Submitted to the Annals of Applied Statistics

SUPPLEMENTARY MATERIAL FOR "COMPUTATIONALLY EFFICIENT WHOLE-GENOME SIGNAL REGION DETECTION FOR QUANTITATIVE AND BINARY TRAITS"

BY FAN WANG[1,a] WEI ZHANG[2,b], AND FANG YAO[2,c]

[1]*Department of Biostatistics, Mailman School of Public Health, Columbia University,* [a]*fw2400@cumc.columbia.edu*

[2]*School of Mathematical Sciences, Center for Statistical Science, Peking University,* [b]*wei_zhang@stu.pku.edu.cn;* [c]*fyao@math.pku.edu.cn*

In this Supplementary Material, we present the theoretical properties, key lemmas and the proofs of the lemmas and theorems in our work. In addition, all codes and instructions for implementing simulations and real data analysis are included in the file "dBiRS_Code". The codes are also accessible in GitHub repository https://github.com/ZWCR7/dBiRS.

S1. Theoretical Analysis. Detailed theoretical results on FWER control and detection accuracy are presented as follows.

S1.1. *Family-wise error rate control.* Recall that $U_n = \boldsymbol{G}^\top (Y - \hat{\eta}_0)/\sqrt{n}$ is the vector of marginal scores and the test statistic is the maxima of marginal scores. Under the null hypothesis, $Y - \hat{\eta}_0$ is asymptotically distributed as $\mathcal{N}(0, P)$, where P is the projection matrix. Let $Z \sim \mathcal{N}(0, P)$, under some regularity assumptions, the following theorem claims that the proposed global test controls the size and consequently, the dBiRS algorithm controls the FWER.

THEOREM 1. *Assume that the following conditions (a)-(b) hold:*
(a) For $i = 1, \ldots, n$*, the function* $w_i(x_0, x_1) = 1/a_i(x_0)\nu(x_1)$ *has finite gradient on a neighborhood of* (ϕ, η_i)*;*
(b) $\log^4(p)/n \to 0$ *as* $n \to \infty$.
Then under the null hypothesis, with probability one, we have

$$|\mathbb{P}(\|U_n\|_\infty > \tilde{c}(\alpha)) - \alpha| \to 0,$$

and consequently, the dBiRS procedure controls the FWER.

Condition (a) is used for consistently estimating Σ by $\hat{\Sigma}$ and is common in asymptotic analysis for GLM. Condition (b) indicates that the data dimension p can grow exponentially in n. Recall that the sBiRS algorithm begins with the global test, then the sBiRS procedure controls the FWER while the size of the global test is controlled. Furthermore, since we perform a sBiRS procedure for blocks in dBiRS algorithm, the FWER of dBiRS is the same as sBiRS, which indicates that the dBiRS procedure also controls the FWER.

S1.2. *Detection accuracy analysis.* Now we concentrate on the accuracy of the detection algorithm in terms of the proportion of true discoveries and false discoveries. Before the rigorous power analysis of our algorithm, we assume genotype data is bounded and centralized and introduce some notations. Given a truncation parameter s, we can split the

Keywords and phrases: family-wise error rate, signal detection, distributed learning, whole genome association studies.

global region into several continuous segments such that the length of each segment is less than 2^s and at least one segment has a length greater than 2^{s-1}. Note that the split is not unique, and let $\mathcal{S} = \{I_1, \ldots, I_S\}$ be the set containing these segments of any split. The regions $I_1, \ldots, I_S$ are neighboring and non-overlapped satisfying $\cup_{j=1}^{K} I_j = \{1, \ldots, p\}$. For a true signal region $I_r^*, r = 1, \ldots, \ell$, we refer to $I_{i_r}, \ldots, I_{i_r+k_r}$ as the minimum cover of I_r^* in $\mathcal{S}$ if $I^* \subset \cup_{j=0}^{k} I_{i+j}$ and $\left|\cup_{j=0}^{k-1} I_{i+j}\right| < |I^*| \leq \left|\cup_{j=0}^{k} I_{i+j}\right|$. Finally, without loss of generality, we suppose that $I_1^*, \ldots, I_\ell^*$ with lengths $p_1, \ldots, p_\ell$ have decayed signals, that is, $\left\|\mu_{I_1^*}\right\|_\infty \geq \left\|\mu_{I_2^*}\right\|_\infty \geq \cdots \geq \left\|\mu_{I_\ell^*}\right\|_\infty$, where $\mu = \mathbb{E}(U)/\sqrt{n}$.

Providing a decay signal strength assumption for the true signal regions, we have the following theorem which gives the approximation results of true discoveries.

THEOREM 2. *Assume that the condition (a)-(b) in Theorem 1 hold, and further assume that*
(c) There exists a sufficiently large constant $C_1 > 0$, for $u = 0, \ldots, k_r$, $r = 1, \ldots, \ell$,

$$\left\|\mu_{I_{i_r+u}}\right\|_\infty \geq C_1 \left[\log\left\{\left(p - \sum_{v=1}^{r-1} p_v\right) n\right\}/n\right]^{1/2}.$$

Let P_ℓ denotes the probability that our dBiRS algorithm detects all ℓ signal regions, when $n \to \infty$,

$$P_\ell \geq 1 - 4\log(p+1)/n^2 \to 1.$$

The result of this theorem states that in both the balanced and decaying signal settings, the dBiRS-detected signal segments can consistently cover the true signal regions. Condition (c) implies that the lower bound of $\left\|\mu_{I_1^*}\right\|_2^2$ for consistent detection is of order $\log(p)\,|I_1^*|\,/2^s/n$. The lower bound in Theorem 3 of Q-SCAN (Li, Liu and Lin, 2020) is weaker than this, however, they assume that the maximum eigenvalue of $\hat{\Sigma}_{I_1^*}$ has a fixed upper bound, which is overly restricted. Furthermore, to the best of our knowledge, condition (c) is the first to allow signal decay settings in GLM model, while the existing work (Jeng, Cai and Li, 2010; Li, Liu and Lin, 2020) assumes that the signal strengths of all regions have a lower bound and Zhang, Wang and Yao (2025) adopts this decay setting under the two-sample test framework.

Next, we focus on the false discovery rate of the dBiRS algorithm and illustrate the consistent detection property of our algorithm. To begin with, we introduce the Jaccard index to quantify the similarity between two regions. Specifically, we define the Jaccard index between sets I_1 and I_2 as

$$J(I_1, I_2) = |I_1 \cap I_2| \,/\, |I_1 \cup I_2|\,.$$

Recall that $\mathcal{I}^* = \{I_1^*, \ldots, I_\ell^*\}$ are the set of true signal regions and $\hat{\mathcal{I}} = \left\{\hat{I}_1, \ldots, \hat{I}_{\ell'}\right\}$ are the estimated signal regions. For a signal region I_i^*, we define that it is consistently detected if for some $\eta(p) = o(1)$, there exists $\hat{I}_{j_i} \in \hat{\mathcal{I}}$ such that

$$\mathbb{P}\left\{J(\hat{I}_{j_i}, I_i^*) \geq 1 - \eta(p)\right\} \to 1,$$

as $p \to \infty$. Note that, even if every signal region is consistently detected, there still could be some additional regions that are incorrectly detected. Let $\tilde{I} = \hat{I} - I^*$ as the set of wrongly detected regions, and we want such regions to be ignorable relative to the true signal regions, i.e., $\left|\tilde{I}\right| / |I^*| \to 0$. Then we say that a detection procedure consistently detected all the true

signal regions, if for a sequence of $\eta_j(p) = o(1), j = 1, \ldots, \ell$ and some $\eta(p) = o(1)$, there exists $\hat{I}_{j_1}, \ldots, \hat{I}_{j_\ell} \in \hat{\mathcal{I}}$ such that

$$\mathbb{P}\left[\left\{\left|\tilde{I}\right| / |I^*| \leq \eta(p)\right\} \cap A\right] \to 1$$

, where $A = \cap_{i=1}^{\ell} \left\{J(\hat{I}_{j_i}, I_i^*) \geq 1 - \eta_j(p)\right\}$.

Under some mild conditions concentrating on the structure of the covariance matrix Σ and the distribution of true signal regions, we derive the detection consistency of dBiRS algorithm in the following theorem.

THEOREM 3. *Assume that conditions (a)-(c) hold. We further assume that*
(d) There exists certain truncation parameter s that results in the set of regions $\mathcal{S} = \{I_1, \ldots, I_S\}$ such that for some $\tilde{\alpha} > 2\alpha$, $\mathcal{O} = \{1, 2, \ldots, S\}$,

$$D(s, R) = \sup_{i_1, \ldots, i_R \in \mathcal{O}} \frac{\mathbb{P}(\cap_{j=1}^{R} A_{i_j})}{\prod_{j=1}^{R} \mathbb{P}(A_{i_j})} \leq \frac{1}{(2\tilde{\alpha})^R},$$

where the events $A_i = \{\|U_n(I_i)\|_\infty \geq c_{I_i}(\alpha)\}$ and $c_{I_i}(\alpha)$ is the $1-\alpha$ quantile of $\|U_n(I_i)\|_\infty$;
(e) The true signal regions are well-separated in the sense that $Gap_{\min} \geq L_{\max}$, where $L_{\max}$ is the maximum length of all true continuous signal regions and $G_{\min}$ is the minimum length of the gaps between any two true continuous signal regions;
(f) There exists a truncation parameter s satisfies (d) and $s = o\{\log_2(L_{\min}/\log \ell)\}$;
(g) There exists $r = o(L_{\min})$ such that for any index $i \in \cup_{l=1}^{\ell} I_l^*$, the non-signal point j satisfies that if $|j - i| > r$, then

$$\left|\boldsymbol{G}_{\cdot j}^{\top} \boldsymbol{G} \beta\right| / \sqrt{n} = o(1).$$

Denote $h_i = \lceil |I_i^*| / 2^s \rceil$ as the cardinality of the minimum cover (from $\mathcal{S}$) of I_i^*, $i = 1, \ldots, \ell$ and $\tilde{r} = \lceil r/2^s \rceil$. Given a significance level α, when $n, p \to \infty$, for any sequence of integers $R_i = o(h_i)$ and $R_i / \log \ell \to \infty$, $i = 1, \ldots, \ell$, there exists $\hat{I}_{j_1}, \ldots, \hat{I}_{j_\ell} \in \hat{\mathcal{I}}$ such that,

$$\mathbb{P}\left[\bigcap_{i=1}^{\ell} \left\{J\left(I_i^*, \hat{I}_{j_i}\right) \geq 1 - \eta_i\right\}\right] \geq 1 - \delta,$$

where $\eta_i = (R_i + 2\tilde{r})/(R_i + h_i + 1) \to 0$ and $\delta = 4\log(p+1)/n^2 + \sum_{i=1}^{\ell} (\alpha/2\tilde{\alpha})^{R_i} \to 0$.

Let $A = \cap_{i=1}^{\ell} \left\{J(\hat{I}_{j_i}, I_i^*) \geq 1 - \eta_j\right\}$, we have that for some integer $R_0 = O(K_1)$ and $R_0 = o(\sum_{i=1}^{\ell} h_i)$,

$$\mathbb{P}\left[\left\{\left|\tilde{I}\right| / |I^*| \leq \eta_0\right\} \cap A\right] \geq (1 - \delta)\left[1 - C_2\left\{\frac{2R_0\alpha}{(R_0 - K_1)\tilde{\alpha}}\right\}^{R_0 - K_1}\right] \to 1, \tag{S1}$$

where $\eta_0 = R_0 / \sum_{i=1}^{\ell} h_i \to 0$, $K_1 \leq K$ is the number of blocks which have signals and C_2 is a constant.

To appreciate this result, we see that $1 - 4\log(p+1)/n^2$ quantifies the probability that there exist estimated signal regions simultaneously covering the true ones as Theorem 2 asserts, while $(\alpha/2\tilde{\alpha})^{R_i}$ indicates how many non-signal segments in $\mathcal{S}$ may be falsely included neighboring to I_i^* resulting from the re-search procedure. The term $1 - C_2\left\{\frac{2R_0\alpha}{(R_0-K_1)\tilde{\alpha}}\right\}^{R_0-K_1}$ in inequality (S1) is the lower bound of the probability that R_0 additional non-signal segments are falsely detected and it relates to the number of blocks K.

Theorem 3 demonstrates that the consistent detection property can be achieved by the dBiRS algorithm under mild conditions. Condition (d) relaxes the common M-dependence assumption in Li, Liu and Lin (2020) and only requires a "weak dependence" assumption under which the Jaccard index consistency of the dBiRS algorithm is still guaranteed. It is straightforward to verify that the M-dependence satisfies condition (d) if we choose the truncation parameter $s \geq \log_2 M$. Moreover, condition (d) includes a larger class of covariance matrices. For instance, the covariance $\Sigma = \left\{(1+|j-k|)^{-\rho}\right\}_{j,k=1}^{p}$, $\rho > 1/2$ and $\Sigma = \left\{\theta^{|j-k|}\right\}_{j,k=1}^{p}, \theta < 1$ satisfy condition (d) but is not M-dependent. Condition (e) imposes the requirement on the distances among the signal regions. The constraint $2^s = o\left(L_{\min}/\log \ell\right)$ in condition (f) guarantees that the union set of the minimum cover of a true signal region is as short as possible. Conditions (e) and (f) are weaker than those in scan-based methods. Specifically, the scan-based method Li, Liu and Lin (2020) assumes that $L_{\min}/\log p \to \infty$, while the dBiRS algorithm has less restriction on the minimum length of signal regions. For instance, consider $L_{\min} = \log p$ and $\ell = \log p$, then we can select $2^s = \sqrt{\log p}$ for consistent detection. This suggests that the dBiRS algorithm can detect shorter (and unbalanced) signal regions than the scan-based method that requires $L_{\min}/\log p \to \infty$ and hence can be more accurate. Moreover, the dBiRS algorithm does not have restrictions on the maximum length of signal regions that the scan method demands and can deal with signal regions with a length of (polynomial) order p. For illustration, let the two signal regions are $I_1^* = \{1, \ldots, \lfloor p/4 \rfloor\}$ and $I_2^* = \{\lfloor 3/4p \rfloor, \ldots, p\}$. According to Theorem 3, dBiRS algorithm can consistently detect the signal regions, whereas the consistency of Q-SCAN (Li, Liu and Lin, 2020) requires $\log L_{\max}/\log p \to 0$. Condition (g) requires the correlations between signal regions and non-signal points to decay with the distance, which is a standard assumption in this field.

S2. Key lemmas.

LEMMA S1. *If* $\min_j \mathbb{E}(F_{n_j})^2 \geq b > 0$ *for some constant* b*, then for any sequence of* $\bar{\Delta}_n > 0$*, on the event* $\left\{\hat{\Delta}_n \leq \bar{\Delta}_n\right\}$*, we have*

$$\rho_n \lesssim \left(\bar{\Delta}_n\right)^{1/3} (\log p)^{2/3}.$$

This lemma is equivalent to lemma 5 in Xue and Yao (2020), one can refer it for more details.

LEMMA S2. *Assume that conditions (a) and (b) hold and only one signal region* I^* *exists. We further assume that for a suitable truncation parameter* s*,*

$$\left\|\mu_{I_{i+u}}\right\|_\infty \geq C_1 \left\{\log(pn)/n\right\}^{1/2},$$

where I_{i+u}*,* $u = 0, \ldots, k$ *is a minimum cover of* I^**. Then we have*

$$\mathbb{P}\left\{\left\|U_n(I_{i+u})\right\|_\infty > \tilde{c}(\alpha)\right\} \geq 1 - \frac{2p_{I_{i+u}}}{n^2 p^2} \to 1,$$

as $n \to \infty$ *for* $u = 0, \ldots, k$*, and then*

$$\mathbb{P}\left\{\cap_{u=0}^{k} \left\|U_n(I_{i+u})\right\|_\infty > \tilde{c}(\alpha)\right\} \geq 1 - \frac{2(k+1)\max_{0\leq u\leq k} p_{I_{i+u}}}{n^2 p^2} \geq 1 - \frac{4}{n^2 p} \to 1.$$

S3. Theoretical Proofs. Proof for Lemma S2 Under alternative hypothesis, $U_n \to \mathcal{N}(\mu, \Sigma)$. We note here that under the alternative hypothesis, the MLEs $\hat{\phi}$ and $\hat{\eta}$ are not consistent to the true dispersion ϕ and conditional mean η but converge to some ϕ_0 and η_0 if $\boldsymbol{G}^\top P\boldsymbol{G}/n$ converges for any ϕ and η. Therefore,

$$\boldsymbol{G}^\top P\boldsymbol{G}/n = \boldsymbol{G}^\top(\Lambda_0^{-1} - \Lambda_0^{-1}\boldsymbol{X}(\boldsymbol{X}^\top\Lambda_0^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^\top\Lambda_0^{-1})\boldsymbol{G}/n,$$

where Λ_0 has the same form as Λ with $\phi = \phi_0$ and $\eta = \eta_0$. Moreover, under this case, $\hat{\Sigma}$ can not well approximate Σ.

Let $F_n \sim \mathcal{N}(0, \Sigma)$, we have

$$P_{I_{i+u}} = \mathbb{P}\left\{\|U_n(I_{i+u})\|_\infty > \tilde{c}(\alpha)\right\} = \mathbb{P}\left\{\left\|F_n(I_{i+u}) + \mu_{I_{i+u}}\right\|_\infty > \tilde{c}(\alpha)\right\} + o(1)$$

By triangle inequality,

$$P_{I_{i+u}} \geq \mathbb{P}\left\{\|F_n(I_{i+u})\|_\infty \leq \left\|\mu_{I_{i+u}}\right\|_\infty - \tilde{c}(\alpha)\right\}.$$

Let $\{b_j; j \leq p\}$ as the natural basis for $\mathbb{R}^p$, we have

$$\begin{aligned}\mathbb{P}\left\{\|F_n(I_{i+u})\|_\infty \geq t\right\} &\leq \sum_{j=1}^{p_{I_{i+u}}} \mathbb{P}\left\{\left|F_{n_j}\right| \geq t\right\} \\ &\leq \sum_{j=1}^{p_{I_{i+u}}} 2\exp\left\{-t^2/\left(2b_j^\top\Sigma b_j\right)\right\} \leq 2p_{I_{i+u}}\exp\left\{-t^2/\left(2\max_{j\leq p_{I_{i+u}}} b_j^\top\Sigma b_j\right)\right\}.\end{aligned}$$

For the threshold $\tilde{c}(\alpha)$, replacing Σ with $\hat{\Sigma}$ and let $I_{i+u} = [1:p]$ in the previous inequality, we have

$$\tilde{c}(\alpha)\left\{2\log(p/2\alpha)\max_{j\leq p}\left(b_j^\top\hat{\Sigma}b_j\right)\right\}^{1/2}.$$

By the convergence of $\boldsymbol{G}^\top P\boldsymbol{G}/n$ for a sufficiently large C_{10}, we have

$$\left\|\mu_{I_{i+u}}\right\|_\infty - \tilde{c}(\alpha) \geq C_{11}\sqrt{2\log(pn)}.$$

Hence, it can be deduced that, with probability tending to one,

$$P_{I_{i+u}} \geq 1 - \mathbb{P}\left\{\|F_n(I_{i+u})\|_\infty \geq \sqrt{2\log(pn)}\right\} \geq 1 - \frac{2p_{I_{i+u}}}{n^2p^2} \to 1.$$

Consequently, we have

$$\mathbb{P}\left\{\cap_{u=0}^k \|U_n(I_{i+u})\|_\infty > \tilde{c}(\alpha)\right\} \geq 1 - \sum_{u=0}^k P_{I_{i+u}} \geq 1 - \frac{2(k+1)\max_{u\leq k} p_{I_{i+u}}}{n^2p^2} \to 1.$$

Then we finish our proof for this lemma. ■

Proof for Theorem 1 To begin with, since $Y - \hat{\eta}_0 \to_d \mathcal{N}(0, P)$, then $\boldsymbol{G}^\top(Y - \hat{\eta}_0)/\sqrt{n} \to \mathcal{N}(0, \boldsymbol{G}^\top P\boldsymbol{G}/n)$. Let $F_n \sim \mathcal{N}(0, \boldsymbol{G}^\top P\boldsymbol{G}/n)$, then we have for any $t > 0$

$$\begin{aligned}&\left|\mathbb{P}\left(\|U_n\|_\infty \leq t\right) - \mathbb{P}\left(\|F_n\|_\infty \leq t\right)\right| \\ &= \left|\mathbb{P}\left(|U_{n_1}| \leq t, \ldots, \left|U_{n_p}\right| \leq t\right) - \mathbb{P}\left(|F_{n_1}| \leq t, \ldots, \left|F_{n_p}\right| \leq t\right)\right| \to 0,\end{aligned}$$

as $n \to \infty$.

Denote that

$$\rho_n = \sup_{A\in\mathcal{A}}\left|\mathbb{P}(F_n \in A) - \mathbb{P}(U_n^e \in A)\right|,$$

where $U_n^e = \boldsymbol{G}^\top \hat{P}^{1/2} e/\sqrt{n}$, $e \sim \mathcal{N}(0, I_n)$ and $\mathcal{A}$ is the collection of rectangles in $\mathbb{R}^p$. Then we have $U_n^e \sim \mathcal{N}(0, \boldsymbol{G}^\top \hat{P} \boldsymbol{G}/n$, let

$$\hat{\Delta}_n = \left\| \boldsymbol{G}^\top P \boldsymbol{G} - \boldsymbol{G}^\top \hat{P} \boldsymbol{G} \right\|_\infty / n.$$

According to condition (a) and Lemma S1, by δ-method, we have that there exists a sufficient large constant C_1 such that for $i = 1, \ldots, n$,

$$\left| \Lambda_{ii} - \hat{\Lambda}_{ii} \right| < \frac{C_1}{\sqrt{n}},$$

with probability tending to one. Additionally,

$$\left\| \left(\boldsymbol{X}^\top \Lambda^{-1} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^\top \Lambda^{-1} \boldsymbol{G} \right\|_\infty = \left\| \left(\boldsymbol{X}^\top \Lambda^{-1} \boldsymbol{X}/n \right)^{-1} \boldsymbol{X}^\top \Lambda^{-1} \boldsymbol{G}/n \right\|_\infty \leq C_2,$$

for some constant C_2. It implies that with probability tending to one,

$$\left\| \left(\boldsymbol{X}^\top \hat{\Lambda}^{-1} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^\top \hat{\Lambda}^{-1} \boldsymbol{G} \right\|_\infty \leq C_3,$$

for some constant C_3. Next, we aim to bound $\hat{\Delta_n}$,

$$\begin{aligned}
\boldsymbol{G}^\top \left(P - \hat{P} \right) \boldsymbol{G} = & \boldsymbol{G}^\top \left(\Lambda^{-1} - \hat{\Lambda}^{-1} \right) \boldsymbol{G} \\
& - \boldsymbol{G}^\top \left(\Lambda^{-1} - \hat{\Lambda}^{-1} \right) \boldsymbol{X} \left(\boldsymbol{X}^\top \Lambda^{-1} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^\top \Lambda^{-1} \boldsymbol{G} \\
& - \boldsymbol{G}^\top \hat{\Lambda}^{-1} \boldsymbol{X} \left\{ \left(\boldsymbol{X}^\top \Lambda^{-1} \boldsymbol{X} \right)^{-1} - \left(\boldsymbol{X}^\top \hat{\Lambda}^{-1} \boldsymbol{X} \right)^{-1} \right\} \boldsymbol{X}^\top \Lambda^{-1} \boldsymbol{G} \\
& - \boldsymbol{G}^\top \hat{\Lambda}^{-1} \boldsymbol{X} \left(\boldsymbol{X}^\top \hat{\Lambda}^{-1} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^\top \left(\Lambda^{-1} - \hat{\Lambda}^{-1} \right) \boldsymbol{G} \\
& = \Delta_1 - \Delta_2 - \Delta_3 - \Delta_4.
\end{aligned}$$

For Δ_1,

$$\|\Delta_1\|_\infty \leq n \|\boldsymbol{G}\|_\infty \left\| \hat{\Lambda}^{-1} - \Lambda^{-1} \right\|_\infty \|\boldsymbol{G}\|_\infty \leq C_4 \sqrt{n}$$

for some constant C_4. For Δ_2, we have

$$\begin{aligned}
\|\Delta_2\|_\infty & \leq d \left\| \boldsymbol{G}^\top \left(\Lambda^{-1} - \hat{\Lambda}^{-1} \right) \boldsymbol{X} \right\|_\infty \left\| \left(\boldsymbol{X}^\top \Lambda^{-1} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^\top \Lambda^{-1} \boldsymbol{G} \right\|_\infty \\
& \leq C_3 n d \|\boldsymbol{G}\|_\infty \left\| \hat{\Lambda}^{-1} - \Lambda^{-1} \right\|_\infty \|\boldsymbol{X}\|_\infty \\
& \leq C_5 d \sqrt{n}
\end{aligned}$$

for some constant C_5. For Δ_3, we have

$$\Delta_3 = \boldsymbol{G}^\top \hat{\Lambda}^{-1} \boldsymbol{X} \left(\boldsymbol{X}^\top \hat{\Lambda}^{-1} \boldsymbol{X} \right)^{-1} \cdot \left\{ \boldsymbol{X}^\top \left(\Lambda^{-1} - \hat{\Lambda}^{-1} \right) \boldsymbol{X} \right\} \cdot \left(\boldsymbol{X}^\top \Lambda^{-1} \boldsymbol{X} \right)^{-1} \boldsymbol{X}^\top \Lambda^{-1} \boldsymbol{G},$$

thus we have

$$\|\Delta_3\|_\infty \leq C_6 d^3 \sqrt{n}$$

for some constant C_6. Similarly, for Δ_4, we have

$$\|\Delta_4\|_\infty \leq C_7 d \sqrt{n}$$

for some constant C_7. Therefore, we have that

$$\hat{\Delta}_n \leq \frac{1}{n}\sum_{j=1}^{4}\|\Delta_j\|_\infty \leq \frac{C_8}{\sqrt{n}}$$

for some constant C_8. Then by Lemma S1, we have with probability tending to one

$$\rho_n \leq C_9\frac{(\log p)^{2/3}}{n^{1/6}} = C_9\left(\frac{\log^4 p}{n}\right)^{1/6}.$$

According to condition (b), $log^4 p/n \to 0$, then we have that

$$|\mathbb{P}\{\|U_n\|_\infty > \tilde{c}(\alpha)\} - \alpha| \to 0.$$

Then we finish the proof. ■

Proof of Theorem 2 Without loss of generality, we assume that the true continuous signal regions reside in some blocks. In order to prove the theorem, we need some algorithm-based inequalities. We first look into the case that there exists two signal regions I_1^*, I_2^*, let,

$$A'_{1g} = \{\text{The binary search step detected } I_1^* \text{ and } g \text{ cover elements for } I_2^*\},$$

Let,

$$A'_2|A'_{1g} = \left\{\text{The first re-search detected the remain } k_2 - g \text{ cover elements under } A'_{1g}\right\}.$$

Next, we specify the upper bound for $\mathbb{P}\left(A'_2|A'_{1g}\right)$: under A'_{1g}, assume that $I_{i_2+u_1}, \ldots, I_{i_2+u_g}$ be found in the binary search algorithm, then the variables and the sample matrices in $I_1^*, I_{i_2+u_1}, \ldots, I_{i_2+u_g}$ are replaced by 0. Then for some remaining cover element $I_{i_2+u'}$, from the proof of Lemma S2, we have the threshold used for the global test in the second round binary search

$$\tilde{c}_2(\alpha) \leq c_{11}\left\{\log\left[\left(p - p_1 - \sum_{j=1}^{g} p_{I_{i_2+u_j}}\right)\right]\right\}^{1/2}.$$

Thus, $\mathbb{P}\left(A'_2|A'_{1g}\right) \geq 1 - \frac{4}{n^2(p-p_1)}$. Then under the case that there exist only two signal regions, we have,

(S2)

$$P_2 = \sum_{g=0}^{k_2}\mathbb{P}\left(A'_{1g}\right)\cdot\mathbb{P}\left(A'_2|A'_{1g}\right) \geq \sum_{g=0}^{k_2}\mathbb{P}\left(A'_{1g}\right)\left\{1 - \frac{4}{n^2(p-p_1)}\right\} \geq P_d\cdot\left\{1 - \frac{4}{n^2(p-p_1)}\right\}.$$

Then by induction and Lemma S2, in general, we have,

$$P_\ell \geq \prod_{r=1}^{\ell}\left\{1 - \frac{4}{n^2\left(p - \sum_{j=1}^{r-1} p_j\right)}\right\} \geq 1 - \frac{4}{n^2}\sum_{i=1}^{p}\frac{1}{i} \geq 1 - \frac{4\log(p+1)}{n^2}.$$

Then we finish our proof. ■

Proof of Theorem 3 We first remark here that for r high-correlated variants around true signal regions, we can suppose that the algorithm can detect them in any case. Since $r = o(L_{\min})$, it will not influence the asymptotic results. Hereafter, we only prove for $r = 0$ for convenience. Let $I_j^{\#} = \cup_{u=0}^{k_j} I_{i_j+u}$ be the minimum cover of $I_j^*, j = 1, \ldots, \ell$. According to

Lemma S2 and condition (c), with probability one, the dBiRS algorithm tends to detect the $I_1^{\#}$, i.e., there exists an estimated $\hat{I}_1 \in \hat{\mathcal{I}}$ (w.l.o.g, we denote it as $\hat{I}_1$) such that,

$$\mathbb{P}\left\{\left|\hat{I}_1 \cup I_1^{\#}\right| = \left|I_1^{\#}\right|\right\} \geq 1 - \frac{4}{n^2 p}. \tag{S3}$$

Since $I_1^* \subset I_1^{\#}$ and according to condition (f), $\left|I_1^{\#}\right| \to |I_1^*|$ as p growth. Then the Jaccard index between $\hat{I}_1$ and $I_1^{\#}$ only depends on how many non-signal segments in $\mathcal{S}$ be detected, under the condition that $\left|\hat{I}_1 \cup I_1^{\#}\right| = \left|I_1^{\#}\right|$, since $\hat{I}_1$ is continuous, let $\hat{I}_1 = \cap_{u=-a_1}^{h_1+b_1} I_{i_1+u}$, then the Jaccard index is (since the lengths of elements in $\mathcal{S}$ just differ in 1, w.l.o.g, we assume that they have the same length),

$$J(I_1^{\#}, \hat{I}_1) = 1 - \frac{a_1 + b_1}{a_1 + b_1 + h_1 + 1}.$$

For $R_1 = a_1 + b_1 < k_1$, then $\hat{I}_1 \cup I_r^{\#} = \emptyset$ according to condition (e) and $\mu^{X_{I_{i_1+u'}}} - \mu^{Y_{I_{i_1+u'}}} = 0$ for $u' = -a_1, \ldots, -1, k_1+1, \ldots, k_1+b_1$. According to the last control procedure in dBiRS, whatever which step $I_{i_1+u'}$ has been selected, it must satisfy that $T_n(I_{i_1+u'}) \geq \hat{c}_{I_{i_1+u'}}(\alpha)$, and the critical value $Thr_{i_1+u'} = \hat{c}_{I_{i_1+u'}}(\alpha)$ is computed under the sample matrices only include the region $I_{i_1+u'}$. For $u' = -a_1, \ldots, -1, h_1+1, \ldots, h_1+b$, then we have that,

$$\begin{aligned}
&\mathbb{P}\left\{J\left(I_1^{\#}, \hat{I}_1\right) \leq 1 - \frac{R_1}{R_1 + h_1 + 1}\right\} \\
&= \mathbb{P}\left[\left\{\cap_{u'=-r}^{-1} T_n\left(I_{i_1+u'}\right) \geq Thr_{i_1+u'}\right\} \cap \left\{\cap_{u'=1}^{R_1-r} T_n\left(I_{i_1+h_1+u'}\right) \geq Thr_{i_1+h_1+u'}\right\}\right. \\
&\qquad \left. \cap \left\{\left|\hat{I}_1 \cap I_1^{\#}\right| = \left|I_1^{\#}\right|\right\}\right] \\
&+ \mathbb{P}\left[J\left(I_1^{\#}, \hat{I}_1\right) \leq 1 - \frac{R_1}{R_1 + h_1 + 1}, \left|\hat{I}_1 \cap I_1^{\#}\right| \neq \left|I_1^{\#}\right|\right] \\
&\leq \mathbb{P}\left[\left\{\cap_{u'=-r}^{-1} T_n\left(I_{i_1+u'}\right) \geq Thr_{i_1+u'}\right\} \cap \left\{\cap_{u'=1}^{R_1-r} T_n\left(I_{i_1+h_1+u'}\right) \geq Thr_{i_1+h_1+u'}\right\}\right] \\
&+ \mathbb{P}\left\{\left|\hat{I}_1 \cap I_1^{\#}\right| \neq \left|I_1^{\#}\right|\right\} \\
&\leq D(s, R_1)\alpha^{R_1} + \frac{4}{n^2 p} \leq \left(\frac{\alpha}{2\beta}\right)^{R_1} + \frac{4}{n^2 p}.
\end{aligned}$$

Mimic the notations above, for I_j^* and $R_i \leq h_i, i = 1, \ldots j-1$, let

$$A_j = \cup_{i=1}^{j-1} \left\{J\left(I_i^{\#}, \hat{I}_i\right) \geq 1 - \frac{R_j}{R_j + h_j + 1}\right\},$$

under A_j, the estimated $\hat{I}_i, i = 1, \ldots, j-1$ disjoint and $\hat{I}_i \cup I_j^{\#} = \emptyset$ according to condition (e), mimic the calculation procedure for $\hat{I}_1$ and according to (S2), (S3), we have there exists

$\hat{I}_j$ such that,

$$\begin{aligned}
&\mathbb{P}\left\{J\left(I_j^{\#},\hat{I}_j\right)\leq 1-\frac{R_j}{R_j+h_j+1}\,|A_j\right\}\\
&=\mathbb{P}\left[\left\{\cap_{u'=-r}^{-1}T_n\left(I_{i_j+u'}\right)\geq Thr_{i_j+u'}\right\}\cap\left\{\cap_{u'=1}^{R_j-r}T_n\left(I_{i_j+h_j+u'}\right)\geq Thr_{i_j+h_j+u'}\right\}\right.\\
&\qquad\left.\cap\left\{\left|\hat{I}_j\cap I_j^{\#}\right|=\left|I_j^{\#}\right|\right\}|A_j\right]\\
&+\mathbb{P}\left[J\left(I_j^{\#},\hat{I}_j\right)\leq 1-\frac{R_j}{R_j+h_j+1},\left|\hat{I}_j\cap I_j^{\#}\right|\neq\left|I_j^{\#}\right|\,|A_j\right]\\
&\leq\mathbb{P}\left[\left\{\cap_{u'=-r}^{-1}T_n\left(I_{i_j+u'}\right)\geq Thr_{i_j+u'}\right\}\cap\left\{\cap_{u'=1}^{R_j-r}T_n\left(I_{i_j+h_j+u'}\right)\geq Thr_{i_j+h_j+u'}\right\}|A_j\right]\\
&+\mathbb{P}\left\{\left|\hat{I}_j\cap I_j^{\#}\right|\neq\left|I_j^{\#}\right|\,|A_j\right\}\\
&\leq D(s,R_j)\alpha^{R_j}+\frac{4}{n^2\left(p-\sum_{r=1}^{j-1}p_r\right)}\leq\left(\frac{\alpha}{2\beta}\right)^{R_j}+\frac{4}{n^2\left(p-\sum_{r=1}^{j-1}p_r\right)}.
\end{aligned}$$

Recall that $\mathbb{P}\left\{J\left(I_1^{\#},\hat{I}_1\right)\leq 1-\frac{R_1}{R_1+h_1+1}\right\}\leq(\alpha/2\beta)^{R_1}+4/n^2p$, according to condition (f), we have

$$\begin{aligned}
P_\ell&=\mathbb{P}\left[\cap_{j=1}^{\ell}\left\{J\left(I_j^{\#},\hat{I}_j\right)\geq 1-\frac{R_j}{R_j+h_j+1}\right\}\right]\\
&=1-\mathbb{P}\left[\cap_{j=1}^{\ell}\left\{J\left(I_j^{\#},\hat{I}_j\right)\leq 1-\frac{R_j}{R_j+h_j+1}\right\}\right]\\
&\geq\prod_{j=1}^{\ell}\left\{1-\left(\frac{\alpha}{2\beta}\right)^{R_j}-\frac{4}{n^2\left(p=\sum_{r=1}^{j-1}p_r\right)}\right\}\geq 1-\sum_{j=1}^{\ell}\left(\frac{\alpha}{2\beta}\right)^{R_j}-\frac{4\log(p+1)}{n^2}\to 1.
\end{aligned}$$

From the previous results and condition (e), we can assume that

$$\hat{\mathcal{I}}=\left\{\hat{I}_1,\ldots,\hat{I}_\ell,\hat{I}_1(1),\ldots,\hat{I}_{\ell_1}(1),\ldots,\hat{I}_1(K),\ldots,\hat{I}_{\ell_K}(K)\right\},$$

where $\hat{I}_j$ covered the signal region I_j^* and $\hat{I}_i(k)$, $i=1,\ldots,\ell_k$ are false detected regions in block k. We first analyze the proportion of false detection in each block, w.l.o.g, for block one.

Let $\hat{I}_v(1)$ combined by $R_v(1)$ elements in $\mathcal{S}$ and $R_0(1)=\sum_{v=1}^{\ell_1}R_v(1)$, let $S=\lceil p(1)/2^s\rceil-\sum_{j=1}^{\ell}h_j$, where $p(1)$ is the dimension of block one.

$P^*_{R_0(1)}=\{\exists R_0(1)$ non-signal segments not neighbor to the true signal regions in $\mathcal{S}$ detected$\}$

$P_{R_0(1)}=\{\exists R_0(1)$ regions in $\mathcal{S}$ been detected under ℓ true signal regions removed$\}$

It obviously that $P^*_{R_0(1)}\leq P_{R_0(1)}$ due to the dimension reduction. Since the consistency of detecting some specific signal region has been proved. We just need to specify the value of $P_{R_0(1)}$.

Assume that the sBiRS algorithm in detected $R_0(1)$ regions in block one after ℓ true signal regions be removed through N whole binary search (the m-th whole binary search is the binary search after $m-1$ re-search procedure), $1\leq N\leq R_0(1)$, the corresponding probability

is $P_N(1)$. Let

$$B(R_{0i}) = \left\{ \exists R_{0i} \text{ segments been detected among the remaining } S - \sum_{j=1}^{i-1} R_{0j} \text{ segments} \right\},$$

$R_0(1) = \sum_{j=1}^{N} R_{0i}$.

Denoting $\Lambda_i = \{$ the index of segments in $\mathcal{S}$ remaining after i whole binary search under the true signal regions removed $\}$, $i = 1, \ldots, N$. Let

$$C_{ik} = \{\text{the segments } k \text{ is selected as signal region at step } i\},$$

$i = 0, \ldots, N-1; k = 1, \ldots, S$. Then we have,

$$\begin{aligned} P_N(1) &\leq \sum_{R_{01}+\cdots+R_{0N}=R_0} \prod_{i=1}^{N} \mathbb{P}\left[B(R_{0i})\right] \\ &\leq \sum_{R_{01}+\cdots+R_{0N}=R_0(1)} \prod_{i=1}^{N} \mathbb{P}\left\{\cup_{\lambda_{k1},\ldots,\lambda_{kR_{0i}} \in \Lambda_{i-1}} \left[C_{\lambda_{k1} i} \cap \cdots \cap C_{\lambda_{kR_{0i}} i}\right]\right\} \qquad \text{(S4)} \\ &\leq \sum_{R_{01}+\cdots+R_{0N}=R_0(1)} \prod_{i=1}^{N} \left\{ \sum_{\lambda_{k1},\ldots,\lambda_{kR_{0i}} \in \Lambda_{i-1}} \mathbb{P}\left[C_{\lambda_{k1} i} \cap \cdots \cap C_{\lambda_{kR_{0i}} i}\right] \right\} \end{aligned}$$

For the probability of C_{ik}, we have that $\mathbb{P}(C_{ik}) \leq \alpha / \left(S - \sum_{j=0}^{i-1} R_{0j}\right)$, continue with (S4), we have,

$$\begin{aligned} P_N(1) &\leq \sum_{R_{01}+\cdots+R_{0N}=R_0(1)} \binom{S}{R_{01}} \cdots \binom{S - \sum_{j=1}^{N-1} R_{0j}}{R_{0N}} \prod_{i=1}^{N} \left(\frac{\alpha}{S - \sum_{j=1}^{i-1} R_{0j}} \right)^{R_{0i}} D(s, R_{0i}) \\ &\leq \sum_{R_{01}+\cdots+R_{0N}=R_0(1)} \binom{S}{R_{01}} \cdots \binom{S - \sum_{j=1}^{N-1} R_{0j}}{R_{0N}} \frac{(S - R_0(1))!}{S!} \alpha^{R_0(1)} \prod_{i=1}^{N} D(s, R_{0i}). \end{aligned}$$

By condition (d), we have $\prod_{i=1}^{N} D(s, R_{0i}) = (1/2\tilde{\alpha})^{R_0(1)}$, and recall that,

$$\begin{aligned} \sum_{R_{01}+\cdots+R_{0N}=R_0(1)} \binom{S}{R_{01}} \cdots \binom{S - \sum_{j=1}^{N-1} R_{0j}}{R_{0N}} &= \frac{S!}{(S - R_0(1))!} \sum_{R_{01}+\cdots+R_{0N}=R_0(1)} \frac{1}{R_{01}! \cdots R_{0N}!} \\ &\leq \frac{S! \binom{R_0(1)-1}{N-1}}{(S - R_0(1))!}. \end{aligned}$$

Then we have $P_N(1) \leq \binom{R_0(1)-1}{N-1} (\alpha / 2\tilde{\alpha})^{R_0(1)}$, thus, we have that,

$$P_{R_0(1)} \leq \sum_{M=R_0(1)}^{S} \sum_{N=1}^{M} \binom{M-1}{N-1} \left(\frac{\alpha}{2\tilde{\alpha}}\right)^{M} \leq \sum_{M=R_0(1)}^{S} \left(\frac{\alpha}{\tilde{\alpha}}\right)^{M} \leq \left(\frac{\alpha}{\tilde{\alpha}}\right)^{R_0(1)-1}.$$

Let R_0 be the total number of non-signal segments, w.l.o.g, we assume that the ℓ signal regions reside in K_1 blocks, then $K \geq K_1$ and there exists at most $\min\{R_0, K\}$ blocks which have detected signal regions. Since the dBiRS algorithm conducts a sBiRS procedure

to determine the significance of blocks, we have that

$$\begin{aligned}
&\mathbb{P}\left\{\frac{\left|\tilde{I}\right|}{|I^*|} \geq \frac{R_0}{\sum_{j=1}^{\ell} h_j}\right\} \\
&\leq c_{12} \sum_{x_1+x_2=R_0; x_1 \geq \ell} \binom{x_1-1}{K_1-1} \left(\frac{\alpha}{\tilde{\alpha}}\right)^{x_1-K_1} \cdot \left\{\sum_{j=1}^{\min(x_2, K-K_1)} \left(\frac{\alpha}{\tilde{\alpha}}\right)^{j} \binom{x_2-1}{j-1} \left(\frac{\alpha}{\tilde{\alpha}}\right)^{x_2-j}\right\} \\
&\leq c_{12} \sum_{x_1+x_2=R_0; x_1 \geq K_1} \binom{x_1-1}{K_1-1} 2^{x_2} \left(\frac{\alpha}{\tilde{\alpha}}\right)^{R_0-K_1} \leq \left(\frac{2\alpha}{\tilde{\alpha}}\right)^{R_0-K_1} \sum_{x_1=K_1}^{R_0} \binom{x_1-1}{K_1-1} \\
&\leq c_{12} \left\{\frac{2R_0\alpha}{(R_0-K_1)\tilde{\alpha}}\right\}^{R_0-K_1},
\end{aligned}$$

for some constant c_{12}. Finally, we have that

$$\begin{aligned}
&\mathbb{P}\left[\left\{\frac{\left|\tilde{I}\right|}{|I^*|} \leq \frac{R_0}{\sum_{j=1}^{\ell} h_j}\right\} \bigcap \left\{\bigcap_{j=1}^{\ell} \left\{J\left(I_j^{\#}, \hat{I}_j\right) \geq 1 - \frac{R_j}{R_j + h_j + 1}\right\}\right\}\right] \\
&\geq \left\{1 - \frac{4\log(p+1)}{n^2} - \sum_{j=1}^{\ell} \left(\frac{\alpha}{2\tilde{\alpha}}\right)^{R_j}\right\} \left\{1 - c_{12}\left(\frac{2R_0\alpha}{(R_0 = K_1)\tilde{\alpha}}\right)^{R_0-K_1}\right\}.
\end{aligned}$$

Then we finish our proof. ■

REFERENCES

JENG, X. J., CAI, T. T. and LI, H. (2010). Optimal Sparse Segment Identification With Application in Copy Number Variation Analysis. *Journal of the American Statistical Association* **105** 1156-1166. PMID: 23543902. https://doi.org/10.1198/jasa.2010.tm10083

LI, Z., LIU, Y. and LIN, X. (2020). Simultaneous Detection of Signal Regions Using Quadratic Scan Statistics With Applications to Whole Genome Association Studies. *Journal of the American Statistical Association* **0** 1-12. https://doi.org/10.1080/01621459.2020.1822849

XUE, K. and YAO, F. (2020). Distribution and correlation-free two-sample test of high-dimensional means. *The Annals of Statistics* **48** 1304 – 1328. https://doi.org/10.1214/19-AOS1848

ZHANG, W., WANG, F. and YAO, F. (2025). Fast Signal Region Detection with Application to Whole Genome Association Studies. *Journal of the American Statistical Association*. https://doi.org/10.1080/01621459.2025.2464271