Submitted to the Annals of Applied Statistics

THE SCREENING AND RANKING ALGORITHM TO DETECT DNA COPY NUMBER VARIATIONS

By Yue S. Niu[†], and Heping Zhang[‡]

University of Arizona[†] *and Yale University*[‡]

*The financial supports from University of Arizona Internal grant and National Institute on Drug Abuse grant R01-DA016750 are greatly acknowledged.

APPENDIX A: GENERAL WEIGHT FUNCTIONS

In this section, we consider a family of local diagnostic functions, which is the weighted average of Y_i's near the point of interest x

$$D(x,h) = \sum_{i=1}^{n} w_i(x) Y_i,$$

where $w_i(x) = w_{i-x}$, and $\mathbf{w} = (\cdots, w_{-1}, w_0, w_1, \cdots)$, satisfying the following conditions:

1. local supportedness: $\exists$ an integer $h \ll n$, such that $w_i = 0$ for $|i| > h$;
2. quasi-symmetry: $i \cdot w_i \leq 0$ and $\sum_{i \leq 0} w_i = -\sum_{i>0} w_i$ i.e. $\sum w_i = 0$;
3. unity: $\sum_{i \leq 0} \mid w_i \mid = \sum_{i>0} \mid w_i \mid = 1$, and hence $\parallel \mathbf{w} \parallel_{\ell_1} = \sum \mid w_i \mid = 2$;
4. negligibility: $\parallel \mathbf{w} \parallel_{\ell_2}^2 / \parallel \mathbf{w} \parallel_{\ell_1}^2 = \sum w_i^2/4 = O(h^{-1})$.

We denote by $\mathcal{W}$ the set of all weight vectors satisfying these four conditions. These four conditions are quite natural. The locally supported condition makes $D(x)$ depend on only those Y_i's within distance h. The quasi-symmetric condition ensures that $D(x)$ measures the difference between the left-hand-side Y_i's and right-hand-side Y_i's. The unity condition is not essential, but helpful for easy presentation. The negligible condition, a little stronger than the traditional negligible condition, prevents the weights from concentrating on few points as the bandwidth h tends to infinity. It is easy to see that all weights introduced in Section 2.2, up to a normalizing constant, are special cases of this family. Moreover, the SaRa with any local diagnostic function in this family satisfies the sure coverage property.

APPENDIX B: PROOFS

We shall prove Theorem 1 in three steps, represented by three lemmas. We introduce the notation and outline the proof first.

A point x is called h-flat if there is no change-point in the h-neighborhood of x, i.e. the interval $(x-h, x+h)$. We omit h and say x is a flat point if h is obvious in the context. Let $\mathcal{F}_h$ be the set of all h-flat points of step function μ. Consider the event $\mathcal{A}_\tau = \left\{ |D(\tau,h)| > \lambda \right\}$ for change-point $\tau \in \mathcal{J}$ and the event $\mathcal{B}_x = \left\{ |D(x,h)| < \lambda \right\}$ for flat point $x \in \mathcal{F}_h$. Define the event

$$\mathcal{E}_n = \left(\bigcap_{\tau \in \mathcal{J}} \mathcal{A}_\tau \right) \bigcap \left(\bigcap_{x \in \mathcal{F}_h} \mathcal{B}_x \right).$$

In Lemma 1, we derive the distribution of $D(x,h)$ at a given point x when there is no change-point other than possibly x in the interval $(x-h, x+h)$.

Then we calculate the probability $\mathbf{P}(\mathcal{E}_n)$ in Lemma 2. In the final step, we show that $\mathcal{J} \subset: \hat{\mathcal{J}} \pm h$ holds under the event $\mathcal{E}_n$.

Lemma 1 *If the noises are i.i.d. Gaussian, then for fixed x and h, $D(x,h)$ is Gaussian. In particular, if x is a flat point, $D(x,h) \sim \mathcal{N}(0,\Delta^2)$. If τ is a change-point with jump size δ, $D(\tau,h) \sim \mathcal{N}(\delta,\Delta^2)$. Here,*

$$\Delta^2 = \sum_i w_i^2 \sigma^2 = O(h^{-1})\sigma^2.$$

Proof of Lemma 1. $D(x,h)$ is a linear combination of Gaussian variables, so it is Gaussian as well. It follows from the quasi-symmetric and unity conditions that the mean of $D(x,h)$ is zero for a flat point and δ for a change-point with jump size δ. The variance is $\sum_i w_i^2\sigma^2$, which is of order $O(h^{-1})\sigma^2$ by the condition 4 on the family $\mathcal{W}$. In particular, for the equally weighted case (**??**), $\Delta^2 = \frac{2}{h}\sigma^2$. □

Lemma 2 *Under Assumption (**??**), there exist h and λ such that*

$$P(\mathcal{E}_n) \to 1 \text{ as } n \to \infty. \tag{B.1}$$

Proof of Lemma 2. It suffices to show that there exist λ and h such that

$$\mathbf{P}(\mathcal{E}_n^c) \le \mathbf{P}\left\{\bigcup_{\tau\in\mathcal{J}} \mathcal{A}_\tau^c\right\} + \mathbf{P}\left\{\bigcup_{x\in\mathcal{F}_h} \mathcal{B}_x^c\right\} \to 0.$$

Take $\lambda = \frac{1}{2}\delta$ and $h = \frac{1}{2}L$ where $\delta = \min|\delta_j|$, $L = \min_{1\le j\le J+1}(\tau_j - \tau_{j-1})$. By Lemma 1, it is obvious that for each $\tau \in \mathcal{J}$ and $x \in \mathcal{F}_h$, $\mathbf{P}(\mathcal{A}_\tau^c) < 1-\Phi(\frac{\delta}{2\Delta})$ and $\mathbf{P}(\mathcal{B}_x^c) = 2(1-\Phi(\frac{\delta}{2\Delta}))$, where Φ is the cumulative distribution function of standard normal distribution. Note the following inequality for the Gaussian tail probability (**?**)

$$1-\Phi(t) < t^{-1}e^{-\frac{1}{2}t^2}.$$

By Bonferroni inequality and $\Delta = \sqrt{2/h}\sigma = 2\sigma/\sqrt{L}$, we have

$$\mathbf{P}(\mathcal{E}_n^c) < 2n\frac{2\Delta}{\delta}e^{-\frac{\delta^2}{8\Delta^2}} = \frac{8n\sigma}{\delta\sqrt{L}}e^{-\frac{L\delta^2}{32\sigma^2}} = \frac{8n}{S}e^{-\frac{S^2}{32}}. \tag{B.2}$$

It is guaranteed by Assumption (**??**) that the right hand side of (B.2) goes to zero as $n \to \infty$. □

Lemma 3 *$\mathcal{J} \subset: \hat{\mathcal{J}} \pm h$ holds under event $\mathcal{E}_n$.*

Proof of Lemma 3. We want to show that there is a one-to-one correspondence between $\mathcal{J}$ and $\hat{\mathcal{J}}$. Under event $\mathcal{E}_n$, no flat points can be selected into $\hat{\mathcal{J}}$ at the screening step. In other words, for any point $\hat{\tau} \in \hat{\mathcal{J}}$, there is at least one change-point in its h-neighborhood $(\hat{\tau} - h, \hat{\tau} + h)$. In fact, there is at most one such change-point by our assumption that $L = 2h$. Therefore, there is exactly one change-point within $(\hat{\tau} - h, \hat{\tau} + h)$ for each $\hat{\tau} \in \hat{\mathcal{J}}$. On the other hand, under event $\mathcal{E}_n$, for every change-point $\tau \in \mathcal{J}$, we have $|D(\tau, h)| > \lambda$. Moreover, $\tau - h$ and $\tau + h$ must be flat points since $L = 2h$. It follows that $\max\left\{|D(\tau - h, h)|, |D(\tau + h, h)|\right\} < \lambda$ and there is a local maximum, say $\hat{\tau}$, which is in $(\tau - h, \tau + h)$ and $|D(\hat{\tau}, h)| \geq |D(\tau, h)| > \lambda$. □

Dr. Yue S. Niu
Department of Mathematics
The University of Arizona
Tucson AZ, 85721
E-mail: yueniu@math.arizona.edu

Professor Heping Zhang
Department of Epidemiology and
Public Health, Yale University
School of Medicine, New Haven CT, 06520
E-mail: heping.zhang@yale.edu