence interval $[c, d]$ at level $1 - \alpha$ for k_{true}, the true underlying copy number aberration, proceeds as follows:

1. For $b = 1, \ldots, B$ form matrices $X^{b,*}$ by taking bootstrap samples of the rows of X.

2. For each matrix $X^{b,*}$, find k^b, the location of the maximum column sum.

3. Let c and d be the $\frac{\alpha}{2}$ and $1 - \frac{\alpha}{2}$ quantiles of $\{k_b\}_{b=1}^B$.

In (2) we choose k^b to be the location of the minimum column sum if the most aberrant marker k in X corresponds to the minimum column sum.

Simulation Results for the Bootstrap-Based Method

We consider 50×2000 simulated copy number matrices $X = \text{Seg}(Y)$, where $Y = ((G1 - G2) * S) + N$, as above. Let ω_1 and ω_2 be the rates of LOH at k_{true1} and k_{true2} in $G1$ and $G2$, respectively. Thus k_{true1} and k_{true2} represent the locations of the true underlying copy number aberrations in X if $\omega_1,\ \omega_2 > \delta$. Because of symmetry, we restrict our attention to the case where $G1$ is simulated under either the null hypothesis $\omega_1 = \delta$ or the alternative hypothesis $\omega_1 > \delta$, whereas $G2$ is always simulated under the null hypothesis $\omega_2 = \delta$. Thus $k_{\text{true1}} = .15$ is the only underlying copy number aberration.

In our simulations we compute the size of the confidence interval produced for each simulated matrix X, as well as the level of coverage, i.e. the proportion of confidence intervals that contain the alternative locus $k_{\text{true1}} = .15$. In particular, we proceed as follows:

1. Use DNAcopy to create $X = \text{Seg}(Y)$, where $Y = ((G1 - G2) * S) + N$ is a matrix of simulated copy number data.

2. Compute $p(T(X))$ using the cyclic shift procedure.

3. If $p(T(X)) < .025$, find the most aberrant marker k in X.

4. Use the bootstrap-based method to find the confidence interval at level $1 - \alpha$ containing k.

5. Record the number of markers in each confidence interval, and determine if it contains $k_{\text{true1}} = .15$.

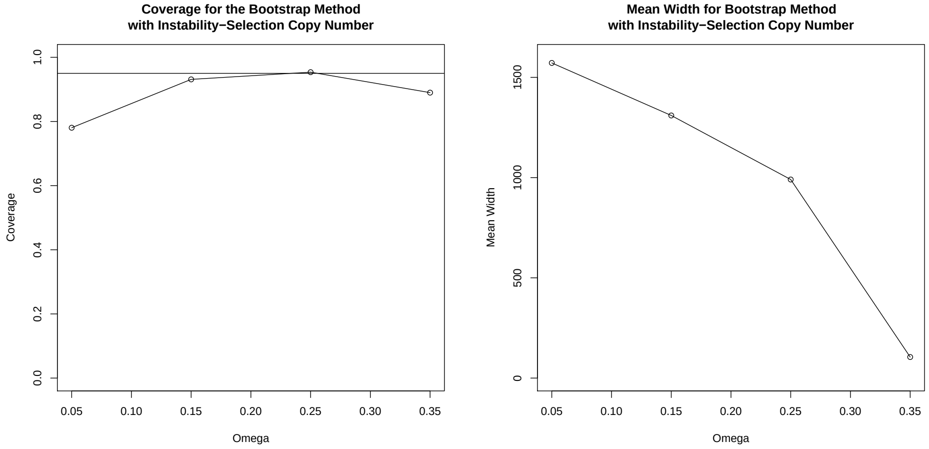

Figure 5.1: Plots Illustrating the Coverage and Median Width of Confidence Intervals Produced by the Bootstrap Procedure for Segmented Copy Number Data Simulated with the Instability-Selection Model

The plots in Figure 5.1 show the coverage and median confidence interval width for the bootstrap method when $\alpha = .05$ based on 1000 simulated segmented matrices X for each effect size $\omega_1 = .05,\ .15,\ .25$, and $.35$. Although the coverage value are near the horizontal line at $y = .95$ when $\omega_1 \geq .15$, the confidence intervals are very large for all but the largest effect size. Because of this, we wish to explore other methods for computing confidence intervals.

Test-Based Methods for Computing Confidence Intervals

In the instability-selection model the maximum likelihood estimate (MLE) $\hat{x}_s$ of x_s is the location that exhibits the greatest frequency of LOH. Moreover, $\hat{\omega}$, the observed frequency of LOH at $\hat{x}_s$, is the MLE of ω, the true rate of LOH at x_s. Since $\hat{\omega}$ is a consistent estimator of ω, the difference $|\hat{\omega} - \omega|$ is small with high probability if the number of samples n is large.

These ideas provide the basis of our test-based approach for constructing confidence intervals. We expect the observed mean copy number at the most aberrant marker k to be close to the mean copy number at the true underlying locus k_{true}, so the difference between the mean copy number values at the two locations should be close to zero. Suppose the most aberrant marker k corresponds to the maximum column sum, let $\overline{x}_{\cdot k}$ be the mean copy number in column k, and write $X_{\cdot j}$ for the j^{th} column of the $n \times m$ matrix X. Briefly, our "narrow t" method for producing a confidence interval containing marker k at level $1 - \alpha$ uses t-tests to compare the mean of $\overline{x}_{\cdot k}\mathbf{1} - X_{\cdot j}$ to zero, where $\mathbf{1}$ is a vector of 1's. Starting at $j = k$ and then moving in either direction, markers j are added to the confidence interval if the one-sided t-test is not significant at level α. We use one-sided t-tests because $\overline{x}_{\cdot k}$ is the largest column mean. Clearly the confidence interval contains k, and the process continues in either direction until the first time a significant p-value is encountered. More specifically, the narrow t method uses the following steps to find the markers contained in the confidence interval:

1. If $k = m$, then proceed to step (4); otherwise, let $j_{\text{right}} = k + 1$.

2. Define $p_{j_{\text{right}}}$ to be the p-value associated with a one-sided t-test to determine if the

mean of $\overline{x}_{\cdot k}\mathbf{1} - X_{\cdot j_{\text{right}}}$ is not equal to zero, where $\mathbf{1}$ is a vector with all entries equal to 1.

3. If $p_{j_{\text{right}}} \geq \alpha$ and $j_{\text{right}} < m$, define $j_{\text{right}} = j_{\text{right}} + 1$ and return to step (2); if $p_{j_{\text{right}}} < \alpha$, then define $j_{\text{right}} = j_{\text{right}} - 1$ and proceed to step (4); if $j_{\text{right}} = m$ and $p_{j_{\text{right}}} \geq \alpha$, proceed to step (4).

4. If $k = 1$, then proceed to step (7); otherwise, let $j_{\text{left}} = k - 1$.

5. Define $p_{j_{\text{left}}}$ to be the p-value associated with a one-sided t-test to determine if the mean of $\overline{x}_{\cdot k}\mathbf{1} - X_{\cdot j_{\text{left}}}$ is not equal to zero.

6. If $p_{j_{\text{left}}} \geq \alpha$ and $j_{\text{left}} > 1$, define $j_{\text{left}} = j_{\text{left}} - 1$ and return to step (5); if $p_{j_{\text{left}}} < \alpha$, then define $j_{\text{left}} = j_{\text{left}} + 1$ and proceed to step (7); if $j_{\text{left}} = 1$ and $p_{j_{\text{left}}} \geq \alpha$, proceed to step (7).

7. The $1 - \alpha$ level confidence interval around marker k is $[j_{\text{left}}, j_{\text{right}}]$.

We now describe two variations of the above method. The first is to create a confidence set J consisting of all markers j in the same chromosome arm as k such that $p_j \geq \alpha$, where p_j is the p-value associated with a one-sided t-test to determine if the mean of $\overline{x}_{\cdot k}\mathbf{1} - X_{\cdot j}$ is different from zero. The construction of J is called the "set t" approach. If both the narrow t and the set t procedures are applied at the same marker k, it is clear that the confidence interval $[j_{\text{left}}, j_{\text{right}}]$ produced by narrow t will be contained in the corresponding confidence set J produced by set t. Although $J = [j_{\text{left}}, j_{\text{right}}]$ in some situations, J could consist of multiple disjoint intervals, one of which is $[j_{\text{left}}, j_{\text{right}}]$. Our second variation is the "wide t" method, which is simply to take the left-most and right-most markers in the interval J found by the set t method. Wide t produces a contiguous interval $[a, b]$ that contains all intervals in J, including $[j_{\text{left}}, j_{\text{right}}]$. Even though narrow t and wide t produce contiguous intervals, we use the term *confidence sets* to refer to the confidence intervals or confidence sets produced by the three procedures.

Simulation Results for the Test-Based Methods

Here we consider two methods for simulating 50×2000 segmented copy number matrices X. The first creates the matrices $Y = ((G1 - G2) * S) + N$ using the instability-selection model, and then uses DNAcopy to produce $X = \text{Seg}(Y)$. The second method produces $X = \text{Seg}(Y)$, where $Y = Y1 - Y2$, and the rows of each of $Y1$ and $Y2$ are generated independently using a multivariate normal model with a common autoregressive correlation structure. Specifically, the entries of the k_{true1} column of $Y1$ are $a_1 + \epsilon_{i1}$, where $a_1 \geq 0$ and the ϵ_{i1} are iid $N(0, .25)$; similarly, the entries of the k_{true2} column of $Y2$ are $a_2 + \epsilon_{i2}$, where $a_2 = 0$ and the ϵ_{i2} are iid $N(0, .25)$. Here we choose $k_{\text{true1}} = 300$ and $k_{\text{true2}} = 1200$, which corresponds to true locations .15 and .6 when we have 2000 equally spaced markers on a chromosome of length 1. Once the k_{true1} entry of a row of $Y1$ is simulated, the remaining entries in the row are simulated using an AR(1) model with correlation $\rho = .9$, and similarly for $Y2$. Setting a_1 equal to 0 corresponds to the null hypothesis that there are no recurrent copy number aberrations. On the other hand, $a_1 > 0$ represents the alternative hypothesis with a single gain locus.

Our simulation scheme for computing confidence intervals is identical to the procedure described above for the bootstrap method, only now in step (4) we use narrow t, set t, and

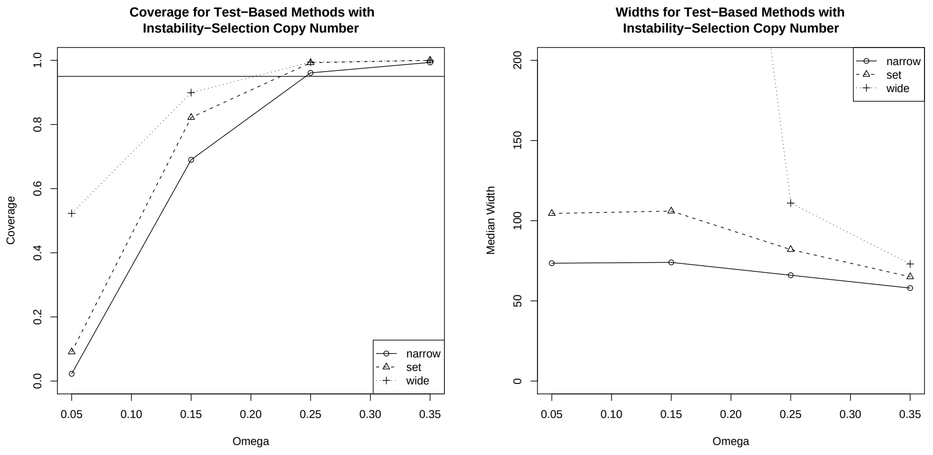

Figure 5.2: Plots Illustrating the Coverage and Median Width of Confidence Intervals Produced by the Test-Based Procedures for Segmented Copy Number Data Simulated with the Instability-Selection Model

wide t. Figures 5.2 and 5.3 show the coverage and median size of the confidence sets produced by the three procedures when $\alpha = .05$ based on 1000 simulated matrices $X = \text{Seg}(Y)$. Both methods of simulating copy number matrices Y are considered. The horizontal lines in the left-hand plots of Figures 5.2 and 5.3 are located at $1 - \alpha$, and clearly the coverage for all three methods exceeds $1 - \alpha$ when the effect size ω_1 or a_1 is sufficiently large. In spite of this, as we see from the right-hand plots of Figures 5.2 and 5.3, the sizes of the confidence intervals produced by narrow t are consistently small over a range of effect sizes. This stands in marked contrast to the simulation results from the bootstrap-based method.

The fact that the narrow t method provides overcoverage merits some discussion. Assume the use of t-tests is appropriate in a matrix X, and suppose the true mean copy number at k_{true}, call it μ, is known. Also assume that μ is larger than μ_j, the mean copy number at any marker $j \neq k_{\text{true}}$. If we perform t-tests of $H_0 : \mu - \overline{x}_{\cdot j} = 0$ vs. $H_a : \mu - \overline{x}_{\cdot j} > 0$ at level α, where $\overline{x}_{\cdot j}$ is the mean of the entries in the j^{th} column of X, then by inverting the test we should obtain confidence sets that provide coverage at level $1 - \alpha$. Moreover, these confidence sets are identical to the ones produced by the set t procedure. Because the confidence interval produced by narrow t at a given marker is contained in the corresponding confidence set produced by set t, the narrow t method should undercover if these assumptions are true.

If k is the marker corresponding to the maximum column sum, then the columns $X_{\cdot k}$ and $X_{\cdot k_{\text{true}}}$ are correlated, and the correlation is potentially very high, because we are analyzing segmented matrices. For example, in one simulated matrix the average value of the observed correlation between neighboring columns was .48, whereas in the segmented version of the same matrix it was .98. This high level of correlation may explain why the use of t-tests provides overcoverage, because the resulting t statistics exhibit less variation than they would if the use of t-tests was appropriate. In an effort to explore this idea we repeat the coverage simulations for 50×2000 matrices $X = ((G1 - G2) * S) + N$, but now we compute the confidence intervals around the most aberrant marker in X, not $\text{Seg}(X)$.

The plots in Figure 5.4 show the coverage and median widths when we use the test-based

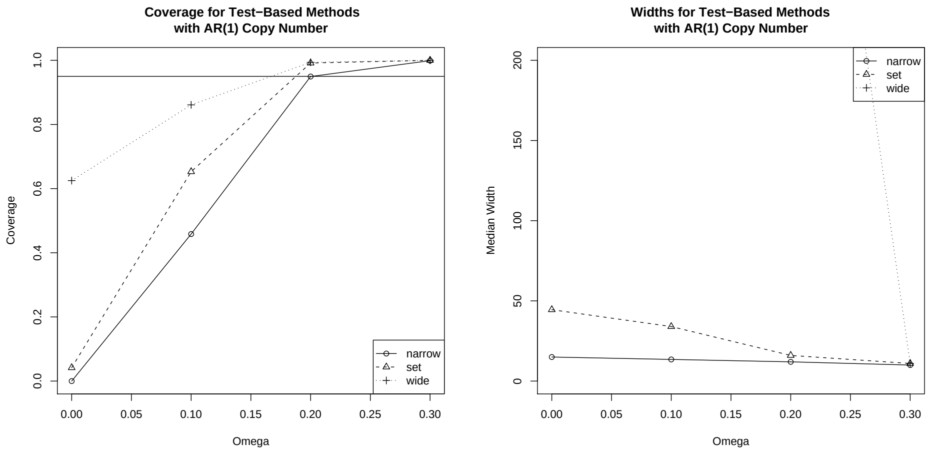

Figure 5.3: Plots Illustrating the Coverage and Median Width of Confidence Intervals Produced by the Test-Based Procedures for Segmented Copy Number Data Simulated with an AR(1) Model

methods to compute confidence intervals in unsegmented matrices X. Unlike the previous simulations, the level of coverage provided by the set t method is approximately correct. However, because the confidence intervals produced by narrow t are contained in the confidence sets produced by set t, the narrow t procedure now provides undercoverage. Although these results do not prove anything, they suggest that correlation in the data, particularly the correlation induced by segmentation, may explain why the test-based methods produce overcoverage. It follows that the t_{n-1} distribution may not be appropriate for assessing the significance of the test statistics that arise when applying the test-based methods. We now present a method for computing confidence intervals that attempts to account for the correlation present in the data.

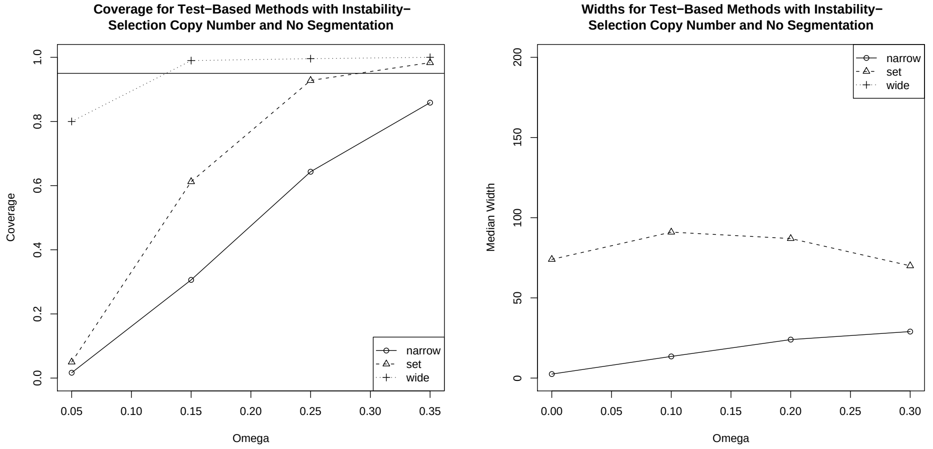

Figure 5.4: Plots Illustrating the Coverage and Median Width of Confidence Intervals Produced by the Test-Based Procedures for Unsegmented Copy Number Data Simulated with the Instability-Selection Model

Bootstrap Test-Based Methods for Computing Confidence Intervals

As noted above, because of the high degree of correlation among the columns of segmented matrices, it may not be appropriate to use standard t-tests in the test-based methods for computing confidence intervals around aberrant markers. Instead of attempting to find the correct parametric distribution, we will use bootstrapping to approximate the distribution of the t-statistics that arise when comparing the mean of $\overline{x}_{\cdot k}\mathbf{1} - X_{\cdot j}$ to zero. By taking the appropriate quantile of these t-statistics, we obtain a threshold τ for the test statistics that is not based on a particular distribution.

We now provide a description of the procedure used to compute the threshold τ in the bootstrap version of the narrow t. Bootstrap versions of the set t and wide t procedures are easily defined once τ is known. Let X be a data matrix, and suppose the most aberrant marker k comes from the maximum column sum of X.

1. Create matrices $X^{b,*}$ for $b = 1, \ldots, B$ by taking boostrap samples of the rows of X.

2. Let k_b be the maximum column sum of $X^{b,*}$.

3. Compute t_b, the t-statistic associated with testing whether the mean of $\overline{x}^{b,*}_{\cdot k_b}\mathbf{1} - X^{b,*}_{\cdot k}$ equals zero.

4. Let τ be the $1 - \alpha$ quantile of the $\{t_b\}$.

In Figure 5.5 we see empirical CDFs of the bootstrap t statistics based on 250 simulated 50×2000 matrices $X = \mathrm{Seg}(Y)$, where $Y = ((G1 - G2) * S) + N$. For the sake of comparison we include the CDF for the standard t_{49} distribution in each plot, as well as a horizontal line at $y = .95$. We see that the empirical CDF changes depending on the effect size, and as a result the bootstrap test threshold τ can be very different from the corresponding threshold for a t_{49} distribution.

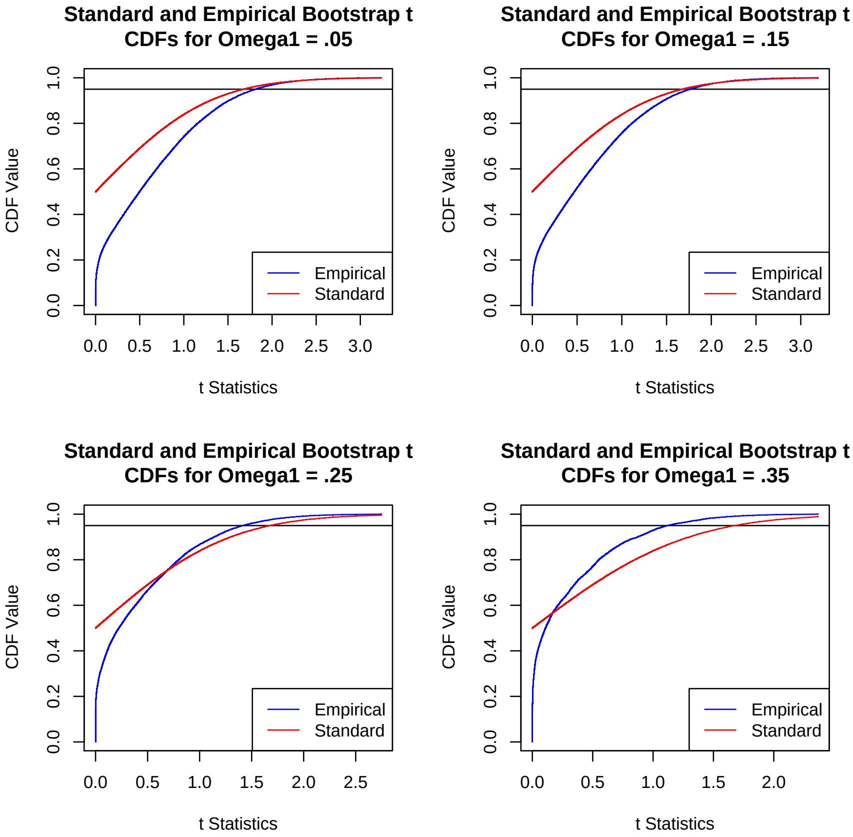

Figure 5.5: Comparison of CDFs of Standard t-Statistics and Empirical t-Statistics Produced by the Bootstrap Test-Based Method

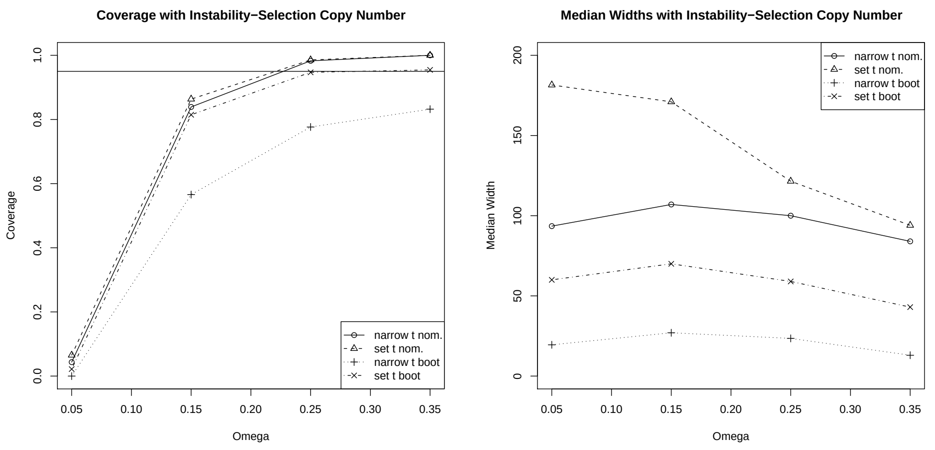

Figure 5.6: Plots Illustrating the Coverage and Median Width of Confidence Intervals Produced by the Bootstrap Test-Based Procedures for Segmented Copy Number Data Simulated with the Instability-Selection Model

Simulation Results for the Bootstrap Test-Based Methods

We now examine the coverage and size of the confidence sets produced by the bootstrap test-based procedures. We use the instability-selection and autoregressive models to simulate our copy number matrices Y, as above, and then use our new methods to find confidence intervals in $X = \text{Seg}(Y)$. Figure 5.6 and 5.7 show the coverage and median size of the level $1 - \alpha$ confidence sets produced by the empirical bootstrap versions of the narrow t and set t procedures when $\alpha = .05$. For the sake of comparison, we also show the values produced by the narrow t and set t methods when ordinary t-tests are performed, and we use the terminology *narrow t nominal* and *set t nominal* when discussing these methods. The values for the wide t versions of both methods are not included, because the confidence intervals they produce are excessively large.

The empirical bootstrap version of set t provides coverage close to $1 - \alpha$, especially when the copy number matrices X are simulated using the instability-selection model, and this is appealing. However, the fact that the median width of the confidence sets for the empirical bootstrap version of set t is consistently larger than the median width of the confidence intervals for the empirical bootstrap version of narrow t suggests that the confidence sets produced by set t regularly consist of multiple disjoint intervals. As noted above, only one of these disjoint intervals contains k, the most aberrant marker. On the other hand, narrow t nominal produces a single interval that always contains k, and this method provides excellent coverage when analyzing segmented matrices. It seems unlikely that researchers using our methods will search for relevant genes in intervals that do not contain the most aberrant marker, so we recommend using the narrow t nominal method, because it produces small confidence intervals that have excellent coverage.

Confidence Intervals in Real Datasets

Although our analysis of confidence sets found in simulated datasets has been informative, we also wish to compute confidence intervals around aberrant markers in real datasets. We

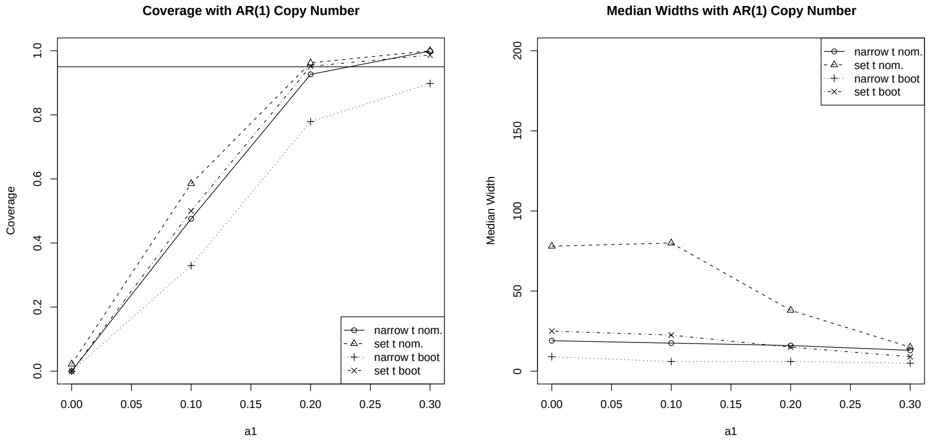

Figure 5.7: Plots Illustrating the Coverage and Median Width of Confidence Intervals Produced by the Bootstrap Test-Based Procedures for Segmented Copy Number Data Simulated with the AR(1) Model

begin by finding the confidence intervals produced by narrow t nominal for the 10 most aberrant markers in the Wilms' tumor dataset of Natrajan et al. (2006). The locations and sizes of the first 10 confidence intervals are presented in Table 5.1, and as we see the sizes vary considerably. Figure 5.8 shows a plot of the column sums for the Wilms' tumor data, and the markers corresponding to the confidence intervals from Table 5.1 are plotted in red. Not surprisingly, the widths of the confidence intervals appear to be correlated with the breadth of the peaks and valleys in the plot of the column sums.

For the sake of comparison, we also examine the confidence intervals produced by narrow t nominal for the 10 most aberrant markers in the glioma dataset of Kotliarov et al. (2006). These confidence intervals are presented in Table 5.2, and one notable difference between these confidence intervals and the ones from the Wilms' tumor dataset is the number of confidence intervals that contain a small number of markers. Here eight of the ten confidence intervals consist of fewer than 10 markers, and none contain more than 30 markers. In contrast, some of the confidence intervals in the Wilms' tumor dataset are considerably larger. Figure 5.9 shows a plot of the column sums of the glioma dataset, and again the markers in the various confidence intervals are plotted in red. Unlike Figure 5.8 above, Figure 5.9 contains no broad peaks or valleys. Thus the aberrations in the glioma dataset appear to be more focal than those in the Wilms' tumor dataset.

In an effort to gain further insight into the differences in the sizes of the confidence intervals found in the Wilms' tumor and glioma datasets, we apply nominal t narrow to the glioblastoma dataset of Veerhaak et al. (2010) and a subset of the lung adenocarcinoma dataset of Weir et al. (2007). The glioblastoma copy number data was computed using Agilent 244K CGH arrays, whereas Affymetrix 250K_Sty arrays were used to produce the lung adenocarcinoma copy number values. Both platforms have approximately 230,000 markers, so unlike the previous two examples, we now have similar marker density. Tables 5.3 and 5.4 list the 10 most aberrant markers for each of the two datasets, along with the lengths of the confidence intervals produced by nominal t narrow. 82 of the 178 samples in the glioma dataset of

Chromosome	Start (bp)	End (bp)	Length (bp)	# Markers
1	142890039	204133558	61243519	107
12	34318876	38270107	3951231	3
8	1790659	7582067	5791408	7
17	31504138	31981051	476913	3
12	28640	34318876	34290236	62
11	26552615	36307918	9755303	17
11	77441058	133753868	56312810	65
7	142791137	142888442	97305	2
1	54039750	113215387	59175637	92
8	48787057	144572963	95785906	91

Table 5.1: Confidence Intervals Produced by Narrow t Nominal for the 10 Most Aberrant Markers in the Wilms' Tumor Dataset of Natrajan et al. (2006)

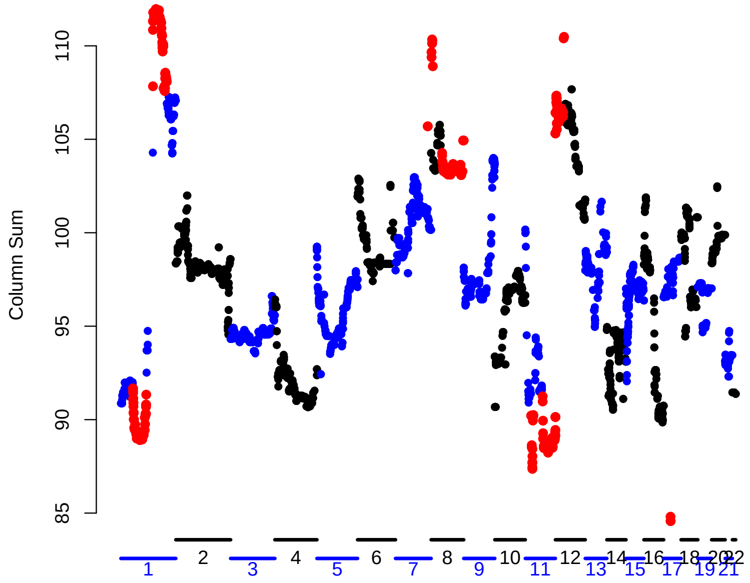

Figure 5.8: Plot of Column Sums for the Wilms' Tumor Data of Natrajan et al. (2006) with Confidence Intervals from Table 7 (Red)

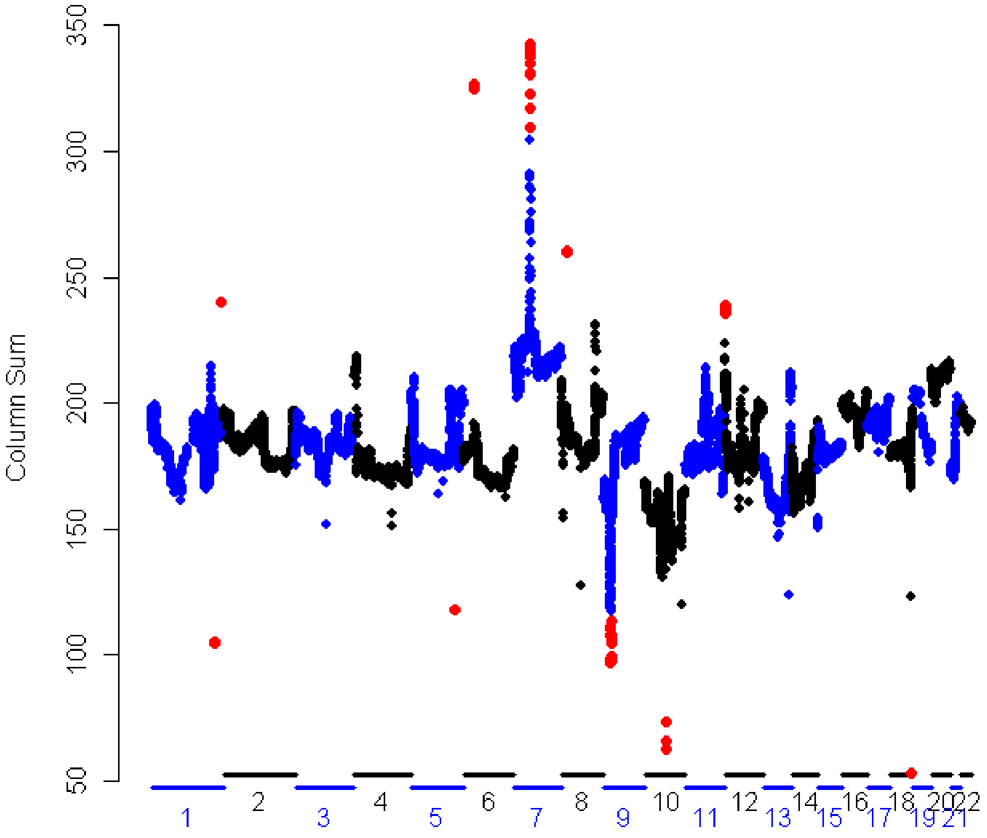

Figure 5.9: Plot of Column Sums for the Glioma Data of Kotliarov et al. (2006) with Confidence Intervals from Table 8 (Red)

Chromosome	Start (bp)	End (bp)	Length (bp)	# Markers
7	54769208	55245458	476250	28
6	32837799	32847290	9491	3
18	67523710	67524503	793	3
10	67452445	67453443	998	7
9	21762317	22268100	505783	23
8	21079131	21084492	5361	4
1	213286968	213287293	325	4
5	147771053	147771548	495	2
1	235706544	235707543	999	2
12	165786	546804	381018	9

Table 5.2: Confidence Intervals Produced by Narrow t Nominal for the 10 Most Aberrant Markers in the Glioma Dataset of Kotliarov et al. (2006)

Chromosome	Start (bp)	End (bp)	Length (bp)	# Markers
7	54816762	55271892	455130	34
9	21899332	22076798	177466	23
6	32586131	32595402	9271	2
8	39363050	39499752	136702	18
14	105594189	105609466	15277	3
11	5738494	5756408	17914	4
7	141687978	141705163	17185	2
12	9452803	9528590	75787	2
10	89478842	89701960	223118	25
20	1506379	1516966	10587	2

Table 5.3: Confidence Intervals Produced by Narrow t Nominal for the 10 Most Aberrant Markers in the Glioblastoma Data of Veerhaak et al. (2010)

Kotliarov et al. (2006) were glioblastoma, which may explain some of the similarities in the sizes of the confidence intervals in the glioblastoma and glioma datasets. In contrast, the lung adenocarcinoma dataset yields a number of large confidence intervals. These results suggest that the differences in the sizes of the confidence interval in the Wilms' tumor and glioma datasets are not solely attributable to differences in array density. Instead, the nature of the underlying aberrations - broad or focal - may also play a significant role.

Chromosome	Start (bp)	End (bp)	Length (bp)	# Markers
5	165712	2193755	2028043	179
8	143949216	146264218	2315002	88
7	1887594	3259036	1371442	149
1	151683544	155570779	3887235	302
20	59800328	62374173	2573845	193
18	62060664	62060963	299	3
19	37779881	41450737	3670856	308
17	69642107	78605474	8963367	686
12	109880996	109884154	3158	2
7	154565104	154576232	11128	2

Table 5.4: Confidence Intervals Produced by Narrow t Nominal for the 10 Most Aberrant Markers in the Lung Adenocarcinoma Dataset of Weir et al. (2007)

Chapter 6

Identifying Subjects that Contribute to Aberrant Markers

Analysis of gene expression data shows that even within the same tumor type, subjects with distinct tumor subtypes may have different expression profiles. For example, Verhaak et al. (2010) use concensus clustering of gene expression data obtained from 202 subjects to identify four distinct subtypes of glioblastoma. These authors also show that different types of copy number aberrations are found in the subtypes. Therefore it is reasonable to expect that different subjects in a given dataset can have distinct copy number profiles. This motivates our interest in determining which subjects "contribute" to a given CNA.

Suppose the most aberrant marker k corresponds to the maximum column sum. Although it is likely that a number of subjects will have high copy number values at marker k, some subjects may have normal or even low copy number. One simple approach for identifying the subjects that have high copy number is to rank the entries in the k^{th} column of the data matrix X. However, because we typically analyze segmented matrices, we do not expect a subject's copy number values to produce a peak that consists of a single marker. Thus it is preferable to rank the mean copy number for each subject over a common interval J. We average over a common interval, as opposed to subject-specific intervals, because this does not allow the shape of any one subject's copy number profile to confer an advantage or disadvantage when the mean copy numbers are ranked. Analysis of subject-specific intervals may be a subject of future research.

Now we examine the common interval J. Large intervals include more markers, and hence more information, but they may lead to a dilution of the effect of the aberrant marker k. Focal aberrations should be more prone to dilution caused by overly large intervals, but the effect on broad aberrations is likely to be minimal. Since we are trying to determine which subjects contribute to the CNA at marker k, we want the effect of marker k on the mean copy number over J, if present for a given subject, to be large. Thus we prefer to make the interval J narrow.

One way to find J is to use the nominal t narrow procedure introduced in Chapter 5 to produce a confidence interval at level $1 - \alpha$. In the simulations below we use $\alpha = .05$. The peeling procedure provides an alternate method for defining J. Let $[a_i, b_i]$ be the intervals defined by the peeling procedure in an $n \times m$ copy number matrix, where $i \in I$ and $I \subseteq \{1, \ldots, n\}$. If $0 < s \leq 100$, the $s\%$ *threshold interval* J is defined to be the set of all markers j that are contained in at least $s\%$ of the intervals $[a_i, b_i]$.

Although ranking average copy number values is conceptually simple, it has some limitations. For example, suppose two subjects have the same mean copy number value on an interval J containing the aberrant marker k. In addition, assume that the first subject has multiple CNAs throughout the genome, whereas the second subject has a single CNA, namely at k. If we rank the mean copy numbers, we cannot distinguish between these two subjects. However, since the first subject has multiple CNAs, we may wish to assess its contribution to the aberration at marker k relative to its contribution to other aberrations. Thus we now present an alternative to ranking average copy numbers.

Suppose J is an interval containing marker k, the location of the maximum column sum, and let $|J|$ be the genomic length of J in base pairs. For any subject i we can compute the empirical p-value of $\overline{x}_{iJ}$, the mean copy number for subject i over the interval J, by comparing $\overline{x}_{iJ}$ to the mean copy number for subject i over all intervals of length $|J|$. We then rank the subjects according to their p-values.

	Alternative True	Null True	
Called Alternative	l_{11}	l_{12}	$l_{1\cdot}$
Called Null	l_{21}	l_{22}	$l_{2\cdot}$
	$l_{\cdot 1}$	$l_{\cdot 2}$	l

Table 6.1: Possible Outcomes for a Binary Classifier

Simulation Results

We wish to determine how well the two ranking approaches identify subjects that contribute to a given aberrant marker. As above, we will consider 50×2000 copy number matrices X whose entries are simulated using either the instability-selection model or an AR(1) model under both the null and alternative hypothesis. Unlike our earlier simulations, however, the rows of X will no longer be iid. Instead, a random subset of rows of size $n_1 < 50$ will be simulated under the alternative hypothesis that a CNA is present at a fixed locus k_{true} with a given effect size. Here we restrict our attention to copy number gains, so we have either $\omega_1 > \delta$ or $a_1 > 0$ when the copy number values are simulated using the instability-selection model or the AR(1) model, respectively. The remaining rows of X will be simulated under the null hypothesis.

Suppose that the methods for defining J and ranking the subjects are fixed. For $1 \leq r \leq n$, let $R \subseteq \{1, \ldots, n\}$ denote the set of the r most highly ranked subjects. Since we know which n_1 rows are simulated under the alternative hypothesis, we can assess the degree to which R reflects the truth. We then record the results of such a binary classifier in a 2×2 table such as Table 6.1.

Our simulated data matrices have 50 rows, and n_1 of them are simulated under the alternative hypothesis. Therefore in Table 6.1 we have $l = 50$, $l_{\cdot 1} = n_1$, and $l_{\cdot 2} = 50 - n_1$. By definition, the *true positive rate* is $\text{TPR} = \dfrac{l_{11}}{l_{\cdot 1}}$, and the *false positive rate* is $\text{FPR} = \dfrac{l_{12}}{l_{\cdot 2}}$. Receiver operating characteristic (ROC) curves are used to assess the accuracy with which R captures the true alternative rows for each value of r in $1, \ldots, n$. We now describe our simulation scheme in detail:

1. Create a 50×2000 copy number matrix Y in which $n_1 < 50$ of the rows are simulated under the alternative hypothesis that a CNA is present at locus k_{true}, and the remaining $50 - n_1$ rows are simulated under the null hypothesis that no CNA is present. We consider $n_1 = 10,\ 20,\ 30$, and 40.

2. Compute $X = \text{Seg}(Y)$, then use the cyclic shift procedure to find $p(T(X))$.

3. If $p(T(X))$ is significant at the $\alpha = .05$ level, use either narrow t nominal or the $s\%$ threshold procedure to define an interval J around k, the marker corresponding to the maximum column sum. Here we consider $s = 90$, but in later simulations we use smaller values of s.

4. Rank the subjects $i = 1, \ldots, n$ according to either (i) their mean copy numbers $\overline{x}_{iJ}$, or (ii) their empirical p-values $p(\overline{x}_{iJ})$.

5. For $r = 1, \ldots, 50$ find the r most highly ranked subjects based on each choice of J and each ranking scheme. Then compute the TPR and FPR for that value of r.

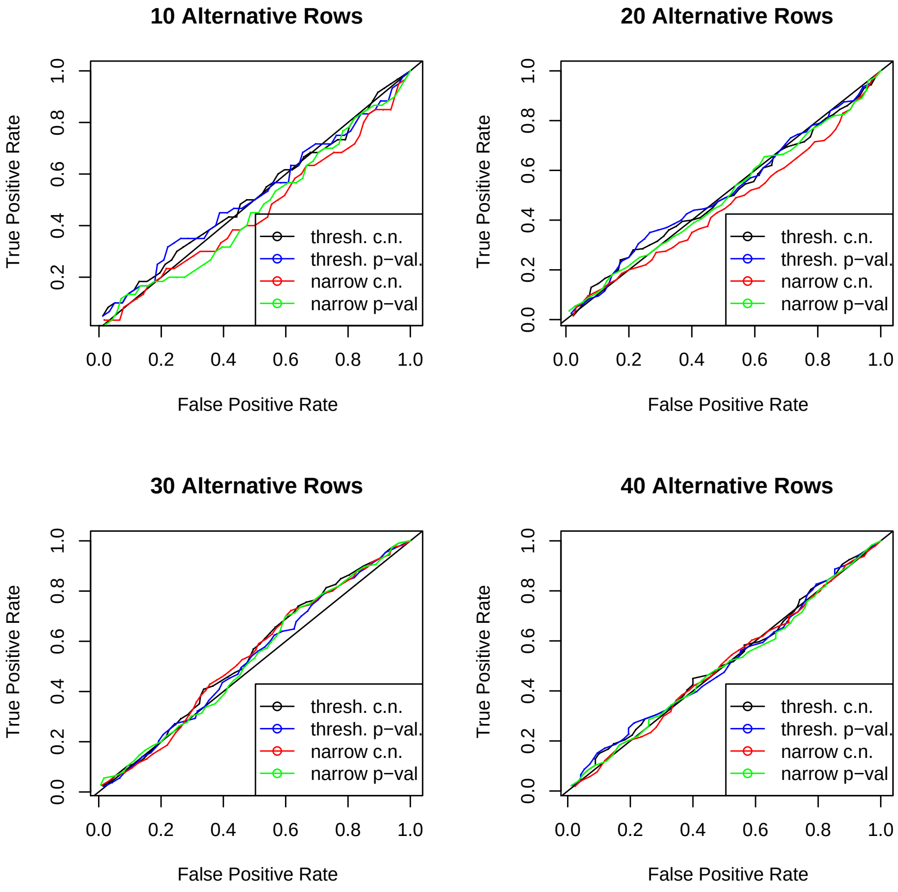

Figure 6.1: ROC Curves Comparing the Classification Methods for Copy Number Data Simulated with an AR(1) Model Under the Null Hypothesis

Plotting the TPR and FPR for $r = 1, \ldots, 50$ gives an ROC curve for a single matrix. The figures below display ROC curves that are averaged over 250 simulated matrices $X = \text{Seg}(Y)$. Figure 6.1 shows the ROC curves when we use an AR(1) model to simulate the copy number values, and the effect size a_1 in the n_1 rows is zero. Here all rows are iid and simulated under the null hypothesis. Because there are no true differences in the rows, we do not expect any of the classification methods to perform better than if they selected rows at random. Thus the ROC curves should follow the $y = x$ line, which they do.

We now examine the performance of the classification methods when the effect size is not zero. Figure 6.2 shows ROC curves for the four methods when we use the AR(1) model to generate the copy number values, and the effect size is $a_1 = .6$. We measure the effectiveness of a classification method by the area under its ROC curve. It follows that ranking mean copy number values over the interval J defined by narrow t nominal (narrow c.n. in the figure) slightly underperforms the other three methods, which are essentially equivalent to each other. However, the difference between this method and the other three decreases as the number of alternative rows increases.

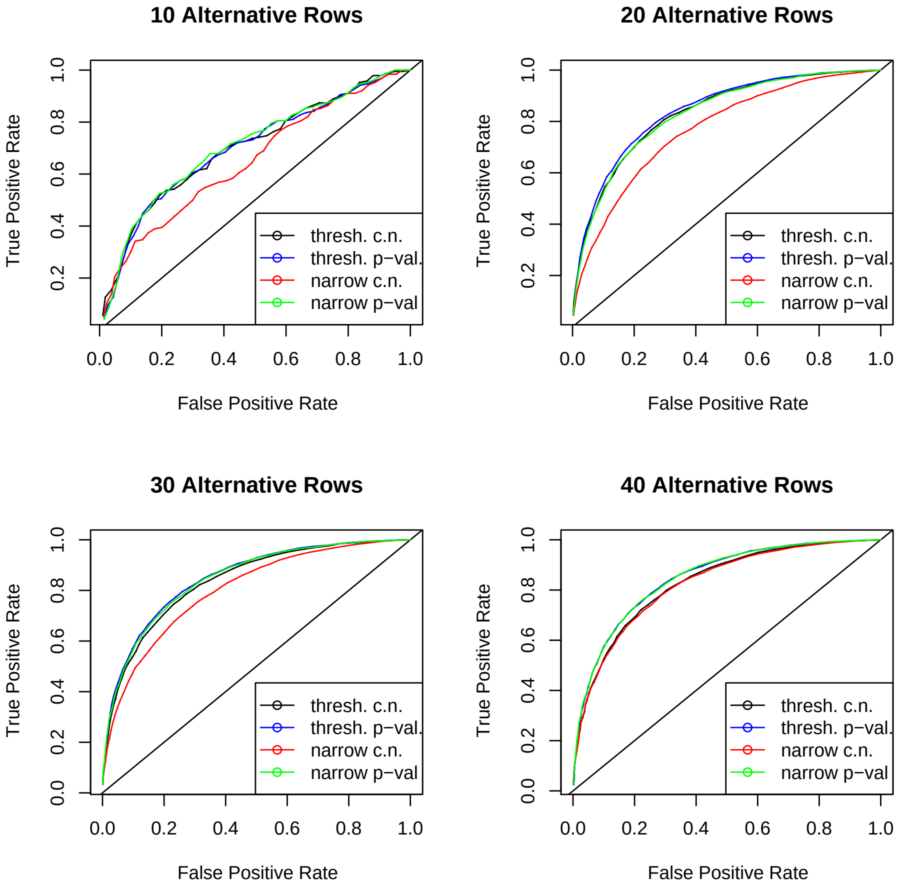

Figure 6.2: ROC Curves Comparing the Classification Methods for Copy Number Data Simulated with an AR(1) Model Under the Alternative Hypothesis

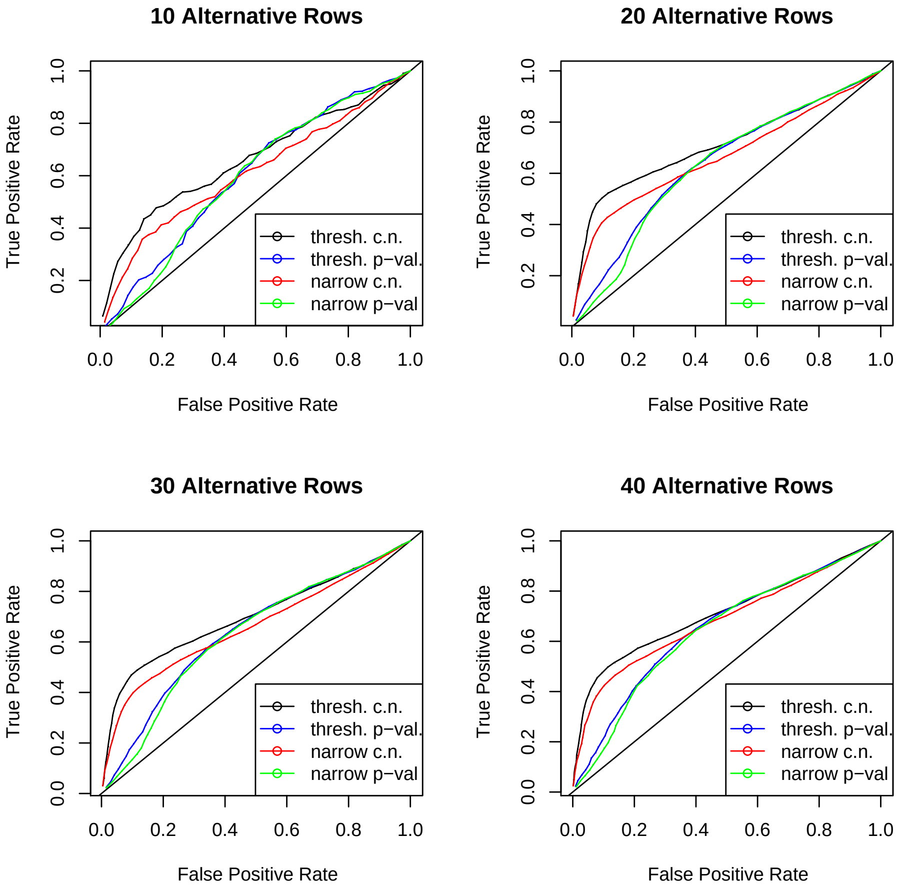

Figure 6.3: ROC Curves Comparing the Classification Methods for Copy Number Data Simulated with the Instability-Selection Model Under the Alternative Hypothesis

Figure 6.3 shows ROC curves when the copy number values are simulated using the instability-selection model, and the effect size is $\omega_1 = .5$. Here we see that the methods based on the use of average copy number over J (thresh c.n. and narrow c.n.) yield better results than the methods based on empirical p-values (thresh p-val and narrow p-val). Moreover, if we restrict our attention to the use of average copy number, the $s\%$ threshold method for defining J is preferable to nominal t narrow. When we use empirical p-values to classify subjects, both methods of defining J yield similar performance.

It is natural to wonder if the performance of classification schemes based on the $s\%$ threshold method for defining J is sensitive to the threshold s. Thus we repeat the simulations described above using $s = 75$ instead of $s = 90$. As s decreases, the size of J increases. Thus we do not expect to see improved results when we use smaller values of s, because the simulated matrices contain focal CNAs. If we compare Figure 6.4 to Figure 6.2 we see that the black curve is lower in Figure 6.4, whereas all others are similar to the corresponding curves in Figure 6.2.

Now we consider the $s\%$ threshold method for defining J with $s = 75$ and matrix entries

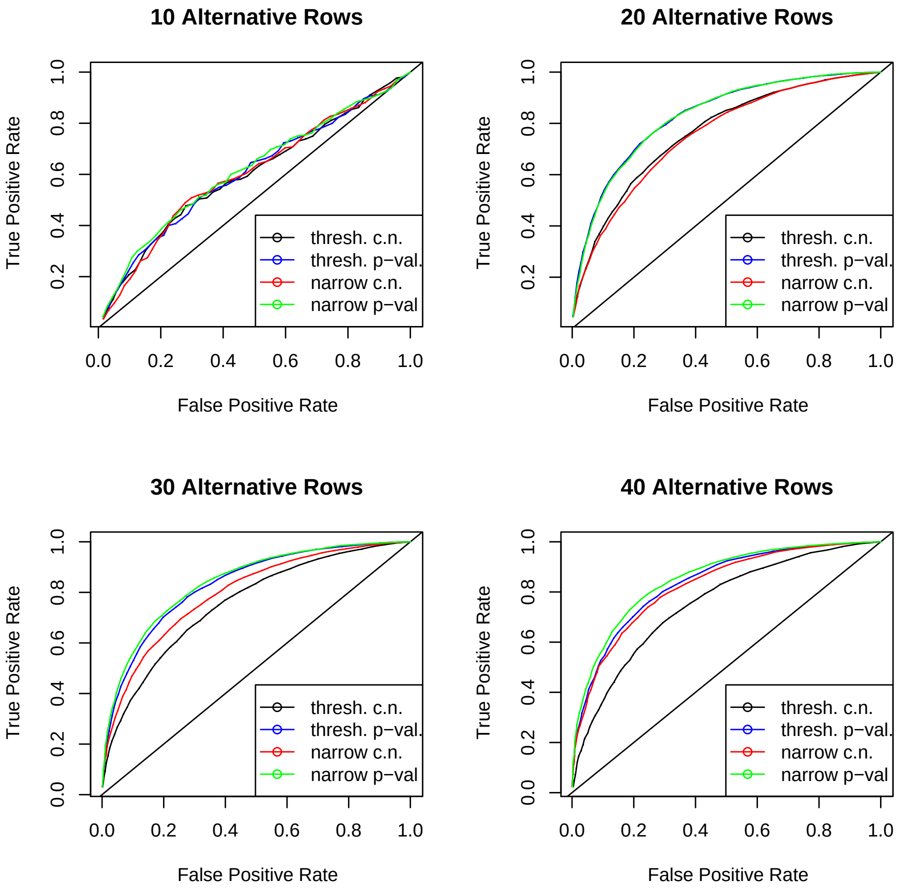

Figure 6.4: ROC Curves Comparing the Classification Methods for Copy Number Data Simulated with an AR(1) Model Under the Alternative Hypothesis and Threshold $s = 75$

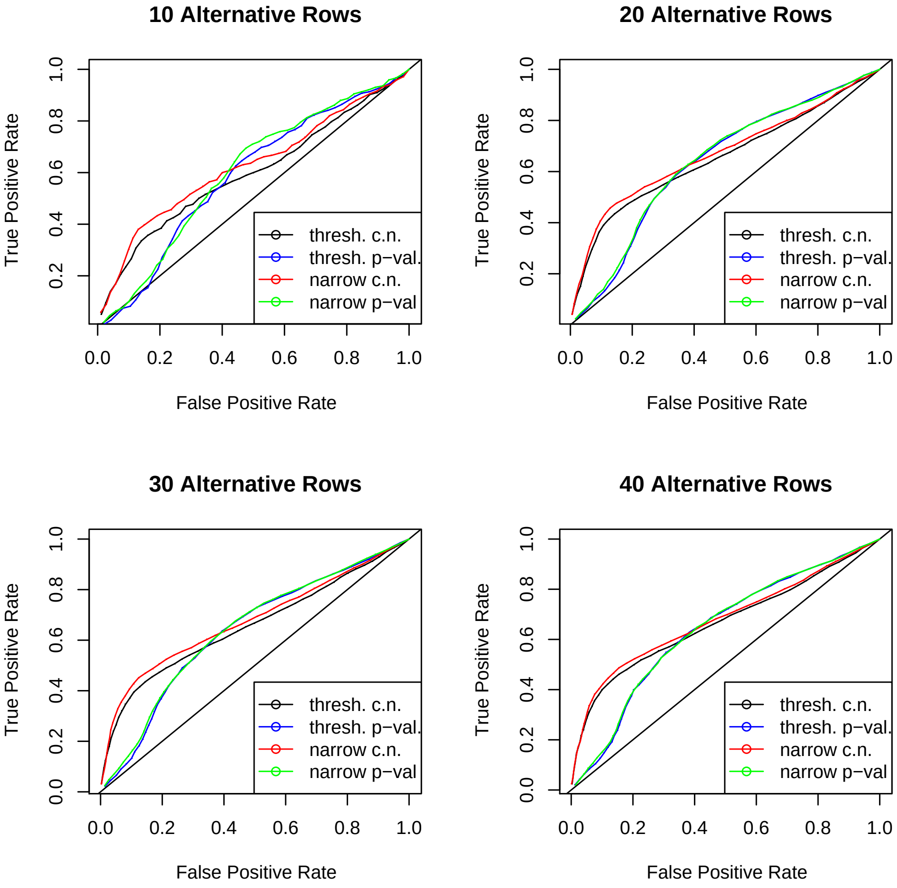

Figure 6.5: ROC Curves Comparing the Classification Methods for Copy Number Data Simulated with the Instability-Selection Model Under the Alternative Hypothesis and Threshold $s = 75$

simulated according to the instability-selection model. If we compare Figure 6.5 with Figure 6.3, we see that the black curve is now lower, but all others are essentially unchanged. These results, combined with the ones in Figure 6.4, suggest that using the mean copy number over J to classify subjects is sensitive to the size of J, whereas the method based on empirical p-values is not.

We conclude this chapter by investigating the degree to which the classification schemes are affected by the underlying level of genomic instability. In the instability-selection model, the parameter δ represents the background probability of LOH. Since we consider copy number matrices $Y = ((G1 - G2) * S) + N$ in which $G1$ and $G2$ are simulated using the instability-selection model with the same parameters λ and δ, the number and size of random CNAs, and hence the level of underlying genomic instability, increase with δ. The segmented matrices $X = \text{Seg}(Y)$ should reflect the level of genomic instability present in Y. Because we only consider datasets simulated with this single model, the scope of our conclusions must be viewed accordingly. However, they provide a basis for future work.

The curves in Figure 6.3 were created using $\delta = .05$. In the following simulations we consider three other possiblities: (1) $\delta = .01$, (2) δ varies randomly across subjects according to a $Unif(.01, .1)$ distribution, and (3) $\delta = .15$. The ROC curves based on these simulations are shown in Figures 6.6, 6.7, and 6.8, respectively, which appear on the following pages.

Comparing Figures 6.3, 6.6, 6.7, and 6.8, we observe decreasing performance among the methods that rank according to mean copy number as the level of genomic instability rises. To see why this might be the case, suppose the level of genomic instability is low. We have $\overline{x}_{iJ}$, the average copy number for row i over the interval J, for $1 \leq i \leq 50$. If row i is simulated under the null hypothesis, then it is unlikely that the rank of $\overline{x}_{iJ}$ will be high, because the low level of variability in the copy number values implies that $\overline{x}_{iJ}$ should be close to $\overline{x}_{\cdot\cdot}$, the grand mean of X. Next consider the rows simulated under the alternative hypothesis, and let $\hat{\omega}_1$ denote the proportion of the n_1 alternative rows that contain an observed CNA at k_{true}. If i is one of the $\hat{\omega}_1 n_1$ rows that contain an observed CNA at k_{true}, then $\overline{x}_{iJ}$ should be much larger than $\overline{x}_{\cdot\cdot}$, and hence highly ranked. Thus when we choose the r most highly ranked subjects based on their mean copy number over J, and $r \leq \hat{\omega}_1 n_1$, we expect to accurately identify subjects that are truly alternative. This level of accuracy should decrease once r becomes sufficiently large, which may explain why the black and red ROC curves in Figure 6.6 change markedly once they reach a certain height.

Now assume the level of genomic instability is high. The increased genomic instability will lead to greater variability of the values $\overline{x}_{iJ}$, regardless of whether row i is simulated under the null or alternative hypothesis. This in turn increases the probability that $\overline{x}_{iJ}$ is highly ranked when row i is simulated under the null hypothesis. As a result, we expect the performance of a classification method based on ranking the $\overline{x}_{iJ}$ to decrease when the level of genomic instability increases.

Next we turn our attention to the methods that rank subjects according to the empirical p-values $p(\overline{x}_{iJ})$. First assume the level of genomic instability is low. If i is one of the $\hat{\omega}_1 n_1$ rows simulated under the alternative hypothesis that contain an observed CNA at k_{true}, then $\overline{x}_{iJ}$ should be large in comparison to the mean copy number over other intervals of length $|J|$ in row i. It follows that $p(\overline{x}_{iJ})$ is small for these i. However, if row i is simulated under the null hypothesis, $p(\overline{x}_{iJ})$ could be small by chance, even if $\overline{x}_{iJ}$ is close to $\overline{x}_{\cdot\cdot}$. Thus classifying subjects according to the ranks of their empirical p-values is potentially more problematic than classifying them according to the ranks of the $\overline{x}_{iJ}$ when we have low levels of genomic instability.

Again assume the level of genomic instability is high. In the instability-selection model, $p_{11} = 1 - (1 - \delta)(1 - e^{-\lambda d})$ is the transition probability for two consecutive 1's when the markers are separated by distance d. Thus as δ increases, so does p_{11}, and hence the expected length of strings of consecutive 1's also increases. Here we are considering $X = \text{Seg}(Y)$, where $Y = ((G1 - G2) * S) + N$. The entries of both X and Y have expected value 0. It follows that 1's in $G1$ or $G2$ produce CNAs in Y and $X = \text{Seg}(Y)$. Thus the expected size of a CNA increases with δ. Under the alternative hypothesis $\omega_1 > \delta$, CNAs at k_{true} will likely be gains. Thus increases in δ lead to larger values of $\overline{x}_{iJ}$ and smaller values of $p(\overline{x}_{iJ})$. Although null rows can still yield small values of $p(\overline{x}_{iJ})$ by chance, these are no more likely now than they were when the level of genomic instability was low. Hence the performance of a classification method based on ranking $p(\overline{x}_{iJ})$ should increase with the level of genomic instability.

Based on our simulations, no single method of identifying subjects is clearly superior to the others. Moreover, the performance of the methods can vary depending on the method used to

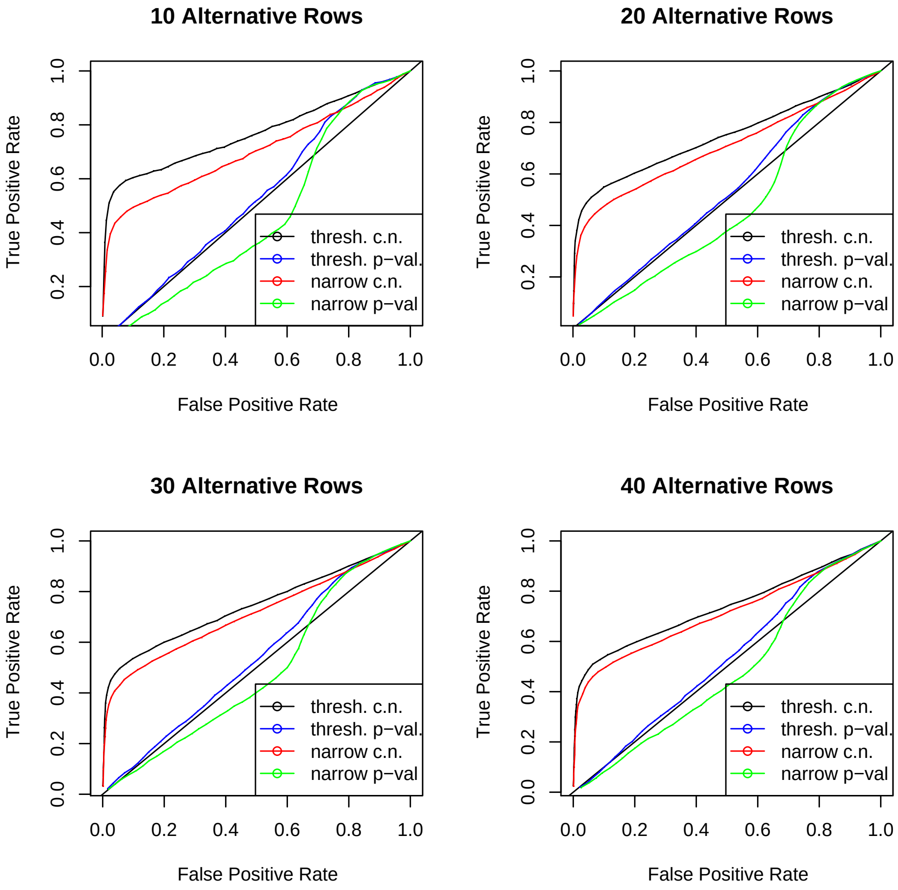

Figure 6.6: ROC Curves Comparing the Classification Methods for Copy Number Data Simulated with the Instability-Selection Model Under the Alternative Hypothesis and $\delta = .01$

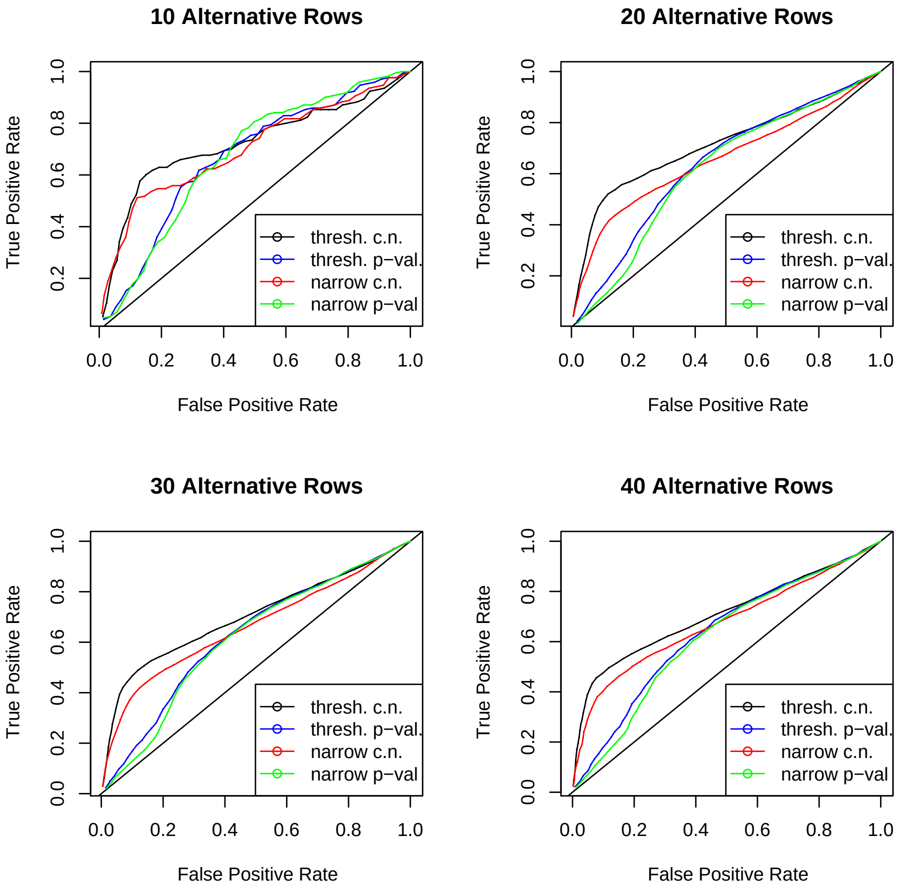

Figure 6.7: ROC Curves Comparing the Classification Methods for Copy Number Data Simulated with the Instability-Selection Model Under the Alternative Hypothesis and $\delta \sim Unif(.01, .1)$

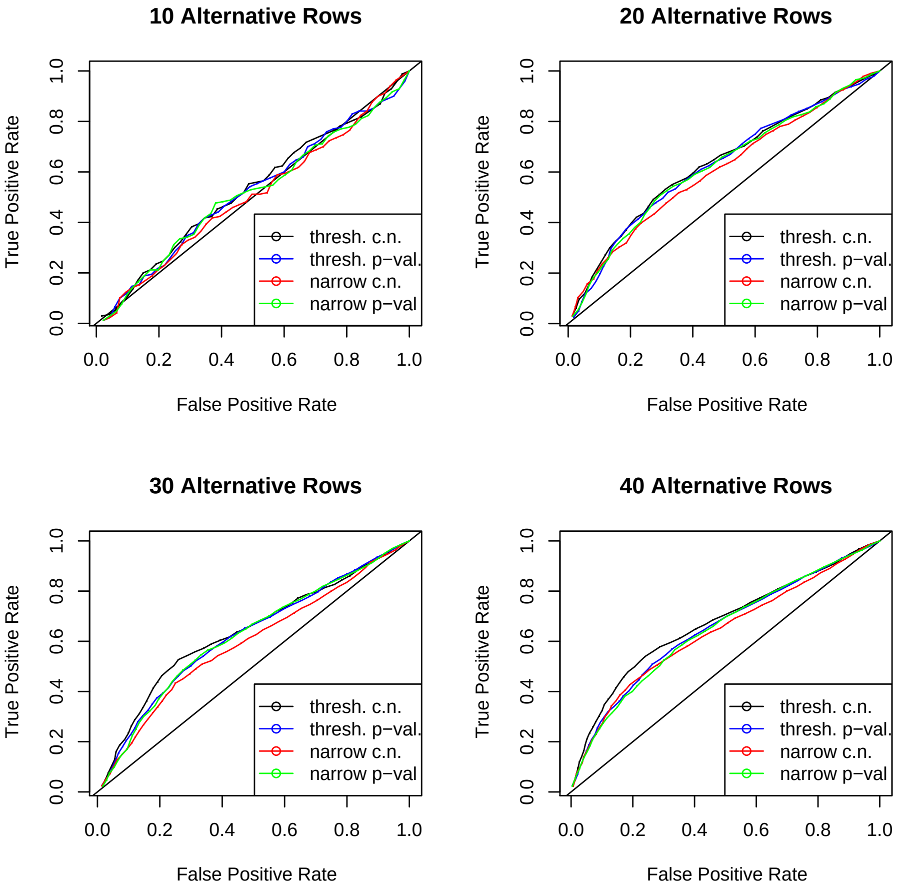

Figure 6.8: ROC Curves Comparing the Classification Methods for Copy Number Data Simulated with the Instability-Selection Model Under the Alternative Hypothesis and $\delta = .15$

simulate the copy number data, the size of the common interval J, and the underlying level of genomic instability. However, our results suggest that (1) small intervals yield better overall performace than large intervals, (2) ranking subjects according to their mean copy number $\overline{x}_{iJ}$ is preferable when the level of genomic instability is low, and (3) ranking subjects according to their empirical p-values $p(\overline{x}_{iJ})$ becomes increasingly appealing as the level of genomic instability rises. Future simulations and analyses of real datasets will undoubtedly provide additional insight into the performance of these or other methods of identifying subjects.

Chapter 7

Joint Analysis of Copy Number and Clinical Data

Historically after the discovery of a tumor of a given type, a patient's overall prognosis was determined by their age, the size of the tumor, the histological type and pathological grade of the tumor, and the level of tumor invasiveness. However, van't Veer et al. (2000) developed a method of predicting patient prognosis based on gene expression data that outperformed the classical methods based on clinical data. Using unsupervised clustering techniques to analyze gene expression data from 295 breast cancer patients, these authors created a gene expression profile based on 70 genes. Their profiling scheme classifies patients as having a "good" or "poor" prognosis, where the two categories are distinguished by the likelihood of developing distant metastases. Because adjuvant therapies are of limited benefit to patients with a good prognosis, physicians can use this classification scheme to match each patient to the most appropriate level of treatment.

It is natural to hope that methods similar to those of van't Veer et al. (2000) can be developed for DNA copy number data. We will not attempt to create profiling methods based on DNA copy number data here, although this may be the subject of future research. Instead, we consider a related problem: Given copy number and clinical data from a set of patients, are there associations between the CNAs and any of the clinical variables? Information about such associations is potentially useful for researchers, because it could provide additional insight into disease progression. We present a testing procedure, as well as some results based on preliminary investigations.

A Testing Procedure for Copy Number and Covariate Data

Tumors in the glioma dataset of Kotliarov et al. (2006) are classified according to tumor type - glioblastoma, astrocytoma, oligodendroglioma, or mixed glioma - as well as tumor grade - 2, 3, or 4. Although we have copy number data for all 178 subjects in the study, tumor grade data is only available for 174 subjects. Thus we restrict our attention to these 174 subjects. In an effort to determine if tumor grade is associated with DNA copy number, for each of the $j = 1, \ldots, 113199$ autosomal markers we construct a linear model $X_{\cdot j} = \beta_{0,j} + \beta_{1,j} Z_{\cdot 1} + \beta_{2,j} Z_{\cdot 2}$. Here $X_{\cdot j} = (X_{1,j}, \ldots, X_{174,j})$ is a vector containing the copy number measurements at the j^{th} marker, $Z_{\cdot 1} = (Z_{1,1}, \ldots, Z_{174,1})$ is a vector of indicator variables $Z_{i,1}$ that equal 1 if and only if the i^{th} subject has tumor grade 3, and $Z_{\cdot 2} = (Z_{1,2}, \ldots, Z_{174,2})$ is a vector of indicator

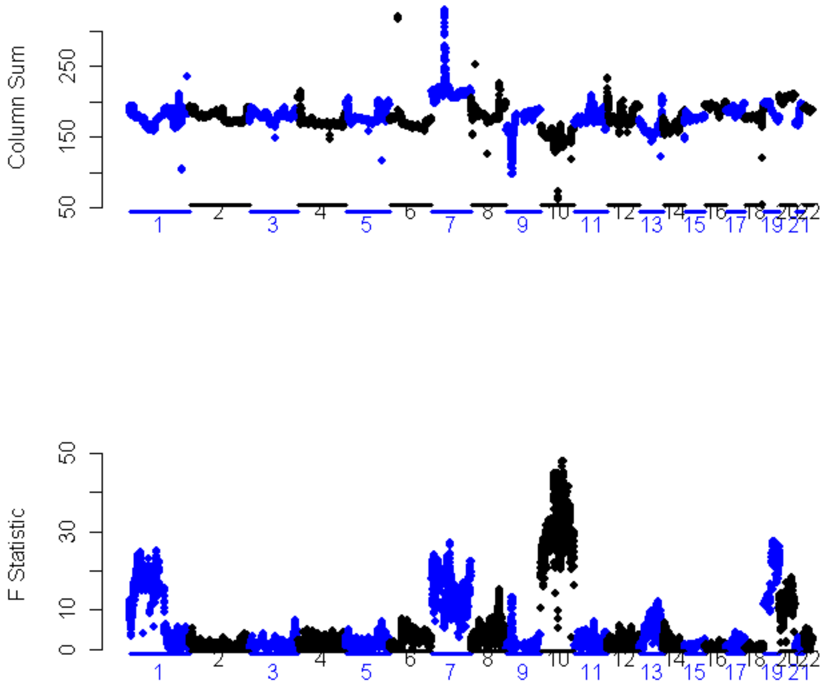

Figure 7.1: Column Sums from the Glioma Dataset of Kotliarov et al. (2006) (top), F Statistics from the Linear Model Regressing Copy Number on Tumor Grade (bottom)

variables $Z_{i,2}$ that equal 1 if and only if the i^{th} subject has tumor grade 4.

The top of Figure 7.1 shows a plot of the column sums of X, and the bottom shows a plot of the F statistics produced by the test of $H_{0,j} : \beta_{1,j} = \beta_{2,j} = 0$ vs. $H_a : \beta_{1,j} \neq 0$ or $\beta_{2,j} \neq 0$ for $j = 1, \ldots, 113199$. Large F statistics appear throughout chromosome 10, and the marker that produces the maximum F statistic is close to the locus of a statistically significant loss. Thus it is natural to wonder if the two sites are the same. We now formalize this question in the context of hypothesis testing.

Clearly the underlying location of a CNA will be different than the locus that produces a large test statistic if the two lie on different chromosomes. Thus suppose that X is an $n \times m_c$ matrix of copy number measurements from n subjects at m_k markers on chromosome c. Next let Z be an $n \times l$ matrix representation of a given clinical variable suitable for use in a linear model. In the above example, Z is a 174×3 matrix whose columns are a vector of 1's representing covariate values for the intercept and the values of indicator variables that classify tumor stage into one of three values. Write $X_{\cdot j}$ for the entries in the j^{th} column of X, and consider the linear model $X_{\cdot j} = Z\beta_j$, where $\beta_j = (\beta_{0,j}, \beta_{1,j}, \ldots, \beta_{l-1,j})$. Finally, let F_j be the test statistic corresponding to the global test that $\beta_{s,j} = 0$ for all $s \geq 1$. If k is the most aberrant marker in X, we wish to test $H_0 : k = arg\ max(F_j)$ vs. $H_a : k \neq arg\ max(F_j)$. We now outline our hypothesis testing procedure when k corresponds to the maximum column sum:

1. For $b = 1, \ldots, B$, form the $n \times m_c$ matrix $X^{b,*}$ and the $n \times l$ matrix $Z^{b,*}$ by taking the same bootstrap sample of the rows of X and Z.

2. Record $k^{b,*}$, the location of the maximum column sum in $X^{b,*}$, and $arg\ max(F_j^{b,*})$ for each value of b. Here $F_j^{b,*}$ is the test statistic for the global test in the linear model $X_{\cdot j}^{b,*} = Z^{b,*}\beta_j$ for $j = 1, \ldots, m_c$.

3. Compute $d_b = k^{b,*} - arg\ max(F_j^{b,*})$. In practice we normalize the marker positions so that the values of j lie on the interval $(0, 1)$.

4. If $n_{pos} = |d_b : d_b > 0|$, and $n_{neg} = |d_b : d_b < 0|$, then the p-value for the hypothesis test is $\frac{2}{B}\min(n_{pos}, n_{neg})$.

The linear model in the above procedure can easily be replaced by with a Cox proportional hazards model if the clinical data consists of censored survival times. Moreover, $-\log_{10}(p-\text{values})$ can be used instead of test statistics.

Analyzing the Glioma Data

Now we apply the testing procedure to specific chromosomes of the glioma dataset of Kotliarov et al. (2006), beginning with chromosome 10. The top two plots in Figure 7.2 show the column sums of the copy number matrix for chromosome 10 for each marker and a plot of the F statistics for each marker. The most aberrant marker is $k = 2965$, and the maximum test statistic appears at marker 3763. The bottom left plot shows a scatterplot of $k^{b,*}$ vs. $arg\ max(F_j^{b,*})$ based on $B = 500$ bootstrap samples, the $y = x$ line, and a red X marking the location of the k and the largest test statistic in the observed data. Finally, in the bottom right we have a histogram of the d_b values, along with a vertical line at 0.

The scatterplot in the lower left shows that the location of $k^{b,*}$ is remarkably stable when we take bootstrap samples. Although the values of $arg\ max(F_j^{b,*})$ exhibit more variation,

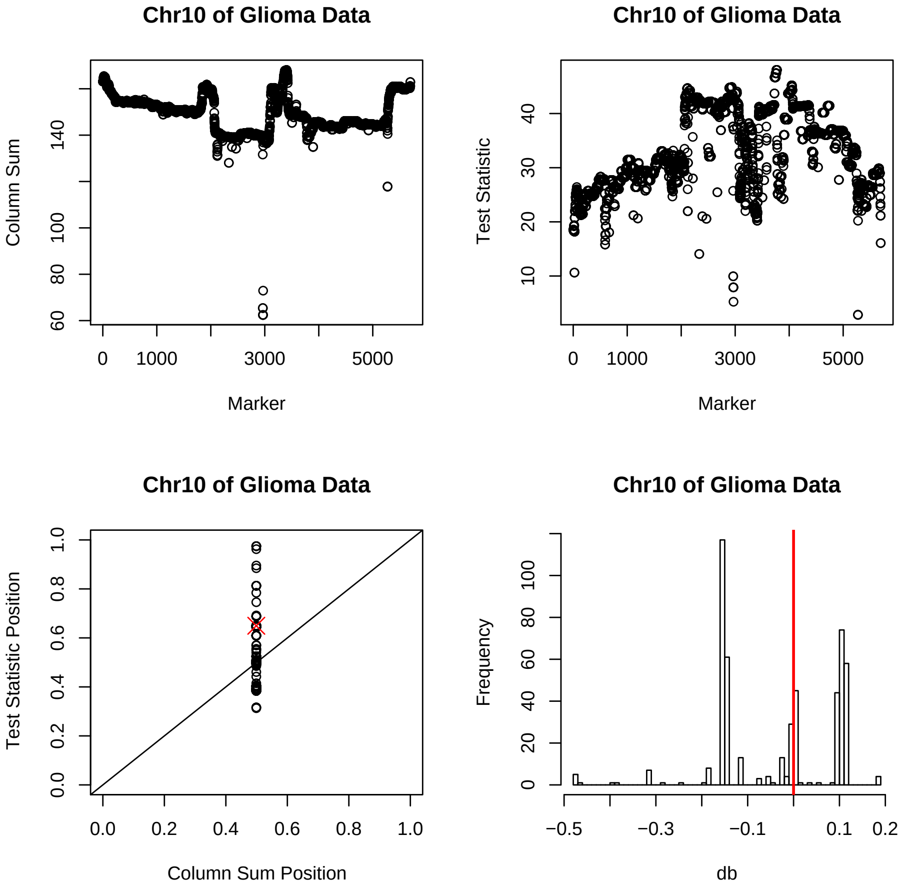

Figure 7.2: Plot of the Column Sums of the Chromosome 10 Glioma Data of Kotliarov et al. (2006) (top left); Plot of the Test Statistics Based on a Regression Model of Copy Number on Tumor Stage (top right); Scatterplot of Locations for the Most Aberrant Marker and Largest Test Statistic Under Bootstrap Resampling (bottom left); Histogram of Values of d_b Under Bootstrap Resampling (bottom right)

the peaks in the histogram of d_b and the fact that the $k^{b,*}$ are essentially constant imply that certain markers repeatedly yield the maximum test statistic under bootstrapping. Our testing procedure yields a p-value of .916, so we fail to reject the null hypothesis that the most aberrant marker produces the largest test statistic.

Next we repeat this analysis for chromosome 1, and here we obtain very different results. Based on the plot of the column sums in Figure 7.3, there appears to be a loss around marker 7818 and a gain around marker 8863, both of which lie in the q arm. Moreover, the scatterplot in the bottom left of Figure 7.3 shows that these two loci consistently reappear when we find the most aberrant marker under bootstrap resampling. While the observed test statistics in the observed data are large throughout most of the p arm of the chromosome, they are small on the q arm. The scatterplot shows that under bootstrap resampling the maximum test statistic is almost always found in the p arm. The resulting values of d_b are all positive, so the p-value for our test is zero. Thus we reject the null hypothesis that the most aberrant marker produces the largest test statistic.

The results of the analysis of chromosome 1 show that the testing procedure has some promise, because for this dataset it is able to distinguish markers that lie on distinct chromosome arms. On the other hand, when analyzing chromosome 10, the procedure produces a very high p-value even though in the observed data there are almost 800 markers separating the most aberrant marker and the marker producing the largest test statistic. At this point it is not clear if the method has limited resolution, or if we can obtain superior results after making some refinements. Only limited simulation studies have been performed to date, but we hope that future investigations will provide additional insight.

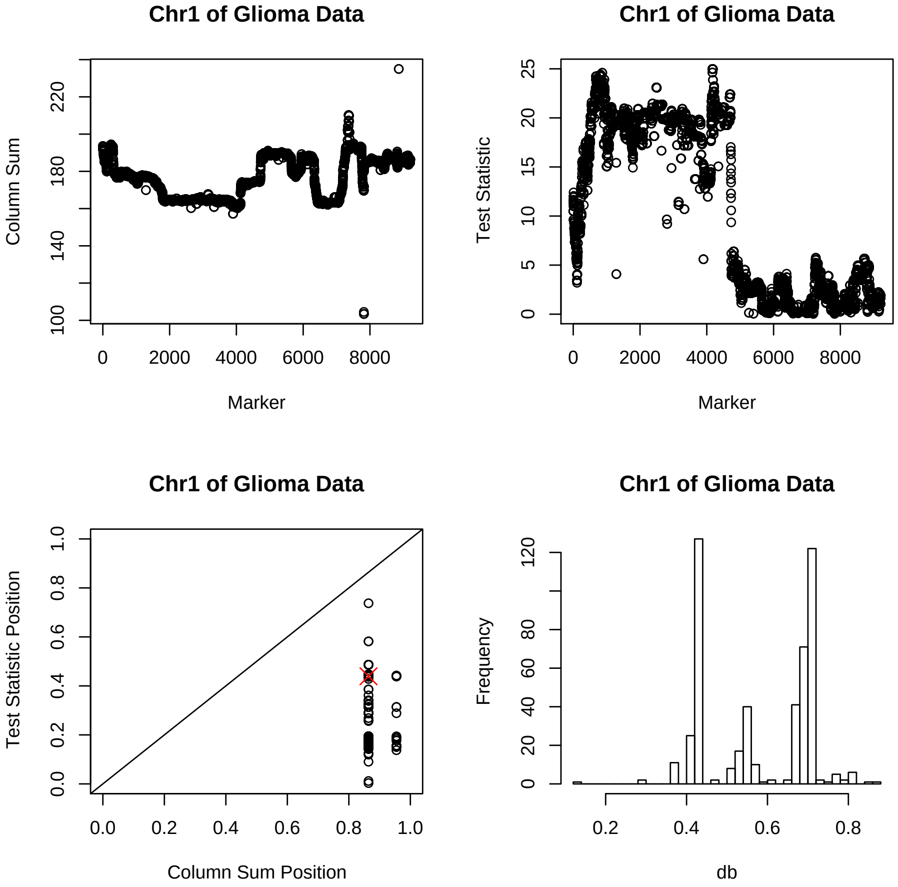

Figure 7.3: Plot of the Column Sums of the Chromosome 1 Glioma Data of Kotliarov et al. (2006) (top left); Plot of the Test Statistics Based on a Regression Model of Copy Number on Tumor Stage (top right); Scatterplot of Locations for the Most Aberrant Marker and Largest Test Statistic Under Bootstrap Resampling (bottom left); Histogram of Values of d_b Under Bootstrap Resampling (bottom right)

Chapter 8

Statistical Theory

We begin this chapter by formally defining different distributions of a global statistic $T(X)$ that arise when resampling the rows of X. Although these definitions are applicable in any resampling scheme, our subsequent discussion focuses on DiNAMIC's cyclic shift procedure. We present some theoretical results when the rows $X_{i\cdot}$ of X are iid and follow a given parametric distribution. Two cases are considered in depth: (1) $X_{i\cdot} \sim MVN(\mu, \Sigma)$, and (2) the $X_{i\cdot}$ are generated by the instability-selection model of Newton et al. (1998). We also present some preliminary results when $X_{i\cdot} \sim MVN(\mu, \Sigma)$ and Σ has an AR(1) correlation structure that may be useful in future studies of DiNAMIC's asymptotic behavior.

The Unconditional and Conditional Distributions

Although the details vary from one method to the next, DiNAMIC, GISTIC, KC-SMART, STAC, and MSA use permutation-based methods for resampling a data matrix X. We wish to define distributions that arise in the context of resampling, and for convenience we write π for a fixed but unspecified resampling procedure. For example, $\pi = \sigma$ when we restrict our attention to cyclic shifts.

Let π be a specific resampling procedure, and define $\mathcal{P}$ to be the set of all possible resampled versions of X. Thus $|\mathcal{P}| = m^n$ if we resample using cyclic shifts, and $|\mathcal{P}| = (m!)^n$ if we resample by taking random permutations of the elements in each row of X, as is done for KC-SMART and GISTIC. Note that $|\mathcal{P}|$ is not uniquely defined for STAC and MSA, because the number of distinct STAC permutation depends on both the number and the size of the aberrant regions in each row of X. For the sake of generality, in the following definitions we assume that X is random and follows a given distribution.

1. We write $X \sim F^n$ when X is an $n \times m$ matrix with rows $X_{1\cdot}, X_{2\cdot}, \ldots X_{n\cdot}$ that are iid with (multivariate) distribution function F. Let f^n be the density function or probability mass function associated with F^n.

2. If $X \sim F^n$, the distribution function of $T(X)$ is

$$G_{TU}(t) = p(T(X) > t),$$

the *true unconditional distribution.*

3. If $X \sim F^n$ and π is a random resampling that is independent of X, then the distribution of $T(\pi(X))$ is

$$G_{RU}(t) = p(T(\pi(X)) > t),$$

the *resampling unconditional distribution.*

4. The *true conditional probability* of a resampling $\tilde{\pi}(X)$ of X is

$$p_{TC}(\tilde{\pi}(X)|\{\pi(X) : \pi \in \mathcal{P}\}) = \frac{f^n(\tilde{\pi}(X))}{\sum_{\pi \in \mathcal{P}} f^n(\pi(X))}.$$

   It is important to note that under F^n the values $\{f^n(\pi(X))\}_{\pi \in \mathcal{P}}$ need not be identical. More generally, the true conditional probability of an event A is

$$p_{TC}(A|\{\pi(X) : \pi \in \mathcal{P}\}) = \sum_{\tilde{\pi} \in \mathcal{P}} p_{TC}(\tilde{\pi}(X)|\{\pi(X) : \pi \in \mathcal{P}\}) I(\tilde{\pi}(X) \in A)).$$

5. The *resampling conditional probability* of a resampling $\tilde{\pi}(X)$ of X is

$$p_{RC}(\tilde{\pi}(X)|\{\pi(X) : \pi \in \mathcal{P}\}) = \frac{f^n(\tilde{\pi}(X))}{\sum_{\pi \in \mathcal{P}} f^n(\pi(X))}$$

   under the assumption that the values $\{f^n(\pi(X))\}_{\pi \in \mathcal{P}}$ are identical. Thus

$$p_{RC}(\tilde{\pi}(X)|\{\pi(X) : \pi \in \mathcal{P}\}) = \frac{1}{|\mathcal{P}|}.$$

   It follows that the resampling conditional probability of an event A is

$$p_{RC}(A|\{\pi(X) : \pi \in \mathcal{P}\}) = \sum_{\tilde{\pi} \in \mathcal{P}} \Big(\frac{1}{|\mathcal{P}|}\Big) I(\tilde{\pi}(X) \in A).$$

6. The *true conditional distribution* of $T(X)$ is

$$\begin{aligned} G_{TC}(t) &= p_{TC}(T(\tilde{\pi}(X)) > t|\{\pi(X) : \pi \in \mathcal{P}\}) \\ &= \sum_{\tilde{\pi} \in \mathcal{P}} p_{TC}(\tilde{\pi}(X)|\{\pi(X) : \pi \in \mathcal{P}\}) I(T(\tilde{\pi}(X)) > t). \end{aligned}$$

7. The *resampling conditional distribution* of $T(X)$ is

$$G_{RC}(t) = p_{RC}(T(\tilde{\pi}(X)) > t|\{\pi(X)\}) = \sum_{\tilde{\pi} \in \mathcal{P}} \Big(\frac{1}{|\mathcal{P}|}\Big) I(T(\tilde{\pi}(X)) > t).$$

8. Let $\tilde{\mathcal{P}}$ be a subset of $\mathcal{P}$ containing N elements. If N is large, then

$$\hat{G}_{RC}(t) = \sum_{\tilde{\pi} \in \tilde{\mathcal{P}}} \Big(\frac{1}{N}\Big) I(T(\tilde{\pi}(X)) > t)$$

is an approximation to $G_{RC}(t)$.

In the remainder of the chapter we illustrate how these distributions arise naturally in the context of the cyclic shift procedure, and why it is important to be aware of the differences between them. In an effort to emphasize that we are restricting our attention to cyclic shifts, we replace the term *resampling* from the preceeding definitions with *cyclic.* Similarly, we replace the subscript R with C. For example, this leads to the definition of the cyclic conditional distribution

$$G_{CC}(t) = p_{CC}(T(\tilde{\sigma}(X)) > t|\{\sigma(X)\}) = \sum_{\tilde{\sigma}\in\mathcal{P}} \Big(\frac{1}{m^n}\Big) I(T(\tilde{\sigma}(X)) > t).$$

Stochastic Processes

Definition. Suppose $(\Omega, \mathcal{F}, P)$ is a probability space, and let T be an index set. A *stochastic process* F with state space X is a collection of X-valued random variables $\{F_t : t \in T\}$. Given any finite subset $\tilde{T} = \{t_1, \ldots, t_n\}$ of T, the distribution function $\tilde{F} = F|_{\tilde{T}}$ is a *finite dimensional distribution* of F.

Definition. A stochastic process F is *stationary* if its finite dimensional distributions are translation-invariant. Specifically, if $\tilde{T} = \{t_1, \ldots, t_n\}$ and $\tilde{F} = F|_{\tilde{T}}$ is a *finite dimensional distribution* of F, then $\tilde{F} = \tilde{F}_t$, where $\tilde{F}_t = F_{\tilde{T}+t}$ and $\tilde{T} + t = \{t_1 + t, \ldots, t_n + t\}$.

Anderson (1960) defined a Gaussian process $X(\theta)$ on a circle with the properties that (i) $0 \le \theta \le 2\pi$, (ii) $X(0) = X(2\pi)$, and (iii) $E(X(\theta)) = 0$. It was shown that $X(\theta)$ has a Markov property if the correct covariance structure is chosen. This motivates the following definition.

Definition. Let Σ be an $n \times n$ covariance matrix for a Gaussian process $X(\theta)$ defined at angles $0 = \theta_1 < \theta_2 < \cdots < \theta_n \le 2\pi$ on a circle. Σ has an *Anderson* covariance structure if $\mathrm{Cov}(X(\theta_i), X(\theta_j)) = \dfrac{\cosh[\lambda(|\theta_i - \theta_j| - \pi)]}{\cosh[\lambda\pi]}$ for some λ.

Some matrices with an Anderson covariance structure satisfy a more general property, which we define now.

Definition. An $n \times n$ matrix M is *circulant* if it has the form

$$M = \begin{pmatrix} m_0 & m_1 & m_2 & \ldots & m_{n-1} \\ m_{n-1} & m_0 & m_1 & \ldots & m_{n-2} \\ m_{n-2} & m_{n-1} & m_0 & \ldots & m_{n-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ m_1 & m_2 & m_3 & \ldots & m_0 \end{pmatrix}.$$

Therefore for $i, j \in \{0, 1, \ldots, n-1\}$ the (i, j) entry of M is $M_{(i,j)} = m_{[j-i]}$, where $[y]$ denotes y *mod* n and $m_{[j-i]}$ is the appropriate entry of the first row of M. Note that we start indexing our rows and columns at 0, not 1.

We now show that some matrices with an Anderson covariance structure are circulant.

Lemma 8.1. *If the angles $0 = \theta_1 < \theta_2 < \cdots < \theta_n \le 2\pi$ are equally spaced, then a matrix with the Anderson covariance structure is circulant.*

Proof. If we define $\theta_k = \dfrac{2\pi(k-1)}{n}$ for $k = 1, \ldots, n$, then the (i, j) entry of the Anderson covariance matrix is $M_{(i,j)} = \mathrm{Cov}(X(\theta_i), X(\theta_j))$. Therefore

$$M_{(i,j)} = \frac{\cosh[\lambda(|\frac{2(i-1)\pi}{n} - \frac{2(j-1)\pi}{n}| - \pi)]}{\cosh[\lambda\pi]} = \frac{\cosh[\lambda(\frac{2\pi}{n}|i-j| - \pi)]}{\cosh[\lambda\pi]}.$$

For circulant matrices $M_{(i,j)} = m_{[j-i]}$, but it is easy to see that when the θ_k are defined as above the $m_{[j-i]}$ entry of the first row is $\dfrac{\cosh[\lambda(\frac{2\pi}{n}|j-i| - \pi)]}{\cosh[\lambda\pi]}$. □

Next we prove a result about multivariate normal densities with circulant covariance matrices.

Lemma 8.2. *Suppose* $\mathbf{z} = (z_0, z_1, \ldots, z_{n-1}) \sim MVN(\mu, \Sigma)$, *where* μ *is a constant vector and* Σ *is circulant. If* f *is the density of* $\mathbf{z}$, *then* $f(z_0, z_1, \ldots, z_{n-1}) = f(z_{n-1}, z_0, z_1, \ldots, z_{n-2})$. *It follows that* f *is invariant under any cyclic shift of the arguments.*

Proof. Assume first that $\mu = \mathbf{0}$ so that $f(\mathbf{z}) = \dfrac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\Big(-\dfrac{1}{2}\mathbf{z}^T\Sigma^{-1}\mathbf{z}\Big)$. It is known that the inverse of a circulant matrix is also circulant. Therefore it suffices to show that $\mathbf{z}^T M\mathbf{z} = \tilde{\mathbf{z}}^T M\tilde{\mathbf{z}}$ whenever M is the circulant matrix given in the above definition, $\mathbf{z} = (z_0, z_1, z_2, \ldots, z_{n-1})$, and $\tilde{\mathbf{z}} = (z_{n-1}, z_0, z_1, \ldots, z_{n-2})$. Both $\mathbf{z}^T M\mathbf{z}$ and $\tilde{\mathbf{z}}^T M\tilde{\mathbf{z}}$ are sums of terms of the form $z_i z_j$ for $i \le j$. We will show that the coefficients of $z_i z_j$ are the same in each expression.

First consider $\mathbf{z}^T M\mathbf{z}$. It is easy to see that the coefficient of z_i^2 is $M_{(i,i)} = m_0$, and when $i < j$ the coefficient of $z_i z_j$ is $M_{(i,j)} + M_{(j,i)} = m_{[j-i]} + m_{[i-j]}$. Next we examine $\tilde{\mathbf{z}}^T M\tilde{\mathbf{z}}$. The coefficient of z_i^2 is now $M_{([i+1],[i+1])}$, but for any i this is m_0 because M is circulant. If $i < j$ the coefficient of $z_i z_j$ is $M_{([i+1],[j+1])} + M_{([j+1],[i+1])}$. However, since M is circulant we may write this as $m_{[(j+1)-(i+1)]} + m_{[(i+1)-(j+1)]} = m_{[j-i]} + m_{[i-j]}$.

For a general constant vector $\mu = (c, \ldots, c)$ we replace the terms $z_i z_j$ in the above argument with $(z_i - c)(z_j - c)$. □

Suppose X is an $n \times m$ matrix whose rows $X_{i\cdot}$ are iid $MVN(\mu, \mathbf{\Sigma})$, where μ is a constant vector and $\mathbf{\Sigma}$ is circulant. If we write f^n for the density of X, then $f^n(X) = \prod_{i=1}^{n} f(X_{i\cdot})$. However, Lemma 8.2 implies that $f^n(\sigma(X)) = f^n(X)$ for any cyclic shift σ. This implies that the true conditional probability $p_{TC}(\sigma(X))$ is $\dfrac{1}{m^n}$ for any cyclic shift σ. As a result, the true conditional distribution $G_{TC}(t)$ of $T(X)$ coincides with the cyclic conditional distribution $G_{CC}(t)$.

Assume that X is an $n \times m$ matrix with rows $X_{1\cdot}, \ldots, X_{n\cdot}$ that are iid $MVN(\mu, \Sigma)$, and let $\{\sigma^i\}_{i=1}^N$ be random cyclic shifts that are independent of X. For any real number t we are interested in $\hat{G}_{CC}(t) = \dfrac{\sum_{i=1}^N I(T(\sigma^i(X)) \ge t)}{N}$ as an estimator of $G_{TU}(t) = p(T(X) > t)$.

Proposition 8.1. *Let* X *be an* $n \times m$ *matrix with rows* $X_{1\cdot}, \ldots, X_{n\cdot}$ *that are iid* $MVN(\mu, \Sigma)$, *where* μ *is a constant vector. If* Σ *is circulant, then* $\hat{G}_{CC}(t)$ *is an unbiased estimator of* $G_{TU}(t)$.

Proof. $E(\hat{G}_{CC}(t)) = E\Big[\dfrac{\sum_{i=1}^N I(T(\sigma^i(X)) \ge t)}{N}\Big] = \dfrac{1}{N}\sum_{i=1}^N E(I(T(\sigma^i(X)) \ge t)) = \dfrac{1}{N}\sum_{i=1}^N p(T(\sigma^i(X)) \ge t) = p(T(\sigma(X)) > t)$. However, since μ is constant and Σ is circulant, Lemma 2 implies that the above expression equals $p(T(X) > t)$, as needed. □

The Instability Selection Model

The *instability selection model*, which was introduced by Newton et al. (1998) and studied by Newton and Lee (2000), is a parsimonious parametric model for allelic loss that accounts for both genetic instability and selection into tumor tissue. It can be used to test for the presence of tumor suppressor genes in binary $n \times m$ matrices X containing LOH data from n tumor samples at a common set of m markers $x_1, \ldots, x_m$ on a chromosome of length 1. Each row $X_{i\cdot}$ of X is called an *allelotype*, and it is assumed that distinct allelotypes are independent and identically distributed.

As noted in Newton and Lee (2000), the distribution of the $X_{i\cdot}$ is determined by four parameters: x, the proposed location of a tumor suppressor; ω, the rate of loss at x; δ, the background rate of loss; and λ, a parameter that governs the rate of transitions between regions of loss and retention. The entries in $X_{i\cdot}$ are modeled using a binary Markov process starting at x and moving towards the telomeres. The transition probabilities are determined by δ, λ, and the distances between the markers x_i. We write the transition probabilities as $p_{ij} = p(x_k = i | x_{k-1} = j)$ for $i, j \in \{0, 1\}$, where 1 represents LOH. If d is the distance between two adjacent markers, the instability-selection model gives $p_{10} = \delta(1 - e^{-\lambda d}), p_{01} = (1 - \delta)(1 - e^{-\lambda d}), p_{00} = 1 - p_{10}$, and $p_{11} = 1 - p_{01}$.

If we view the LOH state at x as a Bernoulli random variable, we can use the conditional probabilities given above to find the unconditional probability of LOH at markers in a given row of a data matrix X. Let $\theta = (x, \omega, \delta, \lambda)$ be the model parameters. The probability of the allelotype $X_{i\cdot}$ is $P_\theta(X_{i\cdot}) = (1 - \omega)P_\theta(X_{i\cdot,left}|L(x) = 0)P_\theta(X_{i\cdot,right}|L(x) = 0) + \omega P_\theta(X_{i\cdot,left}|L(x) = 1)P_\theta(X_{i\cdot,right}|L(x) = 1)$, where $L(x)$ is the loss state at x, and $X_{i\cdot,left}$ and $X_{i\cdot,right}$ denote the loss states at the markers to the left and right of x, respectively. Because of the Markov assumption, the conditional probabilities can be written as products of the transition probabilities between adjacent markers. Moreover, since the rows of X come from unrelated subjects and thus are assumed to be independent, it is possible to compute the probability of an entire data matrix X once we have x, the marker locations, and values for ω, δ, and λ. In all subsequent discussions we assume that $x = x_1$, where x is the presumed location of a tumor suppressor gene.

To illustrate the conditional distributions defined earlier we consider

$$X = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

a matrix of LOH data generated by the instability-selection model. Here we have 10 equally spaced markers on the interval $(0, 1)$, $\lambda = 5$, and either $\omega = \delta = .4$ or $\omega = \delta = .02$. Because there are 10^4 cyclic permutations of X, we can compute the exact true conditional and cyclic conditional distributions. However, the unconditional distributions cannot be evaluated because there are 2^{10^4} binary 4×10 matrices. Tables 8.1 and 8.2 provide the probabilities of each value of $T(X) = \max(S_1, \ldots, S_{10})$ under the true conditional distribution $G_{TC}(t)$ and cyclic conditional distribution $G_{CC}(t)$.

The probabilities given by the cyclic conditional distribution are unaffected by any changes to the parameters in the model, which makes sense in light of the fact that all cyclic shifts are

t	$p_{TC}(T(\tilde{\sigma}(X)) = t \vert \{\sigma(X)\})$	$p_{CC}(T(\tilde{\sigma}(X)) = t \vert \{\sigma(X)\})$
0	0	0
1	.1997	.2
2	.6854	.7
3	.1149	.1
4	0	0

Table 8.1: Probabilities under the True Conditional and Cyclic Conditional Distributions when $\omega = \delta = .4$

t	$p_{TC}(T(\tilde{\sigma}(X)) = t \vert \{\sigma(X)\})$	$p_{CC}(T(\tilde{\sigma}(X)) = t \vert \{\sigma(X)\})$
0	0	0
1	.2230	.2
2	.6803	.7
3	.0967	.1
4	0	0

Table 8.2: Probabilities under the True Conditional and Cyclic Conditional Distributions when $\omega = \delta = .02$

equally likely. Although the probabilities given by the true conditional distribution do change with the parameters, the effect is minor, and in each case they are closely approximated by the probabilities obtained from the cyclic conditional distribution.

As we saw above, for matrices X with iid rows $X_{i\cdot} \sim MVN(\mu, \mathbf{\Sigma})$ the true conditional distribution equals the cyclic conditional distribution if $\mathbf{\Sigma}$ is circulant. We now describe the correlation structure for the instability selection model. Newton and Lee (2000) note that two markers separated by a distance d have correlation $\exp(-\lambda d)$. Under the null hypothesis $\omega = \delta$ it follows that if all markers are equally spaced with common distance d between adjacent markers, then the $X_{i\cdot}$ have an AR(1) correlation structure with correlation $\rho = \exp(-\lambda d)$. Although this appears to be a well-known result in the theory of binary Markov chains, in the Appendix we present an alternate proof that uses techniques from linear algebra.

If X is an $n \times m$ matrix whose rows $X_{i\cdot}$ are independent and generated by the same instability selection model with equally spaced markers, then clearly the column sums also have an autoregressive covariance structure. Our next result describes the covariance structure of X after a applying a random cyclic shift.

Theorem 8.1. *Let X be an $n \times m$ binary matrix whose rows $X_{i\cdot}$ are independent and generated by the instability selection model with equally spaced markers and $\omega = \delta$. If $\sigma(X)$ is a random cyclic shift of X, then the covariance matrix for the column sums of $\sigma(X)$ is circulant.*

Proof. Let $\mathbf{u} = (u_1, u_2, \ldots, u_m)$ be the vector of column sums of X, and write $\tilde{\mathbf{u}} = (\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_m)$ for the vector of column sums of $\tilde{X} = \sigma(X)$. We wish to compute $\mathrm{Cov}(\tilde{u}_i, \tilde{u}_j)$. Because the entries in a given row follow a one-step Markov model, it suffices to find $\mathrm{Cov}(\tilde{u}_1, \tilde{u}_i)$. Now $\mathrm{Cov}(\tilde{u}_1, \tilde{u}_i) = \mathrm{Cov}(\sum_{k=1}^{n} \tilde{x}_{k1}, \sum_{l=1}^{n} \tilde{x}_{li}) = \sum_{k=1}^{n}\sum_{l=1}^{n} \mathrm{Cov}(\tilde{x}_{k1}, \tilde{x}_{li})$, where $\tilde{x}_{ij}$ represents an entry of $\tilde{X}$. However, because the entries in distinct rows are independent the above sum reduces to $\sum_{k=1}^{n} \mathrm{Cov}(\tilde{x}_{k1}, \tilde{x}_{ki})$.

Since all rows behave the same way, we drop the subscript k for ease of notation. $\mathrm{Cov}(\tilde{x}_1, \tilde{x}_i) = E(\tilde{x}_1\tilde{x}_i) - E(\tilde{x}_1)E(\tilde{x}_i) = p(\tilde{x}_1 = 1, \tilde{x}_i = 1) - \delta^2$. Now $p(\tilde{x}_1 = 1, \tilde{x}_i = 1) = p(\tilde{x}_1 = 1, \tilde{x}_i = 1|\text{no break})p(\text{no break}) + p(\tilde{x}_1 = 1, \tilde{x}_i = 1|\text{break})p(\text{break})$, where 'break' represents the event that the cyclic shift introduces a break between columns 1 and i.

When the cyclic shift does not introduce a break between columns 1 and i, Theorem A.1 implies that $\mathrm{Cov}(\tilde{x}_1, \tilde{x}_i) = \delta\Big(1-\delta\Big)\Big(\frac{p_{11}-\delta}{1-\delta}\Big)^{i-1}$. Therefore $p(\tilde{x}_1 = 1, \tilde{x}_i = 1|\text{no break}) = \delta\Big(1-\delta\Big)\Big(\frac{p_{11}-\delta}{1-\delta}\Big)^{i-1} + \delta^2$. Next we consider the case when the cyclic shift does intoduce a break between columns 1 and i. If $(\tilde{x}_1, \ldots, \tilde{x}_m) = \sigma_l(x_1, \ldots, x_m)$, then

$$\tilde{x}_j = \begin{cases} x_{j+l-1} & \text{if } j+l-1 \leq m \\ x_{j+l-m-1} & \text{if } j+l-1 > m \end{cases}$$

Thus $\mathrm{Cov}(\tilde{x}_1, \tilde{x}_i) = \mathrm{Cov}(x_l, x_{l+i-m-1})$. As above, Theorem A.1 implies that $\mathrm{Cov}(x_l, x_{l+i-m-1}) = \delta\Big(1-\delta\Big)\Big(\frac{p_{11}-\delta}{1-\delta}\Big)^{m-i+1}$, because $(x_{l+i-m-1}, \ldots, x_l)$ is a substring of $(x_1, \ldots, x_m)$. Therefore $p(\tilde{x}_1 = 1, \tilde{x}_i = 1|\text{break}) = \delta\Big(1-\delta\Big)\Big(\frac{p_{11}-\delta}{1-\delta}\Big)^{m-i+1} + \delta^2$.

Earlier we noted that

$$\mathrm{Cov}(\tilde{x}_1, \tilde{x}_i) = p(\tilde{x}_1 = 1, \tilde{x}_i = 1|\text{no break})p(\text{no break}) + p(\tilde{x}_1 = 1, \tilde{x}_i = 1|\text{break})p(\text{break}) - \delta^2.$$

Thus if we substitute the values of the conditional probabilities we obtain $\mathrm{Cov}(\tilde{x}_1, \tilde{x}_i) = \delta\Big(1-\delta\Big)\Big[\Big(\frac{p_{11}-\delta}{1-\delta}\Big)^{i-1}\Big(\frac{m-i+1}{m}\Big) + \Big(\frac{p_{11}-\delta}{1-\delta}\Big)^{m-i+1}\big(\frac{i-1}{m}\Big)\Big]$. This simplifies to $\delta\Big(1-\delta\Big)\Big[\exp\Big(-\frac{\lambda(i-1)}{m}\Big)\Big(\frac{m-i+1}{m}\Big) + \exp\Big(-\frac{\lambda(m-i+1)}{m}\Big)\Big(\frac{i-1}{m}\Big)\Big]$. In general the (i,j) entry of the covariance matrix is

$$\begin{aligned} \mathrm{Cov}(u_j, u_i) &= n\delta(1-\delta)\Big[\exp\Big(-\frac{\lambda|i-j|}{m}\Big)\Big(\frac{m-|i-j|}{m}\Big) \\ &+ \exp\Big(-\frac{\lambda(m-|i-j|)}{m}\Big)\Big(\frac{|i-j|}{m}\Big)\Big], \end{aligned}$$

and it is easy to verify that this matrix is circulant. □

In short, the above theorem shows that under the instability selection model with equally spaced markers the covariance of two column sums after cyclic shift is a weighted average of entries of the covariance matrix of the column sums before cyclic shift. It is not hard to see that a corresponding result holds for a AR(1) covariance matrix, or any other banded covariance matrix, for that matter. To illustrate the result, suppose Σ_{AR} is a 5×5 covariance matrix based on an AR(1) correlation structure with correlation ρ and variance 1. Assume X is an $n \times 5$ matrix whose rows $X_{i\cdot} \sim MVN(\mathbf{0}, \Sigma_{\mathrm{AR}})$, and let Σ_{Circ} be the correlation matrix of the column sums of $\sigma(X)$, where σ is a random cyclic shift. Then

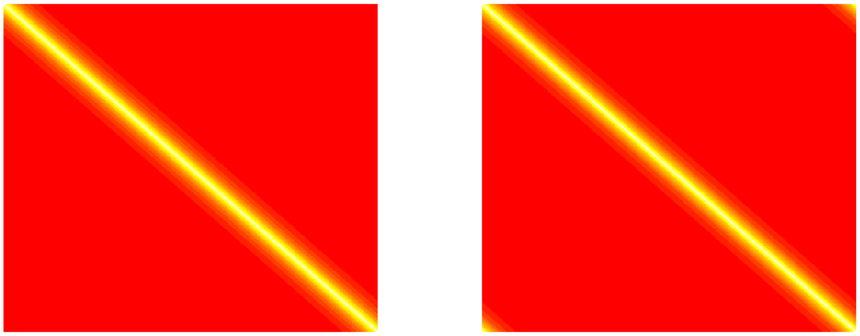

Figure 8.1: Heat Maps for the Correlation Matrices Σ_{AR} (left) and Σ_{Circ} (right)

$$\Sigma_{\text{AR}} = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix} \text{ and } \Sigma_{\text{Circ}} = \begin{pmatrix} 1 & a & b & b & a \\ a & 1 & a & b & b \\ b & a & 1 & a & b \\ b & b & a & 1 & a \\ a & b & b & a & 1 \end{pmatrix},$$

where $a = \dfrac{1}{5}\Big(4\rho + \rho^4\Big)$ and $b = \dfrac{1}{5}\Big(3\rho^2 + 2\rho^3\Big)$.

Earlier we noted that the true conditional and cyclic conditional distributions need not coincide if the covariance matrix Σ is not circulant. However, Figure 8.1 shows heat maps of Σ_{AR} and Σ_{Circ} on the same scale when $m = 250$ and $\rho = .9$. Based on the figures, it appears that the vast majority of the entries in the two matrices are very similar. In fact, the mean value of the entries in $|\Sigma_{\text{AR}} - \Sigma_{\text{Circ}}|$ is .0053, and this mean value decreases to .0013 when $m = 500$ and .00035 when $m = 1000$. The similarity of these matrices suggests that the cyclic conditional distribution may closely approximate the true conditional distribution if Σ is stationary and the number of markers m is large.

Although we do not have any conclusive findings, we now present some preliminary results which suggest the degree to which the cyclic conditional distribution closely approximates the true conditional distribution. Suppose Σ_{AR} is an AR(1) covariance matrix with correlation ρ and variance $\dfrac{1}{1-\rho^2}$. Chan (1997) introduced Σ_{Chan}, a circulant approximation to Σ_{AR}. If X is an $n \times m$ matrix with iid rows $X_{i\cdot} \sim MVN(0, \Sigma)$, we write $p^m_{\text{AR}}(T > t)$ and $p^m_{\text{Chan}}(T > t)$ to denote the probability that $T(X) > t$ under the assumption that Σ is AR(1) or Chan's circulant approximation, respectively. Chan's results imply that for any real number t and for any $\epsilon > 0$ there exists a natural number $m_0(t, \epsilon)$ such that for any $m > m_0(t, \epsilon)$, $|p^m_{\text{AR}}(T > t) - p^m_{\text{Chan}}(T > t)| < \epsilon$.

Theorem 8.1 shows that if X has iid rows $X_{i\cdot} \sim MVN(\mathbf{0}, \Sigma_{\text{AR}})$ and σ is a random cyclic shift, then there exists a circulant matrix Σ_{Circ} such that the column sums of $\sigma(X)$ follow an $MVN(\mathbf{0}, \Sigma_{\text{Circ}})$ distribution. Moreover, the entries of Σ_{Circ} are uniquely defined in terms of the entries in Σ_{AR}. Although Σ_{Chan} is a circulant version of Σ_{AR}, Chan (1997) only defines the entries of $\Sigma_{\text{Chan}}^{-1}$. Our next result gives the entries of Σ_{Chan}.

Lemma 8.3. *Let* $A = \begin{pmatrix} 1+\rho^2 & -\rho & 0 & 0 & \dots & 0 & 0 & -\rho \\ -\rho & 1+\rho^2 & -\rho & 0 & \dots & 0 & 0 & 0 \\ 0 & -\rho & 1+\rho^2 & -\rho & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & -\rho & 1+\rho^2 & -\rho \\ -\rho & 0 & 0 & 0 & \dots & 0 & -\rho & 1+\rho^2 \end{pmatrix}$ *be the* $m \times m$ *matrix defined in equation (5.3) of Chan (1997). If* $B = A^{-1}$*, then the entries of the first row of* B *are*

$$B_{1,j} = \Big(\frac{1}{1-\rho^m}\Big)^2 \Big(\rho^{j-1} \sum_{k=0}^{m-j} \rho^{2k} + \rho^{m-1-j} \sum_{l=1}^{j-1} \rho^{2l}\Big)$$

for $1 \leq j \leq \lceil \frac{m}{2} \rceil$*. Note that the second summand is defined to be zero when* $j = 1$*.*

Proof. Since A is circulant, it follows that $B = A^{-1}$ is also circulant. Thus it suffices to define the elements in the first row of B. However, if the first row of B is $B_{1,1}, B_{1,2}, B_{1,3}, \dots, B_{1,m}$, then $B_{1,2} = B_{1,m}$, $B_{1,3} = B_{1,m-1}$, and so on. Thus we need only define $B_{1,j}$ for $1 \leq j \leq \lceil \frac{m}{2} \rceil$.

To verify that $AB = I$ we must consider the product of the i^{th} row of A and the j^{th} column of B for all i and j. However, since AB is also circulant it suffices to consider the entries in any row or column of AB. We will restrict our attention to the entries in the first column of AB, and here it suffices to consider the dot product of any row of A and the first row of B. Since a circulant covariance matrix equals its transpose, $AB = I$ implies that $BA = B^T A^T = (AB)^T = I^T = I$.

First note that

$$AB_{1,1} = \Big(\frac{1}{1-\rho^m}\Big)^2 \Big[\Big(1+\rho^2\Big)\Big(\sum_{k=0}^{m-1} \rho^{2k}\Big) - 2\rho\Big(\rho \sum_{k=0}^{m-2} \rho^{2k} + \rho^{m-1}\Big)\Big].$$

If we focus our attention on the quantities inside the square brackets we get

$$\sum_{k=0}^{m-1} \rho^{2k} + \sum_{k=1}^{m} \rho^{2k} - 2\rho^2 \sum_{k=0}^{m-2} \rho^{2k} - 2\rho^m = 1 + 2\sum_{k=1}^{m-1} + \rho^{2m} - 2\sum_{k=1}^{m-1} \rho^{2k} - 2\rho^m.$$

However, this simplifies to $(1-\rho^m)^2$. Thus $AB_{1,1} = 1$.

Next consider $AB_{i,1}$ for $i > 1$. We claim that $AB_{i,1} = 0$, so the term $\Big(\dfrac{1}{1-\rho^m}\Big)^2$ will be ignored. As noted above, it suffices to consider the dot product of the i^{th} row of A and the first row of B. The i^{th} row of A has only three non-zero elements, namely in positions $i-1, i$, and $i+1$. If we write $j = i-1$, then there are two cases to consider.

Case 1. The elements $B_{1,j}, B_{1,j+1}$, and $B_{1,j+2}$ are all distinct. In this case the dot product gives

$$(-\rho)\Big[\rho^{j-1} \sum_{k=0}^{m-j} \rho^{2k} + \rho^{m-1-j} \sum_{k=l}^{j-1} \rho^{2l}\Big] + (1+\rho^2)\Big[\rho^j \sum_{k=0}^{m-j-1} \rho^{2k} + \rho^{m-2-j} \sum_{l=1}^{j} \rho^{2l}\Big]$$

$$-\rho\Big[\rho^{j+1}\sum_{k=0}^{m-j-2}\rho^{2k}+\rho^{m-3-j}\sum_{l=1}^{j+1}\rho^{2l}\Big].$$

Rewriting gives

$$-\rho^{j}\sum_{k=0}^{m-j}\rho^{2k}-\rho^{m-j}\sum_{l=1}^{j-1}\rho^{2l}+\rho^{j}\sum_{k=0}^{m-j-1}\rho^{2k}+\rho^{m-j}\sum_{l=0}^{j-1}\rho^{2l}+$$

$$\rho^{j}\sum_{k=1}^{m-j}\rho^{2k}+\rho^{m-j}\sum_{l=1}^{j}\rho^{2l}-\rho^{j}\sum_{k=1}^{m-j-1}\rho^{2k}+\rho^{m-j}\sum_{l=0}^{j}\rho^{2l}.$$

Combining terms involving either ρ^j or ρ^{m-j} shows that the above expression equals 0.
Case 2. The elements $B_{1,j}, B_{1,j+1}$, and $B_{1,j+2}$ are not all distinct. Here we need to consider two subcases.
Subcase A. m is even, so the repeated elements are $B_{1,j} = B_{1,j+2}$ for $j = \dfrac{m}{2}$. Thus the dot product equals

$$(-2\rho)\Big[\rho^{\frac{m}{2}-1}\sum_{k=0}^{m-\frac{m}{2}}\rho^{2k}+\rho^{m-1-\frac{m}{2}}\sum_{l=1}^{\frac{m}{2}-1}\rho^{2l}\Big]+(1+\rho^2)\Big[\rho^{\frac{m}{2}}\sum_{k=0}^{m-\frac{m}{2}-1}\rho^{2k}+\rho^{m-2-\frac{n}{2}}\sum_{l=1}^{\frac{m}{2}}\rho^{2l}\Big].$$

Combining terms and simplifying as in Case 1 shows that this expression equals 0.
Subcase B. m is odd, in which case the repeated elements are either $B_{1,j} = B_{1,j+1}$ or $B_{1,j+1} = B_{1,j+2}$. Both cases are identical, so we consider $B_{1,j} = B_{1,j+1}$ for $j = \dfrac{m+1}{2}$. Here the dot product equals

$$(1-\rho+\rho^2)\Big[\rho^{\frac{m+1}{2}-1}\sum_{k=0}^{m-\frac{m+1}{2}}\rho^{2k}+\rho^{m-1-\frac{m+1}{2}}\sum_{l=1}^{\frac{m+1}{2}-1}\rho^{2l}\Big]+$$

$$(-\rho)\Big[\rho^{\frac{m+1}{2}-2}\sum_{k=0}^{m-(\frac{m+1}{2}-1)}\rho^{2k}+\rho^{m-\frac{m+1}{2}}\sum_{l=1}^{\frac{m+1}{2}-2}\rho^{2l}\Big]$$

Using the same ideas as above, we see that this simplifies to 0. □

The above lemma shows that the entries of Σ_{Chan} and Σ_{Circ} are not identical. We write Σ^m_{Chan} and Σ^m_{Circ} when both matrices are derived from an $m \times m$ covariance matrix Σ^m_{AR}, and suppose Σ^m_{AR} has correlation ρ and variance 1. For $1 \leq j \leq \lceil \frac{m}{2} \rceil$ the elements of Σ^m_{Chan} and Σ^m_{Circ} are

$$(\Sigma^m_{\text{Chan}})_{1j}=\Big(\frac{1}{1-\rho^m}\Big)^2\Big(\rho^{j-1}(1-\rho^{2(m+1-j)})+\rho^{m+1-j}(1-\rho^{2(j-1)})\Big)$$

and

$$(\Sigma^m_{\text{Circ}})_{1,j}=\Big(\frac{m+1-j}{m}\Big)\rho^{j-1}+\Big(\frac{j-1}{m}\Big)\rho^{m+1-j}.$$

Therefore

$$
\begin{aligned}
(\Sigma^m_{\text{Chan}})_{1,j} - (\Sigma^m_{\text{Circ}})_{1,j} &= \rho^{j-1}\Big[\Big(\frac{1}{1-\rho^m}\Big)^2\Big(1-\rho^{2(m+1-j)}\Big) - \frac{m+1-j}{m}\Big] \\
&+ \rho^{m+1-j}\Big[\Big(\frac{1}{1-\rho^m}\Big)^2\Big(1-\rho^{2(j-1)}\Big) - \frac{j-1}{m}\Big].
\end{aligned}
$$

If ρ is fixed, then there exists a natural number $m_0(\epsilon)$ such that for $m > m_0(\epsilon)$ it follows that $|(\Sigma^m_{\text{Chan}})_{1j} - (\Sigma^m_{\text{Circ}})_{1j}| < \epsilon$. Our hope is that this result can be used to show that $p^m_{\text{Chan}}(T > t)$ and $p^m_{\text{Circ}}(T > t)$ become arbitrarily close as $m \to \infty$. If so, Chan's result would imply that the same conclusion holds for $p^m_{\text{AR}}(T > t)$ and $p^m_{\text{Circ}}(T > t)$. It would then follow that $\hat{G}_{CC}(t)$ is an asymptotically unbiased estimator of $G_{TU}(t)$ when Σ has an AR(1) correlation structure.

The above results would imply that true unconditional and cyclic unconditional distributions are asymptotically equivalent when the rows of X are iid multivariate normal with mean $\mathbf{0}$ and have a AR(1) covariance matrix Σ. In practice, however, DiNAMIC conditions on the observed data. Thus we are also interested in whether corresponding results hold for the true conditional and the cyclic conditional distributions. At the moment it is not clear what can be said about these conditional distributions, or if the marginal results described above can be brought to bear.

Appendix

Theorem A.1. *Let* $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ *be a binary vector whose entries are generated by the instability selection model with equally spaced markers and* $\omega = \delta$*. The correlation matrix* $\mathrm{Corr}(\mathbf{x})$ *is autoregressive with common ratio* $\rho = \dfrac{p_{11} - \delta}{1 - \delta}$.

Proof. We begin by noting that all of the transition probabilities may be written in terms of p_{11}, namely $p_{01} = 1 - p_{11}$, $p_{10} = \frac{\delta}{1-\delta} p_{01} = \frac{\delta}{1-\delta}(1 - p_{11})$, and $p_{00} = 1 - p_{10} = 1 - (\frac{\delta}{1-\delta})(1 - p_{11})$. This rewriting will be used later.

Clearly $E(x_1) = \delta$, and from this it follows that $E(x_2) = p(x_2 = 1) = p(x_2 = 1|x_1 = 1)p(x_1 = 1) + p(x_2 = 1|x_1 = 0)p(x_1 = 0) = p_{11}\delta + p_{10}(1 - \delta) = \delta$. This in turn implies that $E(x_k) = \delta$ for $1 \leq k \leq m$. Since x_k only assumes the values 0 and 1, we also note that $\mathrm{Var}(x_k) = E(x_k^2) - E(x_k)^2 = E(x_k) - E(x_k)^2 = \delta - \delta^2$ for $1 \leq k \leq n$.

Next we wish to consider $\mathrm{Cov}(x_j, x_k)$ for $j < k$. If we appeal to the definition and the comments in the previous paragraph, we may write the covariance as $E(x_j x_k) - \delta^2$. Because we have a one-step Markov model, it suffices to find a find a formula for $E(x_1 x_k)$. Now $E(x_1 x_k) = p(x_1 = 1, x_k = 1)$. When $k = 2$ we have $p(x_1 = 1, x_2 = 1) = p_{11}\delta$; if $k = 3$ we have $p(x_1 = 1, x_3 = 1) = p(x_1 = 1, x_2 = 1, x_3 = 1) + p(x_1 = 1, x_2 = 0, x_3 = 1) = p_{11}^2\delta + p_{10}p_{01}\delta$. Thus to compute $p(x_1 = 1, x_k = 1)$ we must find the sum of the probabilities of all binary strings of length k whose first and last entries are 1. From now on we assume that all binary strings have this form, which we write $1x_2 \ldots x_{k-1}1$.

A new binary string $1x_2 \ldots x_{k-1}x_k1$ of length $k+1$ can be obtained from an existing binary string $1x_2 \ldots x_{k-1}1$ of length k by adding a 0 or 1 as the new penultimate digit x_k. Thus any string of length k produces exactly two strings of length $k + 1$. Suppose $x_{k-1} = 0$. Then under the instability selection model $p(1x_2 \ldots x_{k-1}1) = zp_{10}$, where $z = p(1x_2 \ldots x_{k-1})$; this probability would be zp_{11} if $x_{k-1} = 1$. For convenience, suppose $x_{k-1} = 0$. If the $x_k = 0$, then $p(1x_2 \ldots x_{k-1}x_k1) = zp_{00}p_{10}$; this probability is $zp_{10}p_{11}$ if $x_k = 1$. Thus a string of length k whose probability is zp_{10} yields two strings of length $k + 1$, and the sum of their probabilities is $zp_{00}p_{10} + zp_{10}p_{11}$. Similarly it can be shown that a string of length k whose probability is zp_{11} yields two strings of length $k + 1$ whose probabilities sum to $zp_{01}p_{10} + zp_{11}p_{11}$. If we write all of the above expressions in terms of p_{11} we see that

$$z\Big(\frac{\delta}{1-\delta}\Big)\Big(1 - p_{11}\Big) \text{ yields } z\Big[1 - \frac{\delta}{1-\delta}\Big(1 - p_{11}\Big)\Big]\Big(\frac{\delta}{1-\delta}\Big)\Big(1 - p_{11}\Big)$$
$$+z\Big(\frac{\delta}{1-\delta}\Big)\Big[1 - p_{11}\Big]p_{11},$$

and

$$zp_{11} \text{ yields } z\Big[\frac{\delta}{1-\delta}\Big(1-p_{11}\Big)\Big]\Big(1-p_{11}\Big) + z\Big[p_{11}\Big]p_{11}.$$

The notation and lack of simplification in the above expressions are intentional, and if we drop the common terms $z\Big(\dfrac{\delta}{1-\delta}\Big)$ in the top equation and z in the bottom we see the following transitions when we examine the probabilities that arise when we consider strings of length k and $k+1$:

$$(1-p_{11}) \text{ yields } \Big[1-\Big(\frac{\delta}{1-\delta}\Big)\Big(1-p_{11}\Big)\Big]\Big(1-p_{11}\Big) + \Big[1-p_{11}\Big]p_{11},$$

and

$$p_{11} \text{ yields } \Big[\Big(\frac{\delta}{1-\delta}\Big)\Big(1-p_{11}\Big)\Big]\Big(1-p_{11}\Big) + \Big[p_{11}\Big]p_{11},$$

If we let $a = \Big(\dfrac{\delta}{1-\delta}\Big)\Big(1-p_{11}\Big)$ and $b = p_{11}$ for the quantities inside the first and second square brackets, respectively, then we may write these expressions more concisely as

$$(1-p_{11}) \text{ yields } [1-a](1-p_{11}) + [1-b]p_{11}$$

and

$$p_{11} \text{ yields } [a](1-p_{11}) + [b]p_{11}$$

Here the expressions involving a or b in square brackets can be viewed as coefficients of the terms $(1-p_{11})$ and p_{11}.

We now wish to show how to obtain $E(x_1x_{k+1})$ from $E(x_1x_k)$. If we combine all binary strings $1x_2\ldots x_{k-1}1$ in which $x_{k-1}=0$ and all binary strings $1x_2\ldots x_{k-1}1$ in which $x_{k-1}=1$, then we may write $E(x_1x_k) = \alpha p_{10} + \beta p_{11}$. However, rewriting in terms of p_{11} gives

$$E(x_1x_k) = \Big[\alpha\Big(\frac{\delta}{1-\delta}\Big)\Big]\Big(1-p_{11}\Big) + \Big[\beta\Big]p_{11} = [\alpha'](1-p_{11}) + [\beta]p_{11}.$$

The comments in the previous paragraph imply that

$$E(x_1x_{k+1}) = [\alpha']([1-a](1-p_{11}) + [1-b]p_{11}) + [\beta]([a](1-p_{11}) + [b]p_{11}).$$

Combining terms gives

$$[\alpha'(1-a) + \beta a](1-p_{11}) + [\alpha'(1-b) + \beta b]p_{11} = \begin{pmatrix} p_{11} & (1-p_{11}) \end{pmatrix} \begin{pmatrix} b & (1-b) \\ a & (1-a) \end{pmatrix} \begin{pmatrix} \beta \\ \alpha' \end{pmatrix}.$$

Thus by using the matrix $M = \begin{pmatrix} b & (1-b) \\ a & (1-a) \end{pmatrix}$ we can write $E(x_1x_{k+1}) = \mathbf{v}^T M\mathbf{w}$, where $\mathbf{w} = \begin{pmatrix} \beta \\ \alpha' \end{pmatrix}$ is the vector of coefficients of p_{11} and $(1-p_{11})$ in $E(x_1x_k)$ and $\mathbf{v} = \begin{pmatrix} p_{11} \\ 1-p_{11} \end{pmatrix}$.

Using the notation given above, $E(x_1x_2) = \delta p_{11} = \delta\mathbf{v}^T\mathbf{e}_1 = \delta\mathbf{v}^T M^0\mathbf{e}_1$, where $\mathbf{e}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$

and M^0 is the identity matrix. Therefore the argument in the previous paragraph implies that when $s \geq 2$ we have $E(x_1 x_s) = \delta \mathbf{v}^T M^{s-2} \mathbf{e}_1$. The matrix M is diagonalizable when $a \neq b$, or equivalently $\lambda d < \infty$. $M = \begin{pmatrix} 1 & 1 \\ 1 & \frac{a}{b-1} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & b-a \end{pmatrix} \begin{pmatrix} \frac{a}{a-b+1} & \frac{1-b}{a-b+1} \\ \frac{1-b}{a-b+1} & \frac{b-1}{a-b+1} \end{pmatrix}$. Rewriting in terms of δ and p_{11} gives $M = \begin{pmatrix} 1 & 1 \\ 1 & -\frac{\delta}{1-\delta} \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & \frac{p_{11}-\delta}{1-\delta} \end{pmatrix} \begin{pmatrix} \delta & 1-\delta \\ 1-\delta & \delta-1 \end{pmatrix}$. Therefore $E(x_1 x_s) = \delta \Big[p_{11} \Big(\frac{p_{11}-\delta}{1-\delta} \Big)^{s-2} - \delta \Big(\frac{p_{11}-\delta}{1-\delta} \Big)^{s-2} + \delta \Big]$, and hence we conclude that $\mathrm{Corr}(x_1, x_s) = \frac{\mathrm{Cov}(x_1, x_s)}{\sqrt{\mathrm{Var}(x_1)\mathrm{Var}(x_s)}} = \frac{E(x_1 x_s) - \delta^2}{\delta(1-\delta)} = \Big(\frac{p_{11}-\delta}{1-\delta} \Big)^{s-1}$. □

Gain Marker	DiNAMIC	GISTIC	Loss Marker	DiNAMIC	GISTIC
1p36	X		1p21	X	
1q23	X		1p36	X	
1q25	X		1q31	X	
1q32	X		1q41	X	
1q43	X		1q43	X	
2p25	X	X	2q33	X	
2q14	X	X	3p12	X	X
2q37	X	X	3q13	X	
3p26	X		4p14	X	
3p25		X	4p15	X	
3q23	X	X	4q12	X	
3q27	X	X	4q28	X	
4p15	X		4q34		X
4p16	X	X	5p14	X	
4q35		X	5q15	X	
5p15	X	X	5q33	X	
5q31	X	X	6p12	X	
5q35	X	X	6p21	X	
6p21	X	X	6q24	X	X
7p22	X	X	7p11	X	
7p11	X	X	8p21	X	
7q11	X		8p23	X	X
7q36	X	X	8q13	X	
8p21	X	X	9p24		X
8p23	X	X	9p21	X	X
8q24	X	X	10p12	X	
9p22		X	10q11		X
9q34	X	X	10q21	X	X
10p15		X	10q26	X	X
10q22	X	X	11p11	X	
10q26		X	11p14	X	
11q13	X	X	11p15	X	X
11q23	X	X	11q24	X	X
11q25	X	X	12p13	X	X

Table A1. Locations of Significant Markers in the Glioma Dataset of Kotliarov et al. (2006), as Determined by DiNAMIC's Cyclic Shift Procedure and GISTIC

Gain Marker	DiNAMIC	GISTIC
12p13	X	X
12q14	X	X
12q15	X	X
12q24	X	X
13q33	X	X
14q11	X	X
14q32	X	X
15q12		X
16p13		X
16p12	X	
16q24	X	X
17p13		X
17p12	X	
17q12	X	
17q25	X	X
18q23	X	X
19p13	X	X
19q12	X	
20p13	X	X
20q13	X	X
21q22	X	X
22q11	X	X

Loss Marker	DiNAMIC	GISTIC
12q13	X	
12q21	X	
13q21	X	X
13q22	X	
13q33	X	
14q11	X	
14q12	X	X
14q24		X
14q31	X	
15q11	X	X
18q22	X	X
19q13	X	X
21q21	X	X
22q13		X

Table A1. Locations of Significant Markers in the Glioma Dataset of Kotliarov et al. (2006), as Determined by DiNAMIC's Cyclic Shift Procedure and GISTIC

Bibliography

Albertson D.G., Collins C., McCormick F., Gray J.W., 2003. Chromosome aberrations in cancer. *Nat Genet* **34**(4): 369 - 376.

Anderson, T.W. (1960) Some stochastic process methods for intelligence test scores. In *Mathematical Methods in teh Social Sciences, 1959: Proceedings from the First Stanford Symposium* (eds. K.J. Arrow, S. Karlin, P. Suppes). Stanford Mathematical Studies in the Social Sciences: IV, p. 205 - 220. Stanford University Press, Stanford, CA.

Baross, A., *et al.* (2007) Assessment of algorithms for high throughput detection of genomic copy number variation in oligonucleotide microarray data. *BMC Bioinformatics* **8**: 368 doi:10.1186/1471-2105-8-368.

Benjamini, Y. and Yekutieli, D. (2001) The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**(4): 1165 - 1188.

Beroukhim, R., *et al.* (2007) Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proc. Nat. Acad. Sci.* **104**: 20007 - 20012.

Carlson, S., *et al.* (1998) Expression of SET, an inhibitor of protein phosphatase 2A, in renal development and Wilms tumor. *J. Am. Soc. Nephrology* **9**: 1873 - 1880.

Chan, K.S. (1997) On the validity of the method of surrogate data. In *Nonlinear dynamics and time series* (eds. C.D. Cutler and D.T. Kaplan), Fields Institute Communications, vol. 11, 77–97. American Mathematical Society, Providence, RI.

Coe, B.P., *et al.* (2007) Resolving the resolution of array CGH. *Genomics* **89**: 647 - 653.

Davies, J.J., Wilson, I.M., and Lam, W.L., (2005) Array CGH technologies and their applications to cancer genomes. *Chromosome Res.* **13**: 237 - 248.

Diskin, S., *et al.* (2006) STAC: A method for testing the significance of DNA copy number aberrations across multiple array-CGH experiments. *Genome Res.* **16**: 1149 - 1158.

Diskin, S., *et al.* (2008) Adjustment of genomic waves in signal intensities from whole-genome SNP genotyping platforms. *Nucleic Acids Res.* **36**: e126 doi:10.1093/nar/gkn556.

Guttman, M., *et al.* (2007) Assessing the significance of conserved genomic aberrations using high resolution genomic microarrays. *PLoS Genetics* **3**: e143 doi:10.1371/journal.pgen.0030143.

Harada, T., *et al.* (2008) Genome-wide DNA copy number analysis in pancreatic cancer using high-density single nucleotide polymorphism arrays. *Oncogene* **27**: 1951 - 1960.

Heimberger, A.M., *et al.* (2005) The natural history of EGFR and EGFRvIII in glioblastoma patients. *J. Translational Medicine* **3**: 38 doi:10.1186/1479-5876-3-38.

Hupe, P., *et al.* (2004) Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics* **20**: 3413 - 3422.

Iafrate, A.J., *et al.* (2004) Detection of large-scale variation in the human genome. *Nat. Genet.* **36**(9): 949 - 951.

Itsara, A., *et al.* (2009) Population analysis of large copy number variants and hotspots of human genetic disease. *Am. J. Human Genetics* **84**: 148 - 161.

Jackson, M.A., *et al.* (2006) Genetic Alterations in Cancer Knowledge System: Analysis of gene mutations in mouse and human liver and lung tumors. *Toxicological Sci.* **90**(2): 400 - 418.

Joanes, D.N. and Gill, C.A. (1998) Comparing measures of sample skewness and kurtosis. *The Statistician* **47** Part I: 183 - 189.

Kallioniemi, A., *et al.* (1992) Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. *Science* **258**: 818 - 821.

Kent, W.J., *et al.* (2002) The human genome browser at UCSC. *Genome Res.* **12**(6): 996 - 1006.

Klijn, C., *et al.* (2008) Identification of cancer genes using a statistical framework for multiexperiment analysis of nondiscretized array CGH data. *Nucleic Acids Res.* **36**: e13 doi:10.1093/nar/gkm1143.

Knudsen A, 1971. Mutations and cancer: a statistical study of retinoblastoma. *Proc Nat Acad Sci* **78**(4): 820 - 823.

Komura, D., *et al.* (2006) Genome-wide detection of human copy number variations using high-density DNA oligonucleotide arrays. *Genome Res.* **16**: 1575 - 1584.

Kotliarov, Y., *et al.* (2006) High-resolution global genomic survey of 178 gliomas reveals novel regions of copy number alteration and allelic imbalances. *Cancer Res.* **66**: 9428 - 9436.

Kresse, S., *et al.* (2008) DNA copy number changes in high-grade malignant peripheral nerve sheath tumors by array CGH. *Mol. Cancer* **7**: 48 doi:10.1186/1476-4598-7-48.

Lander, E.S. and Botstein, D. (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121**: 185 - 199.

Lucito, R., *et al.* (2003) Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res.* **13**: 2291 - 2305.

Mangin, B., Goffinet, B., and Rebaï, A. (1994) Constructing confidence intervals for QTL location. *Genetics* **138**: 1301 - 1308.

Marioni, J., *et al* (2007) Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Bio.* **8**: R228 doi:10.1186/gb-2007-8-10-r228.

Miller, B.J., *et al.* (2003) Pooled analysis of loss of heterozygosity in breast cancer: a genome scan provides evidence for multiple tumor suppressors and identifies novel candidate regions. *Am. J. Human Genet.* **73**: 748 - 767.

Mitelman, F., Johansson, B., and Mertens, F. (eds) (2010) Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer, http://cgap.nci.nih.gov/Chromosomes/Mitelman.

Myllykangas, S. and Knuutila, S. (2006) Manifestation, mechanisms and mysteries of gene amplifications. *Cancer Let.* **232**: 79 - 89.

Natrajan, R., *et al.* (2006) Array CGH profiling of favourable histology Wilms tumours reveals novel gains and losses associated with relapse. *J. Pathology* **210**: 49 - 58.

Newton, M.A., *et al.* (1998) On the statistical analysis of allelic loss data. *Statist. Med* **17**: 1425 - 1445.

Newton, M.A., and Lee, Y. (2000) Inferring the location and effect of tumor suppressor genes by instability-selection modeling of allelic-loss data. *Biometrics* **56**: 1088 - 1097.

Olshen, A., *et al.* (2004) Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**: 557 - 572.

Pinkel, D., *et al.* (1998) High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* **20**: 207 - 211.

Rahman, N., *et al.* (1996) Evidence for a familial Wilms' tumour gene (FWT1) on 17q12-21. *Nature Genet.* **13**: 461 - 463.

Redon, R., *et al.* (2006) Global variation in copy number in the human genome. *Nature* **444**: 444 - 454.

Rueda, O. and Diaz-Uriarte, R. (2008) Finding recurrent regions of copy number variation: a review. *COBRA Preprint Series*: Paper 42.

Shah, S. (2008) Computational methods for identification of recurrent copy number alteration patterns by array CGH. *Cytogenet. and Genome Res.* **123**: 343-351.

Simes, R.J. (1986) An improved Bonferroni procedure for multiple tests of significance. *Biometrika* **73**(3): 751 - 754.

Sterrett, A. and Wright, F.A., (2007) Inferring the location of tumor suppressor genes by modeling the frequency of allelic loss. *Biometrics* **63**: 33 - 40.

Stratchan, T. and Read, A.P. (1999) *Human Molecular Genetics*, 2^{nd} edition. Wiley-Liss Publishing, New York, NY.

Sun, W., *et al.* (2009) Integrated study of copy number states and genotype calls using high-density SNP arrays. *Nucleic Acids Res.* **37**(16): 5365 - 5377.

van de Wiel, M. and van Wieringen, W. (2007) CGHregions: dimension reduction for array CGH data with minimal information loss. *Cancer Res.* **3**: 55 - 63.

van't Veer, L.J., *et al.* (2000) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**(31): 530 - 536.

Veerhaak, R.G.W., *et al.* (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NH1. *Cancer Cell* **17**: 98 - 110.

Venkatraman, E. and Olshen, A. (2007) A faster circular binary segmentation algorithm for the analysis of aCGH data. *Bioinformatics* **23**: 657-663.

Visscher, P.M., Thompson, R., and Haley C.S. (1996) Confidence intervals in QTL mapping by bootstrapping. *Genetics* **143**: 1013 - 1020.

Weir, B.A., *et al.* (2007) Characterizing the cancer genome in lung adenocarcinoma. *Nature* **450**(6): 893 - 898.

Westfall, P. and Young, S. (1993) *Resampling-based Multiple Testing.* Wiley-Interscience. New York.

Zhao, X., *et al.* (2004) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. *Cancer Res.* **64**: 3060 - 3071.