False Discovery Rates and Copy Number Variation

Bradley Efron* Nancy Zhang†

Department of Statistics
Stanford University

Abstract

Copy number changes, the gains and losses of chromosome segments, are a common type of genetic variation among healthy individuals as well as an important feature in tumor genomes. Microarray technology enables us to simultaneously measure, with moderate accuracy, copy number variation at more than a million chromosome locations and for hundreds of subjects. This leads to massive data sets and complicated inference problems concerning which locations for which subjects are genuinely variable. In this paper we consider a relatively simple false discovery rate approach to cnv analysis. More careful parametric change-point methods can then be focused on promising regions of the genome.

Key words and phrases: False discovery rate, multiple testing, grouped hypotheses, DNA copy number

1 Introduction

Basic genetics says that we have two copies of each bit of chromosomal information. In fact, however, even healthy individuals show occasional variations, displaying stretches of the genome having more or less than two copies. Within the past decade, significant advances in microarray technology have enabled the genome-wide fine scale measurement of DNA copy number in high throughput fashion; see Bignell et al. (2004); Peiffer et al. (2006); Pinkel et al. (1998); Pollack et al. (1999); Snijders et al. (2001). This has led to large-scale studies investigating the role of DNA copy number changes in human disease and phenotypic variation. The studies fall into two main categories: changes in DNA copy number can occur as a form of inherited genetic polymorphism in normal human DNA. They can also accompany somatic mutation, as often observed in cancerous tumors. Inherited copy number changes in normal samples have been called copy number variants (cnv), while those that occur in tumors have been referred to as copy number aberrations (cna), to distinguish the fact that they are "aberrant" forms which do not occur as population-wide variation. This paper discusses a false discovery rate approach to the analysis of DNA copy number data.

The statistical properties of copy number data are quite different in the two cases. In normal samples, the copy number variants, most of which are inherited, are usually short and spaced far apart, whereas in tumor samples, cnas can be quite long, sometimes spanning entire chromosomes. The false discovery rate (fdr) methodology developed in this paper applies to both situations, but our examples will start by focusing on the first. We will discuss tumor samples in more detail, with a data example, in Section 7.

*Research supported in part by NIH grant 8R01 EB002784 and by NSF grant DMS 0804324.
†Research supported in part by NSF Grant DMS 0906394.

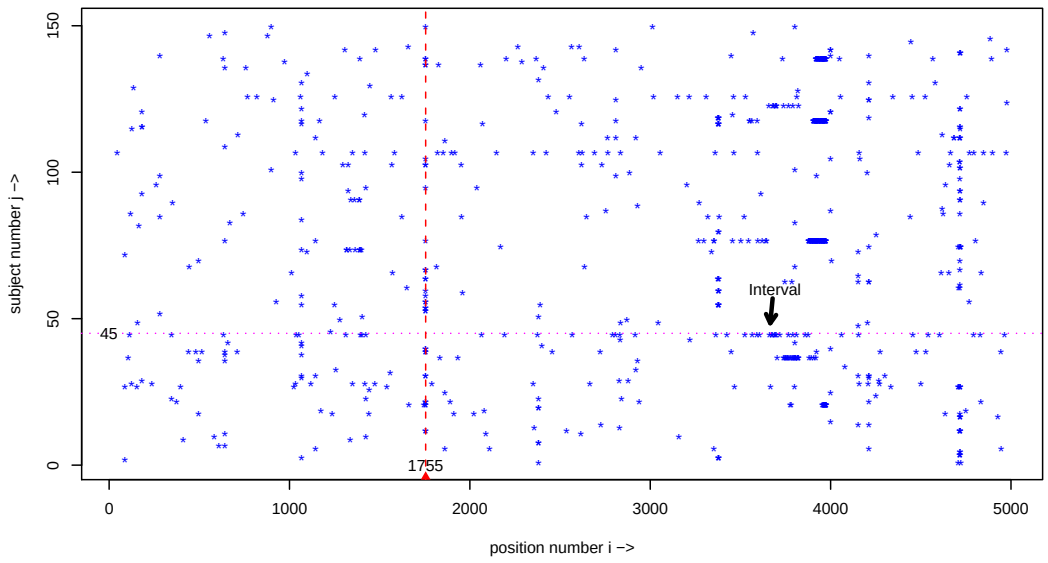

Figure 1: CNV data for 150 healthy subjects measured at 5000 marker positions. Points indicate positions (i, j) with measurement x_{ij} (1.1) in the most negative 1/10 of 1%, perhaps indicating copy numbers less than 2. Subject 45 has a long interval of points around position 3800. The marker at position 1755 seems prone to copy numbers < 2. (Corresponding map for positive x_{ij} values shows less structure.) The matrix has been transposed for convenient display.

Figure 1 concerns a cnv data set we will use to illustrate our methodology. Here $n = 150$ healthy subjects have been each assessed at $N = 5000$ marker positions, yielding cnv measurements

$$x_{ij} = \begin{cases} i = 1, 2, \ldots, N = 5000 \text{ positions} \\ j = 1, 2, \ldots, n = 150 \text{ subjects.} \end{cases} \tag{1.1}$$

Roughly speaking, values of x_{ij} much less than zero indicate *less* than two copies, and values much greater than zero *more* than two copies, but there is considerable measurement error. The histogram of all 750,000 x_{ij} is smoothly unimodal, mean and standard deviation -0.018 and 0.188, showing moderate skewness toward the negative side, with a small percentage of extreme outliers in both directions.

The points in Figure 1 (where the matrix has been transposed) indicate the 750 most negative x_{ij} values, in other words the one-tenth of one percent of (position, subject) combinations giving strongest evidence of copy numbers less than 2. We can see that subject 45, for example, seems to have a long interval of decreased copy numbers around position 3800, while position 1755 might be prone to copy number reductions. A typical question we would like to answer is whether position 1755 is genuinely cnv-prone. A key feature of cnv problems is the availability of information in both directions. Does subject 45 have less than two copies of the marker at position 1755? The methodology introduced in Section 2 combines the horizontal and vertical features in Figure 1 to answer such questions.

Let $\bar{x}_{ij}$ indicate a moving average of the x_{ij} values for subject j,

$$\bar{x}_{ij} = \sum_{i'=i-m}^{i+m} x_{ij} \Big/ (2m+1) \tag{1.2}$$

for some fixed value of m (with obvious modifications for i near 1 or N). Because cnv intervals tend to span a contiguous range of marker positions, $\bar{x}_{ij}$ will be less noisy than x_{ij}; see Section 5 for the specific calculation. It is also helpful to standardize the columns of the $\{\bar{x}_{ij}\}$ matrix, that is, each *subject's* $\bar{x}_{ij}$ values, by defining

$$z_{ij} = (\bar{x}_{ij} - \hat{a}_j) \,/\hat{b}_j \tag{1.3}$$

where $\hat{a}_j$ and $\hat{b}_j$ are the median and robust standard deviation (one-half the distance between the 16th and 84th percentiles) of $\{\bar{x}_{ij} : i = 1, 2, \ldots, N\}$. Most of our numerical examples will be based on z_{ij} values (1.2), (1.3) with $m = 5$. The application of fdr methodology to the z-values renders copy number variations far more visible; see Figure 3.

The paper develops as follows: an iterative algorithm is introduced in Section 2, in which a local false discovery rate estimate (Efron, 2008) is first fit to the combined data, and then modified to take account of differing cnv probabilities at the various positions i. This gives an fdr estimate for each position and subject, as well as an estimate $\hat{k}_i$ of the number of subjects carrying a cnv at position i.

Section 3 and Section 4 develop hypothesis-testing and estimation methods based on the $\hat{k}_i$'s, aimed at answering the question of which, if any, of the positions are cnv-prone. The iterative algorithm is examined more closely in Section 5 and Section 6, and connected to maximum likelihood theory. Section Section 7 examines in more detail the problem of detecting cna-prone regions in tumors. Having located positions prone to copy number changes based on the $\hat{k}_i$ estimates, Section 7 then discusses local change-point methods intended to say which of the subjects are the affected ones, the so-called "carriers."

There are by now many different methods for single-sample analysis of DNA copy number. These methods process each sample (i.e., each column in the matrix (1.1)) separately, as if the method has never seen a similar sample before, and will never see another sample again. Reviews of single-sample methods are given in Lai et al. (2005), Willenbrock and Fridlyand (2005), and Zhang (2010). For both single-sample analysis and the simultaneous processing of multiple samples, global change-point tests, scanning over the entire range of positions, have played a central role in the statistical cnv lierature (Olshen et al., 2004; Siegmund, Yakir and Zhang, 2010; Zhang et al., 2010). The literature leans heavily on Gaussian process theory, and within that realm produces impressively precise testing algorithms. Wang et al. (2005) propose an fdr approach, closer to the methods proposed here.

The identification of cnv-prone regions across a cohort of tumor samples has been a problem of increased scientific interest. Most published methods (Beroukhim et al., 2007; Diskin et al., 2006; Guttman et al., 2007; Newton et al., 1998; Newton and Lee, 2000; Rouveirol et al., 2006; Taylor et al., 2008) take a *post-segmentation* approach: each sample is first segmented individually, which reduces them to piece-wise constant sequences indicating regions of amplification, deletion, or normal copy number. Then the samples are aligned, and a statistical model (Newton et al., 1998; Newton and Lee, 2000) or permutation-based approach (Diskin et al., 2006) is used to identify regions of highly recurrent aberration. These post-segmentation approaches rely on the vagaries of the underlying segmentation model. After segmentation, how evidence for gains and losses should be combined across samples is still much debated. Existing strategies range from counting the number of carriers, without weighting by the strength of evidence of each carrier (e.g., Diskin et al. (2006)), to the "G-score" (Beroukhim et al., 2007), defined as the number of carriers times the average amplitude of the signal among carriers. The fdr-based approach that we describe arises from a natural likelihood model, is simple and computationally

fast, and yields biologically meaningful results.

2 False Discovery Rate Methods

Forgetting about cnv structure for a moment, suppose we have M null hypotheses $H_{01}, H_{02}, \ldots, H_{0M}$ to test, based on possibly correlated test statistics $z_1, z_2, \ldots, z_M$. False discovery rate methods can be motivated by the Bayesian *two-groups model* discussed at length in Efron (2008), in which each case is either null or non-null with prior probability π_0 or $\pi_1 = 1 - \pi_0$, and with the z values having density either $f_0(z)$ or $f_1(z)$,

$$\begin{aligned} \pi_0 &= \Pr\{\text{null}\} & f_0(z) &= \text{density if null} \\ \pi_1 &= \Pr\{\text{non-null}\} & f_1(z) &= \text{density if non-null.} \end{aligned} \tag{2.1}$$

Bayes rule shows that the posterior probability of "null" given z, the *local false discovery rate*, is

$$\text{fdr}(z) = \Pr\{\text{null}|z\} = \pi_0 f_0(z)/f(z) \tag{2.2}$$

where $f(z)$ is the mixture density

$$f(z) = \pi_0 f_0(z) + \pi_1 f_1(z). \tag{2.3}$$

An empirical Bayes approach to multiple testing uses the entire vector $\boldsymbol{z} = (z_1, z_2, \ldots, z_M)$ to estimate $\pi_0, f_0(z), f(z)$, and then $\text{fdr}(z)$,

$$\widehat{\text{fdr}}(z) = \hat{\pi}_0 \hat{f}_0(z)/\hat{f}(z), \tag{2.4}$$

rejecting the mth null hypothesis H_{0m} if $\widehat{\text{fdr}}(z_m)$ is small, perhaps for $\widehat{\text{fdr}}(z_m) \leq 0.1$ or ≤ 0.01. (Replacing densities f_0 and f_1 with their cumulative distribution functions gets us back, almost, to Benjamini and Hochberg's 1995 false discovery rate control algorithm, but here it will be more convenient to deal directly with the densities.)

Figure 2 shows $\widehat{\text{fdr}}(z)$ based on the combined data for all $M = 750,000$ z values z_{ij} (1.3), computed using the `locfdr` algorithm, Efron (2008); `locfdr` assumes that $f_0(z)$ is normal, a reasonable assumption here looking at the central portion of the $\{z_{ij}\}$ histogram, which the averaging in (1.2) renders quite Gaussian. The numerator estimates in (2.4) were

$$\hat{\pi}_0 = 0.954, \qquad \hat{f}_0 \sim \mathcal{N}(0.04, 0.93^2), \tag{2.5}$$

obtained using the "central matching" (geometric) method. Taken literally, this implies 4.6% of the (i, j) pairs represent cnv locations,

$$\hat{\pi}_1 = 0.046. \tag{2.6}$$

As discussed in Efron (2008), we can expect a majority of such pairs to have disappointingly large values of $\widehat{\text{fdr}}(z_{ij})$. Here only 1.5% of the 750,000 have $\widehat{\text{fdr}}(z_{ij}) \leq 0.1$, those with $z_{ij} \leq -3.30$ or ≥ 3.56. However, we can improve power by adapting the fdr methodology to the two-way structure of cnv data.

Let $\mathcal{C}_i$ be the class of pairs (i, j) corresponding to position i,

$$\mathcal{C}_i = \{(i, j) : j = 1, 2, \ldots, n\} \tag{2.7}$$

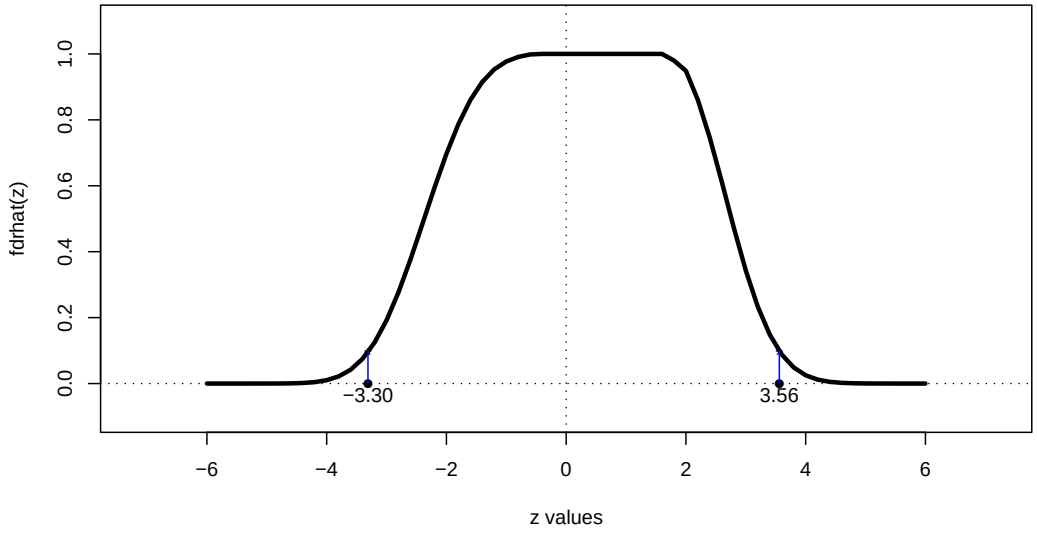

Figure 2: Estimated local false discovery rate $\widehat{\text{fdr}}(z)$ based on all 750,000 values z_{ij} (1.3) for the cnv data; $\widehat{\text{fdr}}(z)$ is ≤ 0.1 for $z \leq -3.30$ and $z \geq 3.56$, respectively 1.2% and 0.3% of the 750,000 cases. Computed from program `locfdr`, Efron (2008).

so the corresponding z values z_{ij} are all those obtained at position i. We now have N classes $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_N$, one for each position, and can imagine fitting a *separate* two-groups model (2.1) for each class, yielding separate false discovery rate functions $\text{fdr}_i(z)$. The trouble is that, unless n is very large, a direct approach as in Figure 2 will produce inaccurate estimates $\widehat{\text{fdr}}_i(z)$.

A compromise between using the combined estimate $\widehat{\text{fdr}}(z)$ or completely separate estimates $\widehat{\text{fdr}}_i(z)$ goes as follows: we assume that the null and non-null densities f_0 and f_1 in (2.1) apply unchanged to each class, but that the null and non-null prior probabilities may differ, say

$$\pi_{i0} = \Pr\{\text{null}|\mathcal{C}_i\} \quad \text{and} \quad \pi_{i1} = 1 - \pi_{i0} = \Pr\{\text{non-null}|\mathcal{C}_i\}. \tag{2.8}$$

So a cnv-prone position would be one having a larger value of π_{i1} than the combined value π_1.

Using $\text{fdr}_i(z) = \pi_{i0} f_0(z)/f_{(i)}(z)$, with $f_{(i)}(z)$ the mixture density applying to $\mathcal{C}_i$,

$$f_{(i)}(z) = \pi_{i0} f_0(z) + \pi_{i1} f_1(z), \tag{2.9}$$

comparison with (2.2) gives

$$\text{fdr}_i(z) = \text{fdr}(z)/[1 + \text{tdr}(z) \cdot R_i] \tag{2.10}$$

where $\text{tdr}(z)$ is the *true discovery rate*

$$\text{tdr}(z) = 1 - \text{fdr}(z) = \Pr\{\text{non-null}|z\} \tag{2.11}$$

and

$$R_i = \frac{\pi_{i1}/\pi_1}{\pi_{i0}/\pi_0} - 1. \tag{2.12}$$

An equivalent form is

$$\text{tdr}_i(z) = \text{tdr}(z)/[1 + \text{fdr}(z) \cdot S_i] \tag{2.13}$$

now where $\text{tdr}_i(z) = 1 - \text{fdr}_i(z)$ is the true discovery rate $\Pr\{\text{non-null}|z, \mathcal{C}_i\}$ applying to $\mathcal{C}_i$, and

$$S_i = \frac{\pi_{i0}/\pi_0}{\pi_{i1}/\pi_1} - 1. \tag{2.14}$$

Section 6 discusses (2.10) and (2.13) in more detail.

None of this seems like a step forward since (2.10) and (2.13) both require knowledge of π_{i1}, the non-null proportion in $\mathcal{C}_i$. There is, however, a simple iterative solution. Given a preliminary estimate $\widehat{\text{tdr}}_i(z)$ of $\text{tdr}_i(z)$, perhaps $\widehat{\text{tdr}}(z)$ from the combined analysis,

$$\hat{k}_i = \sum_{j \in \mathcal{C}_i} \widehat{\text{tdr}}_i(z_{ij}) = \sum_{j=1}^{n} \widehat{\text{tdr}}_i(z_{ij}) \tag{2.15}$$

is the obvious estimate of k_i, the number of non-null cases in $\mathcal{C}_i$, since $\widehat{\text{tdr}}_i(z_{ij})$ estimates the probability that case (i, j) is non-null.

This yields

$$\hat{\pi}_{i1} = \hat{k}_i/n = \sum_{j=1}^{n} \widehat{\text{tdr}}_i(z_{ij}) \Big/ n \tag{2.16}$$

as an estimate of π_{i1}, and

$$\hat{S}_i = \frac{(1 - \hat{\pi}_{i1})/\hat{\pi}_0}{\hat{\pi}_{i1}/\hat{\pi}_1} - 1 \tag{2.17}$$

for (2.14) (with $\hat{\pi}_0$ and $\hat{\pi}_1$ obtained from the combined analysis that gave $\widehat{\text{fdr}}$ and $\widehat{\text{tdr}}$, as in (2.6)). We can now update (2.13) to

$$\widehat{\text{tdr}}_i(z_{ij}) = \widehat{\text{tdr}}(z_{ij}) \Big/ \left[1 + \widehat{\text{fdr}}(z_{ij})\hat{S}_i\right], \tag{2.18}$$

recompute (2.15), etc. The numerical results that follow stopped after five iterations of (2.15)–(2.18), close to the final convergence values; see Section 6. Other examples, involving fewer subjects, required more iterations to reach convergence, though the increase did not noticeably affect subsequent inferences.

Figure 3 displays the results. The top panel shows those pairs (i, j) having $\widehat{\text{tdr}}_i(z_{ij}) \geq 0.99$, or equivalently $\widehat{\text{fdr}}_i(z_{ij}) \leq 0.01$. A very long cnv region is centered around $i = 3800$, with shorter but still prominent regions near 1, 1100, 1300, 3000, and 4700. Position $i = 1755$ is less impressive, but does show some non-null cases. The bottom panel graphs $\hat{k}_i$ as a function of position i, with $\hat{k}_{1755} = 39.1$ showing as a small isolated spike. Is this a "significant" result? The next two sections consider testing and estimation questions for the $\hat{k}_i$ values.

3 Hypothesis Tests for Position-Wise Copy Number Variation

Having obtained estimates $\hat{k}_i,\ i = 1, 2, \ldots, N$, for the number of non-null cnv subjects at marker position i, we wish to decide which, if any, of the positions are unusually prone to copy number variation. For example, $\hat{k}_{1755}$ equals 39.1, compared to the average $\bar{k} = 8.13$ in Figure 3, which might suggest excess variation at position 1755, an hypothesis we would like to test.

An easy permutation test proceeds as follows: let $\boldsymbol{z}_j$ be the jth column of the $\boldsymbol{Z}$ matrix $\{z_{ij}\}$ (1.3), and $\boldsymbol{z}_j^*$ the same vector except shifted left I units (with wraparound),

$$\boldsymbol{z}_j^* = (z_{I+1,j}, z_{I+2,j}, \ldots, z_{N,j}, z_{1j}, z_{2j}, \ldots, z_{Ij})'. \tag{3.1}$$

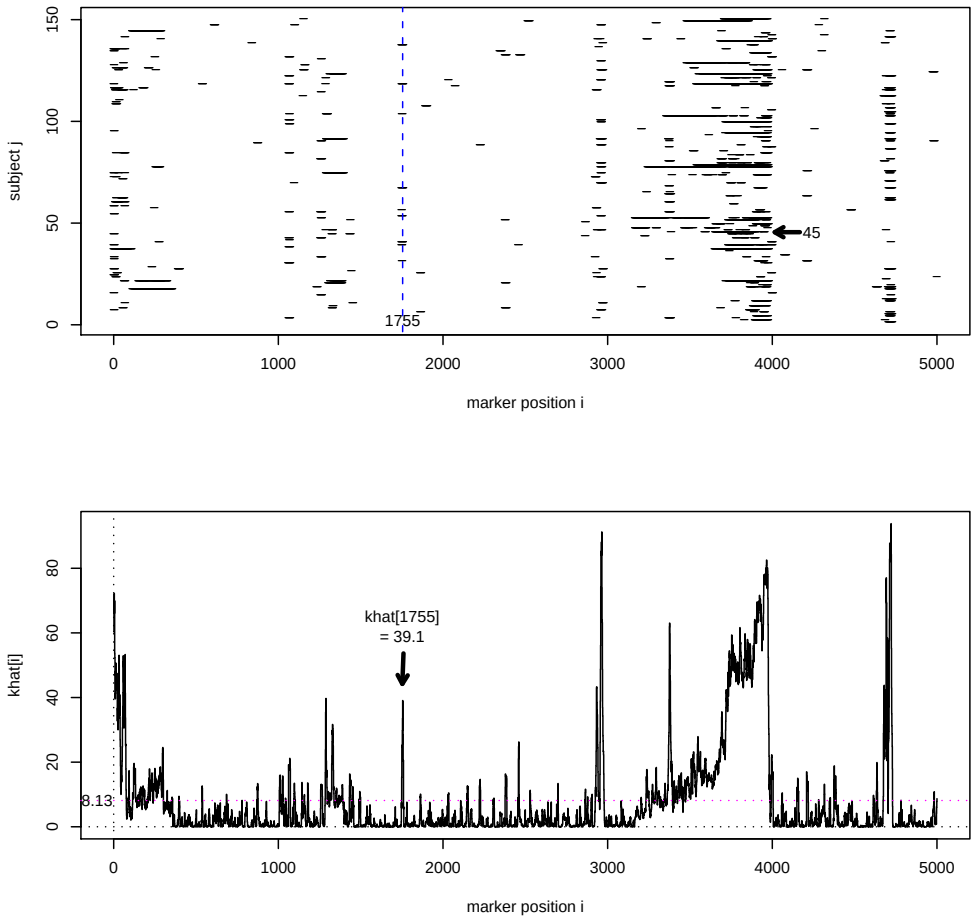

Figure 3: Algorithm (2.15)–(2.18), five iterations, applied to cnv data (1.1). *Top panel*: (position, subject) pairs (i, j) having estimated true discovery rate $\widehat{\text{tdr}}_i(z_{ij}) \geq 0.99$. *Bottom panel*: estimates $\hat{k}_i$ for the number of non-null subjects at marker position i.

Choosing I as an independent and random integer between 0 and $N-1$ for each of the n rows yields a permuted matrix,

$$\boldsymbol{Z}^* = (\boldsymbol{z}_1^*, \boldsymbol{z}_2^*, \ldots, \boldsymbol{z}_n^*) \tag{3.2}$$

in which position-wise structure has been nullified, while any subject-wise structure of cnv intervals is maintained. The permutation test compares $\hat{k}_i$ with the distribution $\{\hat{k}_1^*, \hat{k}_2^*, \ldots, \hat{k}_N^*\}$, where the k^* values are obtained by applying algorithm (2.15)–(2.18) to $\boldsymbol{Z}^*$. (Notice that $\boldsymbol{Z}^*$ has the same elements as $\boldsymbol{Z}$, so that the combined analysis quantities $\hat{\pi}_0, \hat{\pi}_1, \widehat{\text{fdr}}(z)$ and $\widehat{\text{tdr}}(z)$ have the same values as in (2.15)–(2.18).)

Ten independent replications of $\boldsymbol{Z}^*$ were generated for the example of Figure 3, yielding 50,000 $\hat{k}_i^*$ values in total. The line histogram in Figure 4 compares them with the distribution of the 5000 actual $\hat{k}_i$ values: $\max\{k_i^*\} = 23.3$ suggesting, for example, that $\hat{k}_{1755} = 39.1$ is strongly significant evidence for excess variation at position 1755. In less-extreme circumstances we could compute permutation p-values,

$$p_i = \text{proportion of permutation values exceeding } \hat{k}_i \tag{3.3}$$

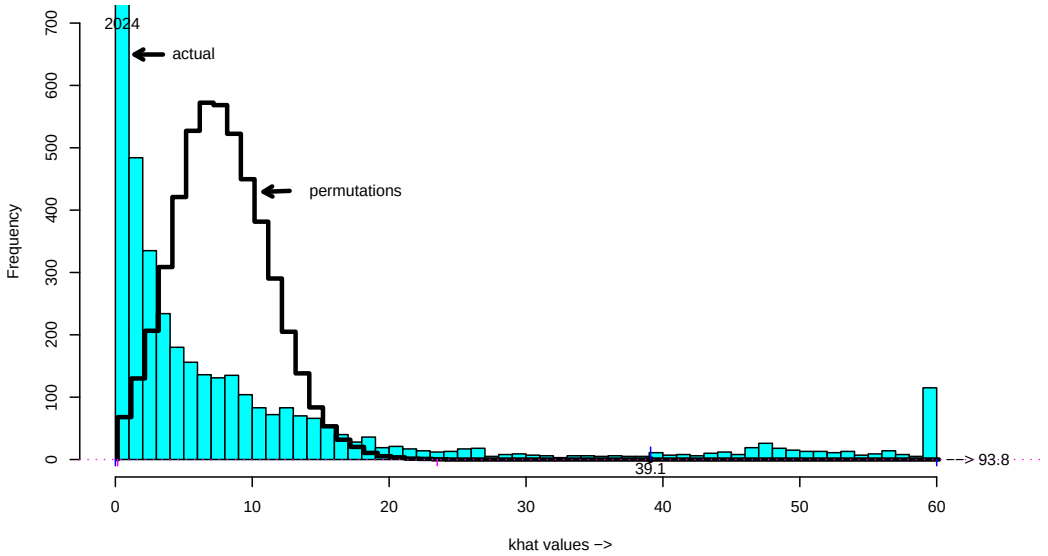

Figure 4: Histogram of the 5000 $\hat{k}_i$ estimates for cnv data, from 5 iterations of (2.15)–(2.18) (solid), compared to 50,000 permutation values $\hat{k}_i^*$ as following (3.2) (line histogram). Maximum permutation value equals 23.3, far less than $\hat{k}_{1755} = 39.1$. Spike at $\hat{k} = 60$ represents 106 $\hat{k}_i$ values ≥ 60, maximum 93.8. The 2024 $\hat{k}_i$ values ≤ 1 are significantly too *small* according to the permutation distribution.

and use a standard false discovery rate procedure to assess significance among the N p-values.

Basing our significance tests on $\hat{k}_i$ values seems reasonable but perhaps *ad hoc*. It can, however, be motivated in terms of the two-groups model (2.1), (2.8). Define

$$r = \pi_{i1}/\pi_1 \tag{3.4}$$

the ratio of the non-null probability at the ith position to the combined value π_1; the null hypothesis H_{0i} that position i is *not* cnv-prone is $H_{0i} : r = 1$.

Observation z_{ij} has density $f(z)$ under H_{0i} and density $f_{(i)}(z)$ under (2.8), (2.9), giving log likelihood ratio

$$\begin{aligned} l(z_{ij}) &= \log\left\{\frac{f_{(i)}(z_{ij})}{f(z_{ij})}\right\} = \log\left\{\frac{\pi_{i0}f_0(z_{ij}) + \pi_{i1}f_1(z_{ij})}{\pi_0 f_0(z_{ij}) + \pi_1 f_1(z_{ij})}\right\} \\ &= \log\left\{1 + (r-1)T(z_{ij})\right\} \end{aligned} \tag{3.5}$$

where some calculation yields

$$T(z) = \frac{\mathrm{tdr}(z) - \pi_1}{\pi_0}, \tag{3.6}$$

$\mathrm{tdr}(z) = 1 - \mathrm{fdr}(z)$ as in (2.11). Assuming that subjects were sampled independently, the n observations in $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots, z_{in})$ are independent of each other. The most powerful test of H_{0i} versus a specific alternative choice of $r > 1$ then rejects H_{0i} for large values of

$$l_r(\boldsymbol{z}_i) = \sum_{j=1}^{n} \log\left\{1 + (r-1)T(z_{ij})\right\}. \tag{3.7}$$

The locally most powerful (lmp) test of $r = 1$ versus $r > 1$ rejects for large values of

$$\left.\frac{\partial l_r(\boldsymbol{z}_i)}{\partial r}\right|_{r=1} = \sum_{j=1}^{n} T(z_{ij}) = \sum_{j=1}^{n} \frac{\mathrm{tdr}(z_{ij}) - \pi_1}{\pi_0}, \tag{3.8}$$

an increasing function of $\sum_1^n \mathrm{tdr}(z_{ij})$. In practice we could reject for large values of

$$\sum_{j=1}^{n} \widehat{\mathrm{tdr}}(z_{ij}) = \hat{k}_i^{(1)} \tag{3.9}$$

where $\hat{k}_i^{(1)}$ is from the *first* iteration of k_i in (2.15), beginning at $\widehat{\mathrm{tdr}}_i(z) = \widehat{\mathrm{tdr}}(z)$. This justifies $\hat{k}_i^{(1)}$ as a preferred test statistic for H_{0i}.

In our cnv example, $\hat{k}_i^{(5)}$, the fifth iterate, performed a little better than $\hat{k}_i^{(1)}$, almost matching the most powerful test statistic (3.5) over the range $1 < r \le 4$. This seems to put the significance of position 1755 as cnv-prone on safe footing.

Figure 4 is not completely reassuring in this regard: the permutation distribution does not look much like a reasonable null hypothesis, since it makes "significant" a majority of the 5000 positions. In particular, the 2024 positions having $\hat{k}_i \le 1$ are significantly too *small* by the permutation criterion. Perhaps we should be estimating the accuracy of the $\hat{k}_i$ values rather than testing them for nullness, a point of view taken up in Section 4.

4 The Accuracy of Position-Wise Estimates

How accurate is $\hat{k}_i$ as an estimate of k_i, the number of non-null cases at position i? A simple answer is obtained by resampling the n subjects and calculating non-parametric bootstrap estimates of standard deviations.

Let $\boldsymbol{z}_j = (z_{1j}, z_{2j}, \ldots, z_{Nj})'$ be the N-vector of data for subject j. A typical bootstrap data matrix is

$$\boldsymbol{Z}^* = (\boldsymbol{z}_{j_1}, \boldsymbol{z}_{j_2}, \ldots, \boldsymbol{z}_{j_n}) \tag{4.1}$$

where $j_1, j_2, \ldots, j_n$ is a random sample taken with replacement from the integers $(1, 2, \ldots, n)$. We calculate $k_i^*,\ i = 1, 2, \ldots, N$, from $\boldsymbol{Z}^*$ according to (2.15)–(2.18), including the five iterations. Doing so B times gives bootstrap standard deviation estimates

$$\widehat{\mathrm{sd}}_i = \sqrt{\frac{\sum_{b=1}^{B} \left(\hat{k}_i^{*b} - \hat{k}_i^{*\cdot}\right)^2}{(B-1)}} \tag{4.2}$$

where $\hat{k}_i^{*\cdot} = \sum_1^B \hat{k}_i^{*b}/B$.

Figure 5 plots $\widehat{\mathrm{sd}}_i$ versus $\hat{k}_i$ for the cnv data (1.1), $i = 1, 2, \ldots, 5000$, based on $B = 200$ bootstrap replications. A smooth curve has been drawn through the 5000 $(\hat{k}_i, \widehat{\mathrm{sd}}_i)$ points, giving for example $\widehat{\mathrm{sd}}_{1755} = 6.5$ at $\hat{k}_{1755} = 39.1$. This yields approximate 95% confidence intervals $\hat{k}_i \pm 2 \cdot \widehat{\mathrm{sd}}_i$, in particular,

$$k_{1755} \in (26.1, 52.1). \tag{4.3}$$

The lower limit is far above $\bar{k} = 8.3$, providing further evidence that position 1755 is cnv-prone. *Note*: The bootstrap calculations did not include recomputation of the combined quantities

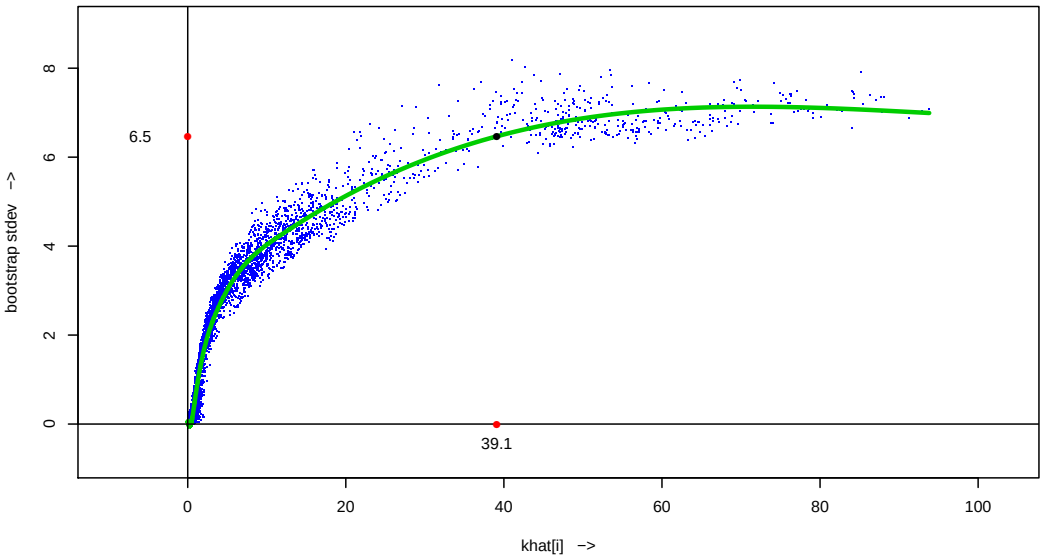

Figure 5: Bootstrap standard deviations of $\hat{k}_i$ estimates (2.15)–(2.18), 5 iterations, for cnv data (1.1), plotted versus $\hat{k}_i$. Smooth curve is natural spline least squares fit, 4 degrees of freedom. At $\hat{k}_{1755} = 39.1$ it gives $\widehat{\text{sd}}_{1755} = 6.5$.

$\hat{\pi}_0, \hat{\pi}_1, \widehat{\text{tdr}}(\cdot)$, and $\widehat{\text{fdr}}(\cdot)$ in (2.15)–(2.18), which were kept at their original values. This amounts to treating them as fixed ancillary statistics, as is effectively done in the permutation test of Section 3. Recomputing the combined quantities for each bootstrap replication considerably increased the standard deviation estimates ($\widehat{\text{sd}}_{1755} = 9.1$ for example), and seemed inappropriately conservative here.

At this point, *selection bias* needs to be considered: positions such as 1755 come to our attention because of their unusual $\hat{k}_i$ values, which can be misleadingly large when selected from thousands of possibilities. Frequentist corrections for bias are difficult here, but a simple empirical Bayes calculation offers some insight.

Consider the univariate Bayesian model in which a parameter μ is drawn from prior density $g(\cdot)$ and then $x \sim \mathcal{N}(\mu, \sigma^2)$ is observed,

$$\mu \sim g(\cdot) \quad \text{and} \quad x|\mu \sim \mathcal{N}(\mu, \sigma^2). \tag{4.4}$$

Let $f(x)$ be the marginal density of x,

$$f(x) = \int_{-\infty}^{\infty} \varphi_\sigma(x-\mu) g(\mu)\, d\mu, \tag{4.5}$$

$\varphi_\sigma(x) = \exp\{-x^2/2\sigma^2\}/\sqrt{2\pi\sigma^2}$, and define $l(x) = \log\{f(x)\}$.

Lemma 1. *The posterior expectation and standard deviation of μ given x are*

$$E\{\mu|x\} = x + \sigma^2 l'(x) \tag{4.6}$$

and

$$\text{Sd}\{\mu|x\} = \sigma \cdot \left[1 + \sigma^2 l''(x)\right]^{1/2} \tag{4.7}$$

where $l'(x)$ and $l''(x)$ are the first and second derivations of $l(x)$.

Proof. According to Bayes theorem, the posterior density of μ given x is

$$\begin{aligned} g(\mu|x) &= \varphi_\sigma(x-\mu)g(\mu)/f(x) \\ &= e^{x\mu/\sigma^2-\psi(x)}\left[g(\mu)e^{-\mu^2/2\sigma^2}\right] \end{aligned} \tag{4.8}$$

with

$$\psi(x) = \log\{f(x)/\varphi_\sigma(x)\}; \tag{4.9}$$

(4.8) is a one-parameter exponential family with canonical parameter x and sufficient statistic μ/σ^2. Differentiating $\psi(x)$ twice yields the mean and variance of μ/σ^2 given x, verifying the lemma. ■

Formula (4.6) goes back, at least, to Robbins (1956), who credits correspondence with M. Tweedie, though (4.7) seems less familiar. They are ideal for empirical Bayes purposes: having observed $x_1, x_2, \ldots, x_N$ from repeated realizations (μ_i, x_i) of (4.4), we can directly estimate $f(x)$ and $l(x)$, and differentiate to get $E\{\mu|x\}$ and $\mathrm{sd}\{\mu|x\}$ from (4.6)–(4.7). The key point is that deconvolution for the estimation of the prior $g(\mu)$ is completely avoided.

Now let $(k_i, \hat{k}_i)$ play the role of (μ_i, x_i) in (4.4). The 200 bootstrap replications for Figure 5 showed

$$\hat{k}_i \doteq \mathcal{N}(k_i, \sigma_i^2) \tag{4.10}$$

to be a reasonable approximation, with σ_i as indicated in Figure 5. A density estimate $\hat{f}(k)$ was obtained by fitting a smooth curve to the histogram heights in Figure 4 (using a Poisson generalized linear model based on a natural spline with five degrees of freedom, as described in Remark D of Efron, 2009). This gave $\hat{l}(k) = \log\{\hat{f}(k)\}, \hat{l}'(k), \hat{l}''(k)$, and then $\hat{E}\{k_i|\hat{k}_i\}$ and $\widehat{\mathrm{Sd}}\{k_i|\hat{k}_i\}$, with σ^2 obtained from the fitted curve in Figure 4.

$\hat{k}$	10	20	30	40	50	60	70	80
$\hat{E} - 2\cdot\widehat{\mathrm{sd}}$	−1.0	5.6	13.4	28.1	41.3	41.7	50.8	68.6
$\hat{\boldsymbol{E}}$	**7.4**	**16.3**	**27.7**	**42.4**	**49.6**	**56.1**	**68.8**	**83.3**
$\hat{E} + 2\cdot\widehat{\mathrm{sd}}$	15.7	27.1	42.0	56.8	57.8	70.4	86.8	98.1

Table 1: Empirical Bayes estimates $\hat{E}\{k|\hat{k}\}$ and posterior limits $\hat{E}\{k|\hat{k}\} \pm 2\cdot\widehat{\mathrm{Sd}}\{k|\hat{k}\}$.

Table 1 displays $\hat{E}\{k|\hat{k}\}$ and the posterior bounds $\hat{E}\{k|\hat{k}\} \pm 2\cdot\widehat{\mathrm{sd}}\{k|\hat{k}\}$. Unlike the examples in Efron (2009), the results here are not very different from the frequentist estimates $\hat{k}_i \pm 2\cdot\widehat{\mathrm{sd}}_i$. In particular, for $\hat{k}_{1755} = 39.1$ we get $\hat{E} = 41.3$ and posterior interval (26.5,56.0), almost the same as (4.3). Figure 4 shows a greater concentration of $\hat{k}_i$ values within a couple of standard deviations to the right of 39.1 than to the left, accounting for the slightly increased Bayesian estimate and interval limits.

Bayes estimates are immune to selection bias. If in fact the posterior expectation of k_{1755} equals 41.3, then it does not matter why position 1755 came to our attention. The reassuring message of Table 1 is that selection bias is not a serious problem here.

There are reasons for skepticism:

- Model (4.4) has σ constant, whereas it varies in our application. More careful calculations show that the effect is small for this situation (only slightly raising the estimates for position 1755).
- At best, the calculations are approxiimating $E\{k_i|\hat{k}_i\}$, not $E\{k_i|\hat{\boldsymbol{k}}\}$, the posterior expectation given *all* the k values.
- The $\hat{k}_i$ estimates are correlated with each other. This does not invalidate the use of Lemma 1, but degrades the accuracy of the empirical Bayes estimates; see Efron (2010a).

These last two points emphasize the fact that empirical Bayes is not actual Bayes, and provides no strict theoretical basis for ignoring selection bias. Nevertheless, the results in Table 1 offer a useful guide for interpreting the estimates $\hat{k}_i$.

Various numerical experiments were carried out investigating the accuracy of $\hat{k}_i$ calculations. The observations x_{ij} in (1.1) actually were each the average of two independent replications x_{ij1} and x_{ij2}. Applying algorithm (2.15)–(2.18) separately to the two sets gave nearly the same results, both being slightly degraded versions of the analysis based on (1.1).

Another test involved "spiking in" artificial cnv signals at non-active positions of data (1.1); for example, adding a square wave signal to 40 of the subjects at positions 2233 through 2239. The corresponding $\hat{k}_i$ values edged up to 40 as the size of the square wave increased, topping out at about 50 for enormous signals. (The window width $2m+1$ in (1.2) was kept at 11 as before.) Large numbers of low values of $\widehat{\text{tdr}}_i(z_i)$ in (2.15) were responsible for the upward bias, which perhaps suggests imposing a cut-off threshold. Section 5 briefly discusses the relation of window width to power and bias.

At this point, we reveal the fact that probe number 1755, which is located at genome base pair position 17,952,757 (NCBI human genome build 36), indeed falls into a region containing previously identified deletions. The deletions in this region have been detected by Conrad et al. (2006) using SNP genotyping arrays and by Mills et al. (2006) and McKernan et al. (2009) using short read sequencing. These studies differ in their estimated boundaries, but all agree that there is a deletion covering probe 1755 in at least one subject in their study.

5 Estimation of False Discovery Rates

The procedures used for the estimation of $\text{fdr}_i(z)$ (2.10), the local false discovery rate applying to position i, raise some questions discussed in this section.

A preliminary question concerns the choice of moving average window width $M = 2m+1$ involved in the construction of the z-values z_{ij} (1.2)–(1.3). Some insight is gained from a simple model in which the observations x_{ij} in (1.1) are independent normal variates with expectation either 0 or μ,

$$\textit{null } x_{ij} \sim \mathcal{N}(0,1) \qquad \textit{non-null } x_{ij} \sim \mathcal{N}(\mu,1) \tag{5.1}$$

and where the non-null cases for subject j occur in contiguous blocks.

The adjusted moving average from (1.2),

$$z_{ij} = \sqrt{M}\bar{x}_{ij} \tag{5.2}$$

has distributions

$$\textit{null } z_{ij} \sim \mathcal{N}(0,1) \qquad \textit{non-null } z_{ij} \sim \mathcal{N}(\mu_M,1) \tag{5.3}$$

where, if the moving average is taken entirely within a contiguous non-null block, we have

$$\mu_M = \sqrt{M}\mu. \tag{5.4}$$

Averaging increases the null/non-null separation in this case, improving the power of our detection procedure, as made explicit next.

The ratio of true to false discovery rates in position i is

$$\frac{\mathrm{tdr}_i(z)}{\mathrm{fdr}_i(z)} = \frac{\Pr\{\text{non-null}|z, \mathcal{C}_i\}}{\Pr\{\text{null}|z, \mathcal{C}_i\}} = \frac{\pi_{i1} f_1(z)}{\pi_{i0} f_0(z)}, \tag{5.5}$$

in the notation of Section 2. Then (5.3) yields

$$\frac{\mathrm{tdr}_i(z)}{\mathrm{fdr}_i(z)} = \frac{\pi_{i1}}{\pi_{i0}} e^{\mu_M(z-\mu_M/2)}. \tag{5.6}$$

Under (5.4),

$$\frac{\mathrm{tdr}_i(z)}{\mathrm{fdr}_i(z)} = \frac{\pi_{i1}}{\pi_{i0}} e^{M\mu^2/2} \tag{5.7}$$

at $z = \mu_M$, its non-null expected value, so increasing the window width M raises the ratio exponentially fast. Put simply, large M produces non-null z-values far from 0, at least at positions inside long non-null blocks.

Suppose though that the non-null block length M_{non} is less than M. The same reasoning as in (5.7) gives

$$\frac{\mathrm{tdr}_i(\mu_M)}{\mathrm{fdr}_i(\mu_M)} = \frac{\pi_{i1}}{\pi_{i0}} e^{(M_{\mathrm{non}}^2/M)\mu^2/2} \tag{5.8}$$

at the block's central position, so that increasing M is now harmful. The ideal choice is $M = M_{\mathrm{non}}$, the well-known *signal matching criterion*, but of course in practice we won't know M_{non}.

Other considerations come into play: larger M improves the normality of z_{ij}, null normality being an important assumption in (2.5); correlation between nearby x_{ij}'s decreases the advantage of averaging; long non-null intervals like those seen near $i = 3800$ in Figure 3 may include sub-blocks of negative as well as positive cnv effect. (See the discussion on one-sided procedures below.)

Changing M from 11 to 21 produced small increases in most of the larger $\hat{k}_i$ values seen in Figure 3, a notable exception being at $i = 1755$ — inside a very short block — where $\hat{k}_i$ was halved. The value $M = 11$ performed satisfactorily on several other data sets, though the specific choice never seemed crucial.

Our data set (1.1) includes copy number variations in both negative and positive directions, that is, having less or more than two copies. This can be seen in Figure 2, where the combined local false discovery rate $\widehat{\mathrm{fdr}}(z)$ decreases to zero at both ends of the z scale. As a consequence, the estimates $\hat{k}_i$ produced by algorithm (2.15)–(2.18) are *two-sided*: if we begin the iteration with $\widehat{\mathrm{tdr}}_i(z) = \widehat{\mathrm{tdr}}(z) = 1 - \widehat{\mathrm{fdr}}(z)$ then $\hat{k}_i$ is increased for z_{ij} values that are extreme in either direction.

As we will show in the following, two-sidedness can have undesirable effects. It is simple, and probably preferable, to calculate instead both *one-sided* $\hat{k}_i$ estimates. Beginning the iteration at (2.15) with

$$\widehat{\mathrm{tdr}}_i(z) = \begin{cases} \widehat{\mathrm{tdr}}(z) & \text{if } z \le 0 \\ 0 & \text{if } z > 0 \end{cases} \tag{5.9}$$

instead of $\widehat{\text{tdr}}(z)$ produces "left-sided" $\hat{k}_i$ estimates, sensitive only to negative z_{ij} values. A similar tactic gives right-sided $\hat{k}_i$ estimates. The sum of the left- and right-sided estimates is similar to the two-sided estimates of Section 2, but there is an interpretive advantage in observing both sides.

Some of the positions for data set (1.1) (though not 1755) displayed large z_{ij} in both directions. These can be genuine, but we might worry that an uncontrolled effect, perhaps a microarray reading difficulty at position i, has artificially broadened the distribution of the n z_{ij} values. A drastic cure is to standardize positions as well as subjects, that is, to perform a second standardization (1.3) with the roles of i and j reversed. Doing so seemed to remove more signal than noise for data (1.1), and is not recommended. Nevertheless, a plot of robust standard deviations as a function of position i may help reveal systematic reading problems.

Formula (2.10), which is the basis of our iterative algorithm (2.15)–(2.18), depends on the strong assumption that $f_0(z)$ and $f_1(z)$, the null and non-null densities in the two-groups model (2.1), apply unchanged to each position $\mathcal{C}_i$. A more general result that allows the non-null density to depend on $\mathcal{C}_i$ (while $f_0(z)$ is still assumed fixed) is developed in Efron (2009) and in Section 10 of Efron (2010b). Define

$$w_i(z) = \Pr\{\mathcal{C}_i|z\}. \tag{5.10}$$

Then, to a good approximation,

$$\text{fdr}_i(z) \doteq \text{fdr}(z)\frac{w_i(0)}{w_i(z)}. \tag{5.11}$$

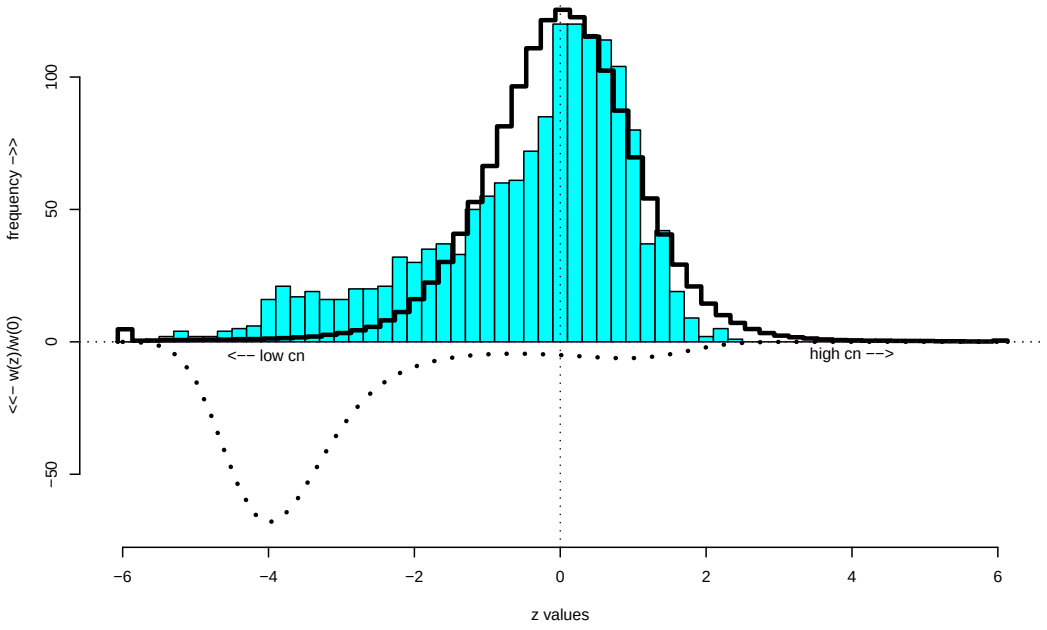

Figure 6: *Solid histogram* the 1500 z_{ij} values for positions i in 1750–1759, cnv data (1.1); *line histogram* for the 748,500 other z_{ij}; *dotted curve* cubic logistic regresssion estimate (5.12) for $w_i(z)/w_i(0)$ (5.10) (multiplied by -5 for display).

Figure 6 and Figure 7 concern the application of (5.11) to the amalgamated set of positions 1750 through 1759. The solid histogram in Figure 6 shows the 1500 z_{ij} at these 10 positions having an excess of negative values, compared to the distribution of all the rest. A logistic regression of the indicator

$$I_{ij} = \begin{cases} 1 & \text{if } i \in 1750:1759 \\ 0 & \text{otherwise} \end{cases} \tag{5.12}$$

as a cubic function of z_{ij} gave estimate $\hat{w}_i(z)$; the ratio $\hat{w}_i(z)/\hat{w}_i(0)$ is plotted below the horizontal axis.

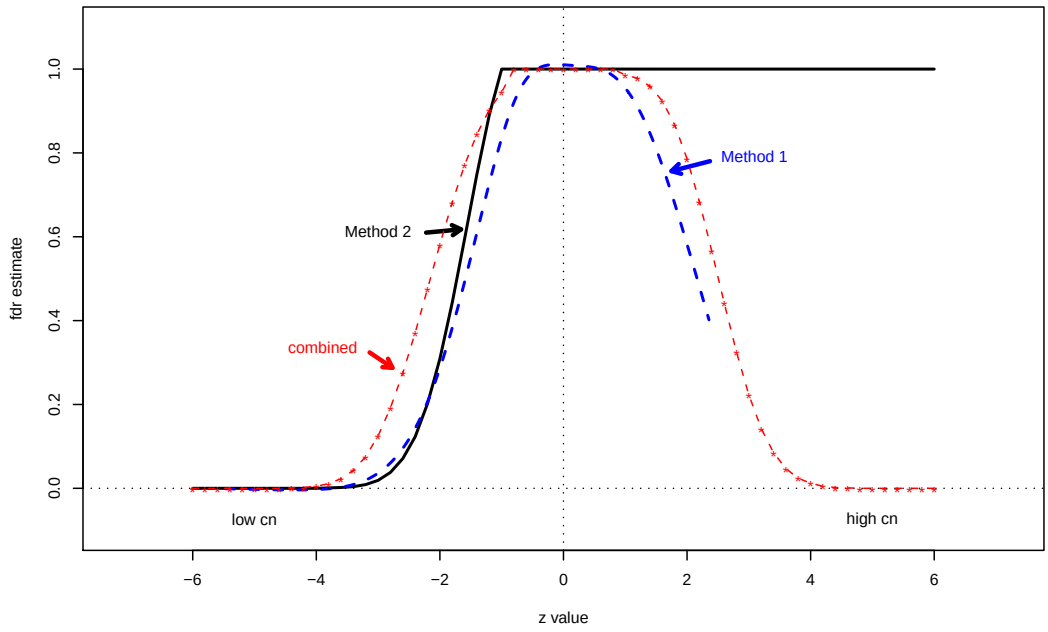

Figure 7: Three estimates of the local false discovery rate for positions 1750–1759 of cnv data (1.1): *combined* from all 750,000 z_{ij}, as in Figure 2; *Method 1* from 5 iterations of (2.15)–(2.18), two-sided; *Method 2* from (5.11), using $\hat{w}_i(z)$ as shown in Figure 6.

Three estimates $\widehat{\text{fdr}}_i(z)$ for the local false discovery rate applying to positions 1750–1759 appear in Figure 7: the combined estimate of Figure 2, obtained from all 750,000 z_{ij} values; *Method 1*, from five iterations of algorithm (2.15)–(2.18), applied in the two-sided fashion of Section 2; and *Method 2*, from formula (5.11), with $w_i(z)/w_i(0)$ as shown in Figure 6.

On the left side, both Method 1 and Method 2 yield much smaller estimates than the combined curve, for instance $\widehat{\text{fdr}}_i(-3) = 0.019$ from Method 2, compared to the combined estimate 0.125. Both Methods are adjusting the combined estimates $\widehat{\text{fdr}}(z_{ij})$ downward to account for the excess cnv activity observed at positions 1750–1759.

The story is different on the right side. Method 2 has $\widehat{\text{fdr}}_i(z) = 1$ for $z > 0$, which is intuitively correct since Figure 6 shows no tendency toward large positive z-values in positions 1750–1759. Many of the other positions *do* show unusually large positive z-values, causing the combined estimate $\widehat{\text{fdr}}(z)$ to decline for large positive z. Method 1's estimate declines even more sharply. It is using non-null density $f_1(z)$ (2.1) as estimated from the combined data, and does not "know" that the non-null values in positions 1750–1759 are only left-sided.

Applying Method 1 separately on the left and right, as in (5.9), resolves the discrepancy with Method 2. Method 2 itself tends to be noisy when applied to individual positions, and the two one-sided versions of Method 1 seem preferable in general.

6 Convergence Properties of the Iterative Algorithm

Algorithm (2.15)–(2.18) was stopped after five iterations since numerical convergence of the $\hat{k}_i$ values had nearly been achieved, producing the results pictured in Figure 3, Figure 4, and Figure 5. This section discusses the theoretical convergence point of the algorithm, leading to a formula for its standard error, and a connection with maximum likelihood estimation in model (3.7). The development will be in terms of $\hat{\pi}_{i1} = \hat{k}_i/n$ (2.16), rather than $\hat{k}_i$ itself.

Returning to the two-groups notation of Section 2, let p_1 and $p_0 = 1-p_1$ take values between 0 and 1, and define

$$f(z, p_1) = p_1 f_1(z) + p_0 f_0(z) \tag{6.1}$$

and

$$\text{tdr}(z, p_1) = \frac{p_1 f_1(z)}{f(z, p_1)} = \frac{1}{1 + \frac{p_0}{p_1} L(z)} \tag{6.2}$$

where $L(z)$ is the likelihood ratio

$$L(z) = f_0(z)/f_1(z). \tag{6.3}$$

The actual true discovery rate in class $\mathcal{C}_i$, $\Pr\{\text{non-null}|z, \mathcal{C}_i\}$, is

$$\text{tdr}_i(z) = \text{tdr}(z, \pi_{i1}) = 1\big/\left[1 + (\pi_{i0}/\pi_{i1}) L(z)\right]. \tag{6.4}$$

(A little algebra shows that (6.4) equals (2.13).)

Finally, define

$$h_i(p_1) = p_1 - \int_{\mathcal{Z}} \text{tdr}(z, p_1) f_{(i)}(z)\, dz \tag{6.5}$$

where $f_{(i)}(z)$ equals $f(z, \pi_{i1})$, the mixture distribution (2.9) of z in $\mathcal{C}_i$, and the integral is taken over $\mathcal{Z}$, the sample space of z. Since

$$\int_{\mathcal{Z}} \text{tdr}(z, \pi_{i1}) f_{(i)}(z)\, dz = \int_{\mathcal{Z}} \frac{\pi_{i1} f_1(z)}{f_{(i)}(z)} f_{(i)}(z)\, dz = \pi_{i1}, \tag{6.6}$$

the function $h_i(p_1)$ satisfies

$$h_i(\pi_{i1}) = 0; \tag{6.7}$$

$h_i(\cdot)$ will turn out to determine the convergence point of algorithm (2.15)–(2.18), and also the delta-method standard error of the converged estimate.

Lemma 2. *The derivative of* $h_i(p_1)$ *is*

$$h_i'(p_1) = 1 - \int_{\mathcal{Z}} \text{tdr}(z, p_1)\, \text{fdr}(z, p_i) f_{(i)}(z)\, dz / p_1 p_0 \tag{6.8}$$

where $\text{fdr}(z, p_1) = 1 - \text{tdr}(z, p_1)$ (6.2).

Proof. From (6.2), we calculate

$$\begin{aligned}\frac{\partial \operatorname{tdr}(z,p_1)}{\partial p_1} &= \frac{1}{\left[1+\left(\frac{1}{p_1}-1\right)L(z)\right]^2}\frac{L(z)}{p_1^2} = \frac{\operatorname{tdr}(z,p_1)^2}{p_1^2}\frac{f_0(z)}{f_1(z)} \\ &= \frac{\operatorname{tdr}(z,p_1)}{p_1^2}\frac{p_1 f_1(z)}{f(z,p_1)}\frac{f_0(z)}{f_1(z)} = \frac{\operatorname{tdr}(z,p_1)\operatorname{fdr}(z,p_1)}{p_1 p_0}\end{aligned} \tag{6.9}$$

which gives (6.8) from (6.2). ■

The derivative $h'(p_1)$ takes on a convenient form at $p_1 = \pi_{i1}$ (the actual non-null probability in class $\mathcal{C}_i$), where $h_i(\pi_{i1}) = 0$.

Lemma 3.

$$h_i'(\pi_{i1}) = \xi_i/\pi_{i1}\pi_{i0} \tag{6.10}$$

with

$$\xi_i = \int_{\mathcal{Z}} \left[\operatorname{tdr}(z,\pi_{i1}) - \pi_{i1}\right]^2 f_{(i)}(z)\, dz, \tag{6.11}$$

the variance of $\operatorname{tdr}(z,\pi_{i1}) = \operatorname{tdr}_i(z)$ *in* $\mathcal{C}_i$ *(which is also the variance of* $\operatorname{fdr}_i(z)$ *in* $\mathcal{C}_i$*).*

Proof. Define I to be the null indicator for a random case in $\mathcal{C}_i$,

$$I = \begin{cases} 1 & \text{if null} \\ 0 & \text{if non-null,} \end{cases} \tag{6.12}$$

so

$$\pi_{i1} = \Pr\{I = 0|\mathcal{C}_i\} \quad \text{and} \quad \operatorname{tdr}_i(z) = \Pr\{I = 0|z_i, \mathcal{C}_i\}. \tag{6.13}$$

At $p_1 = \pi_{i1}$, (6.8) becomes

$$h_i'(\pi_{i1}) = 1 - \int_{\mathcal{Z}} \operatorname{tdr}_i(z)\operatorname{fdr}_i(z) f_{(i)}(z)\, dz/\pi_{i1}\pi_{i0}. \tag{6.14}$$

But $\operatorname{tdr}_i(z)\operatorname{fdr}_i(z)$ equals $\operatorname{var}_i\{I|z\}$, the conditional variance of the Bernoulli random variable I, so

$$h_i'(\pi_{i1}) = 1 - E_i\left\{\operatorname{var}_i\{I|z\}\right\}/\pi_{i1}\pi_{i0}, \tag{6.15}$$

E_i indicating expectation with respect to $f_{(i)}(z)$.

A standard relationship between conditional and unconditional variances is

$$\operatorname{var}_i\{I\} = E_i\left\{\operatorname{var}_i\{I|z\}\right\} + \operatorname{var}_i\left\{E_i\{I|z\}\right\}, \tag{6.16}$$

var_i indicating variance with respect to $f_{(i)}(z)$. Since $E_i\{I|z\} = \operatorname{fdr}_i(z)$, (6.15)–(6.16) imply $h_i'(\pi_{i1}) = \operatorname{var}_i\{\operatorname{fdr}_i(z)\}/\pi_{i1}\pi_{i0}$, which is the same as (6.10). ■

Note. Since $\pi_{i1}\pi_{i0} = \operatorname{var}_i\{I\}$, we can also write (6.8) as

$$h_i'(\pi_{i1}) = \operatorname{var}_i\left\{E_i\{I|z\}\right\}/\operatorname{var}_i\{I\} \le 1. \tag{6.17}$$

An empirical version of these theoretical results brings us back to algorithm (2.15)–(2.18). Let $\boldsymbol{z}_i = (z_{i1}, z_{i2}, \ldots, z_{in})$ be the vector of n observations for position i, and define

$$\hat{h}_i(p_1) = p_1 - \frac{1}{n}\sum_{j=1}^{n} \mathrm{tdr}(z_{ij}, p_1), \tag{6.18}$$

with $\mathrm{tdr}(z, p_1)$ as in (6.2). The value $\hat{\pi}_{i1}$ having $\hat{h}_i(\hat{\pi}_{i1}) = 0$ satisfies

$$\hat{\pi}_{i1} = \frac{1}{n}\sum_{j=1}^{n} \mathrm{tdr}\,(z_{ij}, \hat{\pi}_{i1}) = \frac{1}{n}\sum_{j=1}^{n} \widehat{\mathrm{tdr}}_i(z_{ij}), \tag{6.19}$$

showing that $\hat{\pi}_{i1}$ is the stable point of (2.16). A familiar estimating-equation argument provides an approximate standard error for $\hat{\pi}_{i1}$ (or equivalently for the convergent value of $\hat{k}_i = n\hat{\pi}_{i1}$).

Theorem 4. *The standard deviation of $\hat{\pi}_{i1}$ is approximated by*

$$\mathrm{sd}\,(\hat{\pi}_{i1}) \doteq \pi_{i1}\pi_{i0}/(n\xi_i)^{1/2} \tag{6.20}$$

with ξ_i as in (6.11).

Proof. Only a heuristic derivation of (6.20) will be given here. The random variable $\mathrm{tdr}(z, \pi_{i1})$ has mean (6.6) and standard deviation (6.11),

$$\mathrm{tdr}(z, \pi_{i1}) \sim (\pi_{i1}, \xi_i) \tag{6.21}$$

under the distribution $F_{(i)}$ corresponding to density $f_{(i)}(z)$. Therefore

$$\hat{h}_i(\pi_{i1}) = \pi_{i1} - \frac{1}{n}\sum_{j=1}^{n} \mathrm{tdr}(z_{ij}, \pi_{i1}) \sim (0, \xi_i/n). \tag{6.22}$$

The first Newton–Raphson step to find $\hat{\pi}_{i1}$ gives

$$\hat{\pi}_{i1} - \pi_{i1} = -\hat{h}_i(\pi_{i1})/\hat{h}_i'(\pi_{i1}) \doteq -\hat{h}_i\,(\hat{\pi}_{i1})\,/h_i'(\pi_{i1}) = \frac{\xi_i}{\pi_{i1}\pi_{i0}}\hat{h}_i\,(\hat{\pi}_{i1})\,; \tag{6.23}$$

(6.22) and (6.23) yield

$$\hat{\pi}_{i1} - \pi_{i1} \sim \left(0, (\pi_{i1}\pi_{i0})^2/\xi_i\right), \tag{6.24}$$

which gives (6.20).

The second step in (6.23) substitutes $h_i'(\pi_{i1})$ for $\hat{h}_i'(\pi_{i1})$. Using (6.9),

$$\hat{h}_i'(\pi_{i1}) - h_i'(\pi_{i1}) = \int_{\mathcal{Z}} \frac{\mathrm{tdr}(z, \pi_{i1})\,\mathrm{fdr}(z, \pi_{i1})}{\pi_{i1}\pi_{i0}}\, d\left(F_{(i)} - \hat{F}_{(i)}\right)(z), \tag{6.25}$$

$F_{(i)} - \hat{F}_{(i)}$ being the difference between the true and empirical distributions in $\mathcal{C}_i$. Under standard conditions, this will append a factor of only $1 + O(n^{-1/2})$ to the approximation (6.23). ■

Several relevant points are raised by the previous discussion:

- The key assumption for algorithm (2.15)–(2.18) is that the same likelihood ratio $L(z) = f_0(z)/f_1(z)$ applies to all classes $\mathcal{C}_i$, which is a weaker assumption than $f_0(z)$ and $f_1(z)$ both staying the same. This follows for (6.2) and (6.19).

- The stable point $\hat{\pi}_{i1}$ (6.19) can be found by Newton–Raphson updating, $dp_1 = -\hat{h}_i(p_1)/\hat{h}_i'(p_1)$, (6.18) and (6.8). Theoretically, this should converge faster than the EM-type steps in (2.15)–(2.18).

- The convergence estimates $\hat{k}_i = n\hat{\pi}_i$ were nearly the same as those shown in Figure 3, for example 39.3 compared to 39.1 at position 1755.

- The standard deviation estimates for $\hat{k}_i$ based on the empirical version of (6.20) were a good match to those in Figure 5 for positions having $\hat{k}_i \geq 15$. However, (6.20) gave quite erratic results for $\hat{k}_i < 15$, and is not recommended in general.

- The standard deviation estimate (6.20) equals the Cramér–Rao lower bound at $r = 1$ in parametric family (3.7), but not for $r \neq 1$.

- A possible competitor to $\hat{\pi}_{i1}$ would be

$$\tilde{\pi}_{i1} = \hat{\pi}_1 \hat{r}_i \tag{6.26}$$

  where $\hat{r}_i$ is the maximum likelihood estimate of r in (3.4), (3.7). Example 7 of Efron (1975) implies that $\tilde{\pi}_{i1}$ would be fully efficient at $r = 1$ but far more variable than Fisher information calculations suggest when r much exceeds 1.

- For our cnv example (1.1), the ML estimates $\tilde{\pi}_{i1}$ were a nearly perfect linear function of the converged iterative estimates $\hat{\pi}_{i1}$,

$$\tilde{\pi}_{i1} \doteq 1.06 \cdot \hat{\pi}_{i1}. \tag{6.27}$$

  In other words, the $\hat{k}_i$ estimates of Figure 3 nearly equal MLEs from the class-wise two-groups model (2.1), (2.8).

7 Identifying CNV-Prone Regions in Tumors

Analysis of chromosome copy number aberrations in tumor samples is now a staple of cancer studies. A central question in this paper has been "Which locations are more prone to be gained or lost?" The meaning and motivation of this question in the analysis of tumor samples differs from that in the analysis of normal samples. Since tumorigenesis involves the breakdown of DNA repair and maintenance systems, the accumulation of many chromosomal gains and losses in tumors are hypothesized to be random events that occur as a due effect of the development of the tumor. In this sense, many of the cnas we detect are "passenger" mutations, that, unlike "driver" mutations, do not play a functional role in driving tumor progression. For a recent review, see Stratton, Campbell and Futreal (2009). An important goal in the analysis of tumor samples is to find the driver mutations. Since passenger mutations tend to occur more or less randomly throughout the genome, and driver mutations tend to favor certain genome positions containing functionally relevant genes, driver mutations can be identified by finding positions that are more cna-prone than "random" in a cross-sample analysis. This is the scientific problem that motivates our analysis of tumor samples.

As an example, we analyze chromosome 1 of 207 glioblastoma subjects from the Cancer Genome Atlas project (The Cancer Genome Atlas, 2008). This data is a 42,075 by 207 matrix, derived from the 42,075 probes that map to chromosome 1 on the Illumina HumanHap 550k

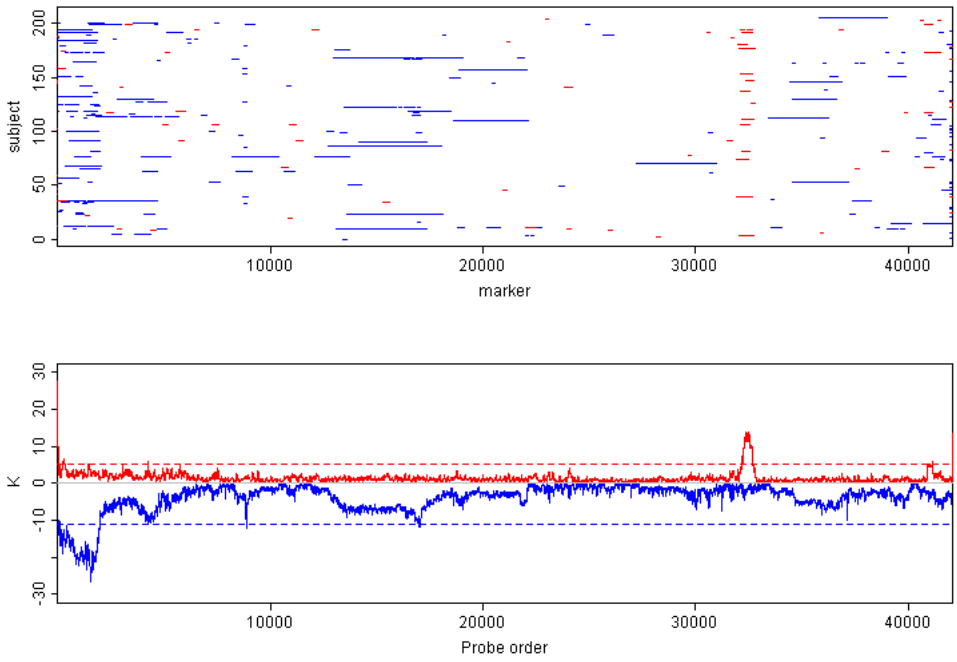

Figure 8: Chromosome 1 of TCGA glioblastoma data. The dash-plot shows the candidate gains and losses by a 0.05 threshold on the local false discovery rate. The bottom plot shows $\hat{k}_i$ profile for gains (blue, positive axis) and losses (red, plotted inverted on the negative axis), with the horizontal dashed line showing the 95% quantile of $\max_k \hat{k}_i$ computed by block bootstrap.

array. We applied the hypothesis-testing framework of Section 3 to identify locations prone to gains and losses. The estimates $\hat{k}_i$ are computed at each location by equation (3.9), which was shown to be a score statistic under a likelihood model. Since gains and losses have completely different biological ramifications in tumors, the two types of changes are treated separately. That is, in computing $\hat{k}_i$, we set the true discovery rate of subjects with $z_{ij} < 0$ to 0 for gains, and vice versa for losses as in (5.9).

The top plot of Figure 8 shows the locations where the one-sided false discovery rates are lower than 0.05; blue for gains and red for losses. The bottom plot shows the $\hat{k}_i$ estimates, plotted in the positive direction for gains and plotted inverted in the negative direction for losses. Here we employed the block bootstrap method on the $\{\text{tdr}_{ij}\}$ matrix (Künsch, 1989) to find the distribution of

$$k^{\max} = \max_i \hat{k}_i$$

(a different calculation from, but in similar spirit to, the one used in Figure 5). Briefly, N/L blocks of size L are sampled with replacement from subject j and concatenated to form $\widehat{\mathbf{tdr}}_j^*$, one bootstrap realization of the $\widehat{\text{tdr}}$ vector for sample j. This is done for $j = 1, \ldots, n$, forming one bootstrap realization $\widehat{\mathbf{tdr}}^*$ of the original matrix. This sampling process effectively eliminates the position alignment of the samples, while for large enough L the local correlation structure across positions for each sample is preserved. Equation (3.9) is applied to the $\widehat{\mathbf{tdr}}^*$ matrix to obtain $\{k_i^*\}$, and $k^{\max,*}$, separately for gains and losses. The horizontal dashed lines in the bottom plot of Figure 8 are the 95% quantiles of the distribution of $k^{\max}$ (separately for gains

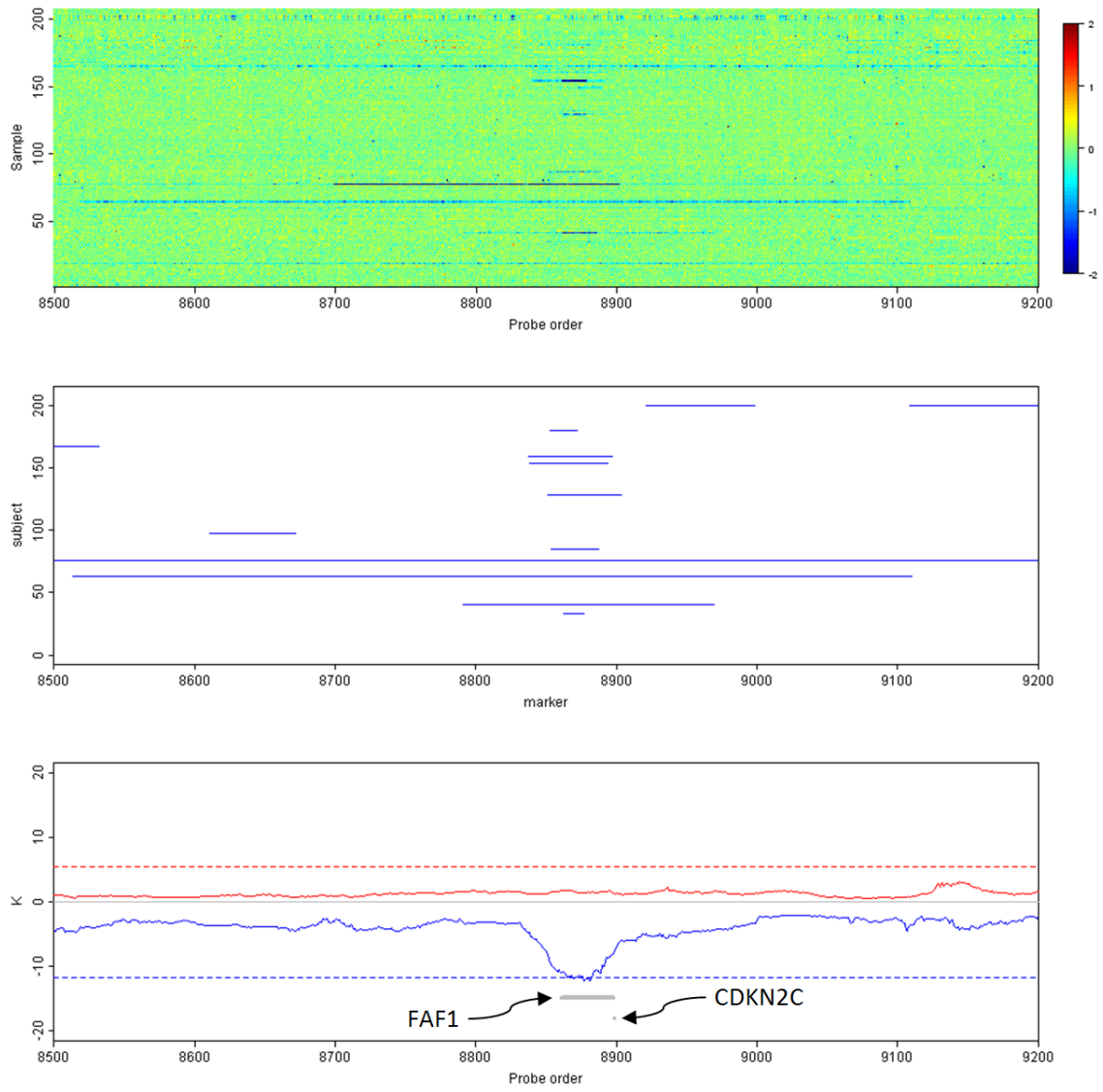

Figure 9: The region between 8500 and 9200 of Chromosome 1 of TCGA glioblastoma data. The heatmap on top shows the Illumina log ratios. The dash-plot in the middle shows the candidate gains and losses by a 0.05 threshold on the local false discovery rate. The bottom plot shows $\hat{k}_i$ profile for gains (blue, positive axis) and losses (red, negative axis), with the horizontal dashed line showing the 95% quantile of $\max_k \hat{k}_i$ computed by block bootstrap. The locations of the genes FAF1 and CDKN2C are shown in the bottom plot.

and losses) estimated from 100 bootstrap samples. The block size of the block-bootstrap was set to 1000, with block sizes between 500 and 10,000 yielding essentially the same results. Since only the tdr values are bootstrapped, not the original x matrix, this procedure assumes that the parameters estimated in the original computation of tdr (before the iterative position-wise fitting) are fixed. This assumption is not unrealistic, since with 42075×207 data points, the parameter estimates of the mixture distribution in the `locfdr` method should have more or less converged to their asymptotic values.

The results in Figure 8 show that, for both gains and losses, there are several significant spikes in the $\{\hat{k}_i\}$ profile. There is one very prominent peak for gains at around 32,000. The signals that contribute to this spike are noticeable in the dash-plot as a column of overlapping dashes. A more interesting case is presented by the less conspicuous spike at around position 9,000 for losses. In Figure 9, we zoom in to the region $[8500, 9200]$ containing this spike. For this 700 marker region, we also show the heatmap of the original data (the x matrix). The color scheme of the heatmap has the baseline value of 0 as green, with losses visible as horizontal streaks of blue. Dark blue streaks, such as the one for sample 77 from around 8700 to around 8900, are evidence for loss of both copies of the chromosome (homozygous deletions). Less conspicuous light blue streaks are evidence for loss of one of the two chromosomal copies (hemizygous deletions). The dash-plot below the heatmap shows those signals that fall below an FDR threshold of 0.05. All of the homozygous deletions seem to be captured at this threshold. Some of the hemizygous deletions fall above this threshold and are thus not shown in the dash-plot, but they contribute to $\hat{k}_i$, which takes on a maximum value of 12 in this region. The dash-plot clearly shows that this region contains overlapping deletions in quite a few of the samples, with $\hat{k}_i$ peaking at the cross sample intersection of the deleted areas.

The peak (or "valley" in the inverted plot) of the $\hat{k}_i$ profile between markers 8800 and 8900 covers the coding regions for two genes: Fas-associated factor 1 (FAF1) and Cyclin-dependent kinase 4 inhibitor C (CDKN2C), the locations of which are marked in the bottom plot of Figure 8. Notice that the width of FAF1 coincide very well with the width of the peak. FAF1 codes for a protein that enhances FAS-induced programmed cell death (apoptosis). It is well known that cells attain uncontrolled growth in tumors by disrupting apoptosis. CDKN2C also encodes a cell growth regulator protein that prevents tumorigenesis. Thus, it is plausible that deletion of the region within probes 8800–8900 (mapping to 50Mb–51Mb on the p-arm of chromosome 1) is an event that plays a driving role in the tumorigenesis of its carriers in this cohort.

8 Local Tests for CNV

Once a genomic region has been identified to contain a candidate cnv by the fdr-based method, other methods can be used to extract more detailed information. At this stage, many questions remain: When multiple nearby locations within a sample fall below the fdr threshold, do they belong to a single contiguous stretch of cnv? If they do, can we accurately estimate the locations of the break-points? For a cnv-prone location, identified by a high $\hat{k}_i$ value, can the carriers be identified more accurately than by thresholding the local fdr? Also, the set of candidates reported by the fdr-based method would no doubt contain false positives. Could we achieve better accuracy by a more detailed follow-up analysis, that examines each candidate cnv and tosses out those that look like imposters? In this section, we seek solutions to these remaining problems. Since our analysis is limited only to those genomic regions that are labeled as

"interesting" by the fdr method, we will call the ensuing analysis "local", as opposed to a "global" analysis covering the entire data matrix.

In the local analysis, we return to the original $\{x_{ij}\}$ matrix of normalized intensity values. The fdr method described in the previous sections works off the matrix of z-values, which were obtained from the normalized data through a smoothing step that averages adjacent probes. The original x matrix, if properly normalized, contains entries that are approximately i.i.d. standard normal under the null hypothesis. An effective normalization procedure based on a low-rank factorization followed by probe-specific standardization is given in Siegmund et al. (2010).

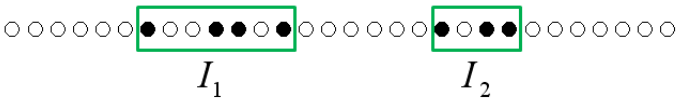

Figure 10: Schematic of a hypothetical region containing 30 positions, with the positions that fall below the fdr threshold marked in black. If we define "nearby" to be ≤ 5 markers, then this region would contain two index sets, I_1 and I_2, as shown.

We consider the situation where a set of nearby positions

$$\mathcal{I} = \{i_1, i_2, \ldots, i_l\}$$

fall below the fdr threshold in a given sample. By "nearby", we mean that they are close enough for us to suspect that they may belong to the same contiguous cnv. Since it is common for cnvs in normal samples to cover 10 kilobases or more, and very uncommon for two *different* cnvs to be separated by less than 10 kilobases, we might define "nearby" to be within 10 kilobases of genomic distance, which equates to 1–10 probes depending on the microarray platform. If a position falls below the fdr threshold, and no nearby positions are significant, then we would have $l = 1$, in which case $\mathcal{I}$ would contain only the single position. We assume that the index set $\mathcal{I}$ can not be expanded further, that is, there is no fdr-significant position in the given sample that is nearby, but that does not belong to $\mathcal{I}$. It is easy to see that the set of all called positions for a given sample can be uniquely partitioned in this way into non-overlapping index sets. An example is shown in Figure 10.

For each index set $\mathcal{I}$ (say, corresponding to a sample j), a change-point model can be used to estimate the location(s) of one or more possible change-points in the genomic region containing $\mathcal{I}$. Let the indices in $\mathcal{I}$ be ordered by genome position, and let $s = i_1 - L$, $t = i_l + L$, where L is a value that is large but much smaller N. We then extract the values $\{x_{s,j}, x_{s+1,j}, \ldots, x_{t,j}\}$ from the x matrix, which we re-name element-wise as $y_1, \ldots, y_T$, $T = t - s + 1$, for convenience. If we were to take a hypothesis-testing approach this step, the null hypothesis that there is actually nothing going on in this region can be formulated as

$$H_0 : y_i \sim \mathcal{N}(0, 1) \qquad i = 1, \ldots, T,$$

with the alternative hypothesis that there is a cnv interval at $[\tau_1, \tau_2)$ formulated as

$$H_A : y_i \sim \mathcal{N}(\mu_i, 1), \qquad \mu_i = \begin{cases} \mu & i = \tau_1, \ldots, \tau_2 - 1 \\ 0 & \text{otherwise.} \end{cases}$$

The parameters μ, τ_1, τ_2 are not known. For some platforms, it has been noted that the noise variance increases for CNV regions, which may motivate the addition of an extra variance term σ^2 to the observations within $[\tau_1, \tau_2)$ under the alternative. However, we have found, as does Olshen et al. (2004) and Wen et al. (2006), that the heterogeneous variance model does not significantly improve detection accuracy.

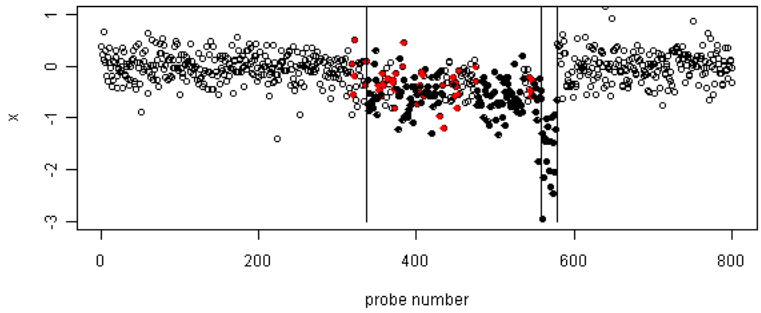

Figure 11: Example of a local cnv analysis. This region contains probes 3400–4200 of sample 41 of the data shown in Figure 1. The vertical axis is the normalized (but unsmoothed) intensity values (the x_{ij}'s). The red points are locations with fdr < 0.05, and the black points are locations with fdr < 0.005. The vertical lines show the change-points determined by the modified BIC method (Zhang and Siegmund, 2007).

Under the above model, the generalized likelihood ratio assuming known τ_1, τ_2 and maximized over μ has the form

$$L(\tau_1, \tau_2) = \sum_{i=\tau_1}^{\tau_2 - 1} y_i / (\tau_2 - \tau_1).$$

Maximizing over τ_1 and τ_2, the generalized likelihood ratio test of H_0 versus H_A is $L(\hat{\tau}_1, \hat{\tau}_2)$, where

$$(\hat{\tau}_1, \hat{\tau}_2) = \text{argmax}_{1 \le \tau_1 < \tau_2 \le M} L(\tau_1, \tau_2).$$

The CBS algorithm of Olshen et al. (2004) and the MBIC method of Zhang and Siegmund (2007) use a similar statistic, but adjusted for an unknown baseline mean. Significance values for tests using $L(\hat{\tau}_1, \hat{\tau}_2)$ are given in Siegmund (2007). Since this region has already passed a global filter based on $\widehat{\text{fdr}}$, a less conservative test is more appropriate for the local analysis. In our experience, thresholds of 0.05 or 0.1, without adjusting for the multiple testing across regions, work well. Since the set of regions reported by the fdr procedure should be heavily enriched for true cnvs, it seems more fitting to treat the analysis of each region as an estimation problem rather than as a testing problem. Instead of asking the question, "Is there a CNV in this region?" we instead ask, "How many break-points does this region contain, and what are their locations?" This framework is especially fitting for index sets that contain multiple cnvs, or complex variants with nested changes. In this sense, the BIC approach described in Yao (1988) and Zhang and Siegmund (2007) seems to be appropriate. The models in Yao (1988) and Zhang and Siegmund (2007) assume that there are m change-points $\tau_1, \ldots, \tau_m$ (m is unknown). The data is assumed Gaussian, with the mean shifting at each change-point, but with the variance remaining constant. While Yao (1988) showed that the traditional BIC

(Schwarz, 1978) is consistent for the estimation of m, Zhang and Siegmund (2007) showed that it is not consistent in estimating the Bayes factor, the quantity that underlies the classic BIC. Zhang and Siegmund (2007) gave a modified form of the BIC, which improves the small-sample accuracy for estimating m.

An example, shown in Figure 11, is the well-known cnv region on the p-arm of chromosome 22. The figure shows the break-points estimated by maximum likelihood under this Gaussian model, with the number of break-points estimated using the modified BIC criterion of Zhang and Siegmund (2007). As indicated by the coloring of points, while most of the locations in the nested homozygous deletion (between 559 and 578) pass the 0.005 fdr threshold, some of the locations in the hemizygous deletion (between 337 and 559) do not even pass the 0.05 threshold. Thus, a local change-point analysis is useful for refining the fdr result and building a more complete picture for each cnv region.

The BIC may report that the candidate region contains no change-points. There would be two possible reasons for this: the region is a false positive, or the signal is so weak that it is missed by the local analysis. Local analysis with a well formulated change-point model should be more powerful than global analyses, because the multiplicity has been much reduced. Thus, if the BIC reports 0 change-points, we conclude that the region is a false positive.

References

Beroukhim, R., Getz, G., Nghiemphu, L., Barretina, J., Hsueh, T., Linhart, D., Vivanco, I., Lee, J. C., Huang, J. H., Alexander, S., Du, J., Kau, T., Thomas, R. K., Shah, K., Soto, H., Perner, S., Prensner, J., Debiasi, R. M., Demichelis, F., Hatton, C., Rubin, M. A., Garraway, L. A., Nelson, S. F., Liau, L., Mischel, Cloughesy, T. F., Meyerson, M., Golub, T. A., Lander, E. S., Mellinghoff, I. K. and Sellers, W. R. (2007). Assessing the significance of chromosomal aberrations in cancer: Methodology and application to glioma. *Proc. Natl. Acad. Sci. USA* : 0710052104+.

Bignell, G. R., Huang, J., Greshock, J., Watt, S., Butler, A., West, S., Grigorova, M., Jones, K. W., Wei, W., Stratton, M. R., Futreal, P. A., Weber, B., Shapero, M. H. and Wooster, R. (2004). High-resolution analysis of DNA copy number using oligonucleotide microarrays. *Genome Res.* 14: 287–295.

Conrad, D., Andrews, T., Carter, N., Hurles, M., and Pritchard, J. (2006). A high-resolution survey of deletion polymorphism in the human genome. *Nat. Genet.* 38: 75–81.

Diskin, S. J., Eck, T., Greshock, J., Mosse, Y. P., Naylor, T., Stoeckert Jr., C. J., Weber, B. L., Maris, J. M. and Grant, G. R. (2006). Stac: A method for testing the significance of DNA copy number aberrations across multiple array-cgh experiments. *Genome Res.* 16: 1149–1158.

Efron, B. (1975). Defining the curvature of a statistical problem (with applications to second order efficiency). *Ann. Statist.* 3: 1189–1242, with a discussion by C. R. Rao, Don A. Pierce, D. R. Cox, D. V. Lindley, Lucien LeCam, J. K. Ghosh, J. Pfanzagl, Niels Keiding, A. P. Dawid, Jim Reeds and with a reply by the author.

Efron, B. (2008). Microarrays, empirical Bayes and the two-groups model. *Statist. Sci.* 23: 1–22.

Efron, B. (2009). Empirical bayes estimates for large-scale prediction problems. *J. Amer. Statist. Assoc.* 104: 1015–1028.

Efron, B. (2010a). Correlated z-values and the accuracy of large-scale statistical estimates. *J. Amer. Statist. Assoc.* To appear.

Efron, B. (2010b). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*, Institute of Mathematical Statistics Monographs I. Cambridge: Cambridge University Press, to be published August 2010.

Guttman, M., Mies, C., Dudycz-Sulicz, K., Diskin, S. J., Baldwin, D. A., Stoeckert, C. J. and Grant, G. R. (2007). Assessing the significance of conserved genomic aberrations using high resolution genomic microarrays. *PLoS Genetics* 3: e143+.

Künsch, H. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* 17: 1217–1241.

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

McKernan, K. J., Peckham, H. E., Costa, G. L., McLaughlin, S. F., Fu, Y., Tsung, E. F., Clouser, C. R., Duncan, C., Ichikawa, J. K., Lee, C. C., Zhang, Z., Ranade, S. S., Dimalanta, E. T., Hyland, F. C., Sokolsky, T. D., Zhang, L., Sheridan, A., Fu, H., Hendrickson, C. L., Li, B., Kotler, L., Stuart, J. R., Malek, J. A., Manning, J. M., Antipova, A. A., Perez, D. S., Moore, M. P., Hayashibara, K. C., Lyons, M. R., Beaudoin, R. E., Coleman, B. E., Laptewicz, M. W., Sannicandro, A. E., Rhodes, M. D., Gottimukkala, R. K., Yang, S., Bafna, V., Bashir, A., MacBride, A., Alkan, C., Kidd, J. M., Eichler, E. E., Reese, M. G., De La Vega, F. M. and Blanchard, A. P. (2009). Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. *Genome Res.* 19: 1527–1541.

Mills, R. E. E., Luttig, C. T. T., Larkins, C. E. E., Beauchamp, A., Tsui, C., Pittard, W. S. S. and Devine, S. E. E. (2006). An initial map of insertion and deletion (indel) variation in the human genome. *Genome Res* 16: 1182–1190.

Newton, M., Gould, M., Reznikoff, C. and Haag, J. (1998). On the statistical analysis of allelic-loss data. *Stat. Med.* 17: 1425–1445.

Newton, M. 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. B., Venkatraman, E. S., Lucito, R. and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5: 557–572.

Peiffer, D. A., Le, J. M., Steemers, F. J., Chang, W., Jenniges, T., Garcia, F., Haden, K., Li, J., Shaw, C. A., Belmont, J., Cheung, S. W., Shen, R. M., Barker, D. L. and Gunderson, K. L. (2006). High-resolution genomic profiling of chromosomal aberrations using infinium whole-genome genotyping. *Genome Res.* 16: 1136–1148.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W. L., Chen, C., Zhai, Y., Dairkee, S. H., Ljung, B. M., Gray, J. W. and Albertson, D. G. (1998). High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. *Nat. Genet.* 20: 207–11.

Pollack, J., Perou, C., Alizadeh, A., Eisen, M., Pergamenschikov, A., Williams, C., Jeffrey, S., Botstein, D. and Brown, P. (1999). Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* 23: 41–46.

Robbins, H. (1956). An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I*. Berkeley and Los Angeles: University of California Press, 157–163.

Rouveirol, C., Stransky, N., Hupé, P., La Rosa, P., Viara, E., Barillot, E. and Radvanyi, F. (2006). Computation of recurrent minimal genomic alterations from array-CGH data. *Bioinformatics* 22: 849–856.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* 6: 461–464.

Siegmund, D., Yakir, B. and Zhang, N. (2010). Detecting simultaneous variant intervals in aligned sequences, submitted.

Siegmund, D. O. (2007). Approximate tail probabilities for the maxima of some random fields. *Ann. Probab.* 16: 487–501.

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

Stratton, M. R., Campbell, P. J. and Futreal, P. A. (2009). The cancer genome. *Nature* 458: 719–724.

Taylor, B. S., Barretina, J., Socci, N. D., Decarolis, P., Ladanyi, M., Meyerson, M., Singer, S. and Sander, C. (2008). Functional copy-number alterations in cancer. *PLoS ONE* 3: e3179+.

The Cancer Genome Atlas (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455: 1061–1068.

Wang, P., Kim, Y., Pollack, J., Narasimhan, B. and Tibshirani, R. (2005). A method for calling gains and losses in array-CGH data. *Biostatistics* 6: 45–58.

Wen, C., Wu, Y., Huang, Y., Chen, W., Liu, S., Jiang, S., Juang, J., Lin, C., Fang, W., Hsiung, C. and Chang, I. (2006). A Bayes regression approach to array-CGH data. *Stat. Appl. Mol. Biol.* 5.

Willenbrock, H. and Fridlyand, J. (2005). A comparison study: Applying segmentation to array-CGH data for downstream analyses. *Bioinformatics* 21: 4084–4091.

Yao, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Stat. Probab. Lett.* 6: 181–189.

Zhang, N. (2010). DNA Copy Number Profiling in Normal and Tumor Genomes. In Fu, W. (ed.), *Probability and Statistics and Their Applications to Biology*. Springer-Verlag.

Zhang, N. and Siegmund, D. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* 63: 22–32.

Zhang, N., Siegmund, D., Ji, H. and Li, J. Z. (2010). Detecting simultaneous change-points in multiple sequences. *Biometrika* In press.