Joint Estimation of DNA Copy Number from Multiple Platforms

Nancy R. Zhang, Yasin Senbabaoglu, and Jun Z. Li

Supplementary Appendices

1 Derivation of the Likelihood Ratio Statistic (3.3)

To show that the likelihood ratio statistic gives (3.3): For simplicity of notation we suppress the location indices $[s,t]$. Since this is a Gaussian mean shift model, the log likelihood ratio between H_A and H_0 is

$$l_A(\mu) - l_0 = \sum_{k=1}^{K} [\mu \delta_k X_k/\sigma_k - \mu^2 \delta_k^2/(2\sigma_k^2)]. \tag{1.1}$$

Differentiating the above with respect to μ and setting the derivative to 0, we get $\hat{\mu} = \tilde{\delta}' X/\tilde{\delta}'\tilde{\delta}$, where $\tilde{\delta} = (\delta_1/\sigma_1, \ldots, \delta_K/\sigma_K)$. Plugging this value back into (1.1), we have $l_A(\hat{\mu}) - l_0$ equals $(\tilde{\delta}' X)^2/(2\tilde{\delta}'\tilde{\delta})$, which is one-half of (3.3).

2 Pseudo-code for MPCBS Segmentation Algorithm

Initialize:

Set $k \leftarrow 0$, $S_0 \leftarrow \{0, T\}$,

$$Z_{\max} = \max_{0<i<j<T} Z(i,j), \quad (s^*, t^*) = \arg\max_{0<i<j<T} Z(i,j),$$

Set $\mathcal{Z} \leftarrow Z_{\max}$, $\mathcal{R} \leftarrow (s^*, t^*)$, $BIC(0) \leftarrow 0$.

While $|S_k| - 2 < M$ **repeat:**

1. Let $i^* \leftarrow \arg\max_i \mathcal{Z}[i]$, $(s^*, t^*) \leftarrow \mathcal{R}[i^*]$,

   $$s \leftarrow \max\{i \in S_k, i < s^*\}, \quad t \leftarrow \min\{i \in S_k, i > t^*\}.$$

   For each of $(i,j) \in \{[s, s^*), [s^*, t^*), [t^*, t)\}$, compute

   $$Z_{\max} = \max_{i<a<b<j} Z(a,b), \quad (s^*, t^*) = \arg\max_{i<a<b<j} Z(a,b).$$

   Let Z_L, Z_C, and Z_R be respectively the value of $Z_{\max}$ computed for the left segment $[s, s^*)$, the center segment $[s^*, t^*)$, and the right segment $[t^*, t)$. Similarly, let R_L, R_C, R_R be respectively the maximizer for the left, center, and right segments.

2. Let $L = |\mathcal{Z}|$, Set:

   $$k \leftarrow k + 1,$$
   $$S_k \leftarrow S_{k-1} \bigcup \{s^*, t^*\},$$
   $$\mathcal{Z} \leftarrow \{\mathcal{Z}[1 : i^* - 1], Z_L, Z_C, Z_R, \mathcal{Z}[i^* + 1, L]\},$$
   $$\mathcal{R} \leftarrow \{\mathcal{R}[1 : i^* - 1], R_L, R_C, R_R, \mathcal{R}[i^* + 1, L]\}.$$

   Set $BIC(k)$ to be the BIC criterion (**??**) of the estimated change-points S_k.

Finally, let $k^* = \arg\max_{0 \le k \le M} BIC(k)$**. Return** S_{k^*}**.**

3 Block-update procedure for estimating platform response ratio

Let K be the number of platforms, m be the number of regions. We are fitting

$$\frac{1}{2}\sum_{i=0}^{m}\sum_{k=1}^{K} v_{k,i}^{-1}(\hat{f}_{i,k} - \alpha_k - r_k\theta_i)^2 \tag{3.1}$$

with the response ratio r_K for platform K constrained to be 1.

Initialize $t \leftarrow 0$,

$$r^0 \leftarrow (1, \ldots, 1)_{1\times K},$$
$$\alpha^0 \leftarrow (0, \ldots, 0)_{1\times K}.$$

Repeat:

1. $t \leftarrow t + 1$
2. Given r^{t-1}, estimate by weighted least squares

$$\theta^t \leftarrow \arg\min_{\theta} \sum_{i=0}^{m}\sum_{k=1}^{K} v_{k,i}^{-1}(\hat{f}_{i,k} - \alpha_k^t - r_k^{t-1}\theta_i)^2.$$

3. Given θ^t, estimate by weighted least squares

$$(\alpha_{1:K-1}^t, r_{1:K-1}^t) \leftarrow \arg\min_{\alpha,r} \sum_{i=0}^{m}\sum_{k=1}^{K-1} v_{k,i}^{-1}(\hat{f}_{i,k} - \alpha_k - r_k\theta_i^t)^2.$$

4. For the K-th platform, keep r_K^t at 1 and set $\alpha_K^t \leftarrow \sum_{i=1}^{m}(\hat{f}_{i,K} - \theta_i)$.
5. If $\|r^t - r^{t-1}\|/m < \epsilon$ exit loop.

Report $r = r^t$, $\theta = \theta^t$, $\alpha = \alpha^t$.

4 Normalization of Hapmap samples

For Affymetrix data, we requested GeneChip 6.0 CEL files for HapMap samples from Affymetrix, InC. We used the software package Aroma to analyze preprocess the eight HapMap samples used in this study. The resulting logR values were used in CBS and MPCBS analysis without further normalization. For Illumina Human1M-Duo Beadchip data, we downloaded from Illumina's public FTP site the logR values for the 270 HapMap samples. We first median-centered logR values for each sample and in each chromosome. To correct for the long range oscillations in baseline logR levels (a phenomenon known as "genomic waves", which is related to local GC content and whole-genome amplification conditions prior to array hybridization), we developed a normalization procedure based on a principal component analysis (PCA) in the entire cohort of 270 samples. Each sample is assessed PCA scores that correspond to the magnitude and phase of the "wave" for that sample. For each SNP we performed a linear regression of logR values in all samples against the samples' first three PCA scores. The residuals from the regression were taken as logR values corrected for the "genomic waves", and used for segmentation by CBS or MPCBS.