A nearly optimal sequential testing approach to permutation-based association testing

Julian Hecker[1], Ingo Ruczinski[2], Brent Coull[1], Christoph Lange[1,3]

1) Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

2) Department of Biostatistics, Bloomberg School of Public Health, Baltimore, MD, USA

3) Division of Network Medicine, Brigham and Women's Hospital, Boston, MA, USA

The following technical report describes the technical details for the implementation of a sequential testing approach to permutation-based association testing in whole-genome sequencing studies. The sequential testing approach enables to control the probability of a type 1 and type 2 error at arbitrary small pre-specified levels and approaches the theoretical minimum of expected number of required permutations as these levels go to zero.

In practice, since it is not feasible to go through all permutations of a genetic data set, the permutation-based p-value is usually estimated from a large number of random permutations.

The procedure of the re-calculation of the association test statistic for permuted data and comparison with the observed test statistic can be described by a sequence $x_1, x_2, \ldots$, where $x_i = 1$ if the i-th permuted test statistic is larger or equal than the observed statistic and $x_i = 0$ otherwise. Denote the true and unknown association p-value, computed by the evaluation of all permutations (not feasible in practice), by θ. The scientific interest is mainly summarized by the question if $\theta \leq p_1$, for a pre-specified significance level p_1. For example, $p_1 = 5 \cdot 10^{-8}$ in classical GWAS. Given the extremely large number of

possible permutations and assuming an appropriate generation of random permutations, we interprete the sequence $x_1, x_2, \ldots$ as independent, identically distributed Bernoulli random variables with success probability θ. In the following, we describe how significance testing can be performed efficiently by a sequential testing approach.

Setting

Let $(\Omega, \mathcal{F}, \mathbf{P}_\theta)$ be a probability space,

$x_1, x_2, \cdots \sim$ Bernoulli(θ), a sequence of independent identically distributed Bernoulli random variables with success parameter θ, and $\mathcal{F}_n = \sigma(x_1, \ldots, x_n) \subset \mathcal{F}$, for $n = 1, 2, \ldots$.

$0 \leq \theta \leq 1$ is the true p-value, that can computed (theoretically) by evaluating all permutations of the data set. We extended the standard Bernoulli distribution for $\theta \in (0, 1)$ to the inclusion of the extreme cases $\theta = 1$ and $\theta = 0$.

We would like to differentiate between two hypotheses:

$$H_1 : \quad \theta \leq p_1 \qquad\qquad H_2 : \quad \theta \geq p_2,$$

where $p_2 - p_1 = d > 0$, $p_1 > 0$ and $p_2 < 0.5$. In practice, we choose, for example, $p_1 = 5 \cdot 10^{-8}$

(genome-wide significane in classical GWAS), $d = 10^{-8}$(resolution level of 10^8 permutations). d is chosen small and affects the worst case expected run time, as described below. The interval (p_1, p_2) is the so-called indifference zone, where both hypotheses are plausible.

Sequential testing framework: Corresponding objects and results

We utilize the work of Pavlov (1991) [1] and Tartakovsky (2014) [2] for sequential testing theory. The following strategies and results are strongly related to the work in [1] and adapt these results from the general setting to our specific scenario. In particular, we show that our estimator for θ is an appropriate choice and deal with the problem of the

degenerated cases $\theta = 1$ or $\theta = 0$.

Let $D_1 := [0, p_1]$ and $D_2 := [p_2, 1]$. Let $(\alpha_1, \alpha_2) = (\alpha t_1, \alpha t_2)$ for positive t_1, t_2, α such that $\alpha_1 + \alpha_2 < 1$.

Introduce

$$\tau_1(\alpha_1) := \min\left\{ n \middle| \pi_n / \sup_{\theta \in D_1} p_n(\theta, x^n) \geq \alpha_1^{-1} \right\}$$

and

$$\tau_2(\alpha_2) := \min\left\{ n \middle| \pi_n / \sup_{\theta \in D_2} p_n(\theta, x^n) \geq \alpha_2^{-1} \right\},$$

where $\pi_n := \prod_{r=1}^{n} p(\hat{\theta}_{r-1}, x_r)$ with $\pi_1 := 0.5$ and $p_n(\theta, x^n) = \prod_{r=1}^{n} p(\theta, x_r)$. $\hat{\theta}_{r-1}$ is an estimate of θ which depends only on the first $r-1$ observations and is specified by $\hat{\theta}_{r-1} := \frac{\sum_{k=1}^{r-1} x_k + \frac{1}{2}}{r}$.

A decision test is described by $\mathcal{F}_n$-stopping time N and a $\mathcal{F}_N$-measurable function δ, which can take the values 1 and 2.

We define the decision test for our approach as

$$(N, \delta) = \begin{cases} (\tau_1(\alpha_1), 2) & \text{if } \tau_1(\alpha_1) \leq \tau_2(\alpha_2) \\ (\tau_2(\alpha_2), 1) & \text{otherwise.} \end{cases}$$

Denote by $\rho(\theta_1, \theta_2)$ the Kullback-Leibler Distance for the Bernoulli distribution, defined by $\rho(\theta_1, \theta_2) := \theta_1 \log\left[\frac{\theta_1}{\theta_2}\right] + (1 - \theta_1) \log\left[\frac{1-\theta_1}{1-\theta_2}\right]$, for $\theta_1, \theta_2 \in (0, 1)$.
From our choice for the decision test, we get the following results.

Theorem. *1.) For the error probabilities, we have*

$$\mathbf{P}_\theta\left[\delta = 2\right] \leq \alpha_1 \text{ for } \theta \in D_1$$

and

$$\mathbf{P}_\theta\left[\delta = 1\right] \leq \alpha_2 \text{ for } \theta \in D_2.$$

2.) For the expected number of permutations, we have

$$\mathbf{E}_\theta[N] \leq \begin{cases} \frac{|\log(\alpha_1)|}{\rho(\theta,p_1)}(1+o(1)) & \text{if } \theta \in D_2 - \{1\} \\ \frac{|\log(\alpha_1)|}{\log\frac{1}{p_1}}(1+o(1)) & \text{if } \theta = 1 \\ \frac{|\log(\alpha_2)|}{\rho(\theta,p_2)}(1+o(1)) & \text{if } \theta \in D_1 - \{0\} \\ \frac{|\log(\alpha_2)|}{\log\frac{1}{1-p_2}}(1+o(1)) & \text{if } \theta = 0 \\ \min\left[\frac{|\log(\alpha_1)|}{\rho(\theta,p_1)}, \frac{|\log(\alpha_2)|}{\rho(\theta,p_2)}\right](1+o(1)) & \text{if } \theta \in (p_1,p_2), \end{cases}$$

as $\alpha \to 0$.

3.) Let $K(\alpha,t_1,t_2)$ *be the class of all decision tests* (N',δ') *such that* $\mathbf{P}_\theta\left[\delta' = 2\right] \leq \alpha t_1$ *for* $\theta \in D_1$ *and* $\mathbf{P}_\theta\left[\delta' = 1\right] \leq \alpha t_2$ *for* $\theta \in D_2$, *then*

$$\mathbf{E}_\theta[N] / \inf_{(N',\delta') \in K(\alpha,t_1,t_2)} \mathbf{E}_\theta[N'] = 1 + o(1) \text{ for all } \theta \in [0,1],$$

as $\alpha \to 0$.

Remark 1. *Note that, for given* $p_2 \in (0,1)$, *we have* $\lim_{p\to 0} \rho(p,p_2) = \log[\frac{1}{1-p_2}]$. *This shows that the results for* $\theta = 0$ *and* $\theta = 1$ *are the natural extensions.*

Proof of the Theorem

The proof of the Theorem is strongly related to the derivations in [1]. One important difference is that Pavlov derived uniform bounds, whereas our estimates will depend on θ. We use the explicit form of $\hat{\theta}_{r-1}$ and show how we can deal with the cases $\theta = 1$ and $\theta = 0$.

Lemma 1. *For all* $\theta \in D_1$

$$\mathbf{P}_\theta[\tau_1(\alpha_1) < \infty] \leq \alpha_1$$

and for all $\theta \in D_2$

$$\mathbf{P}_\theta[\tau_2(\alpha_2) < \infty] \leq \alpha_2.$$

Proof. We use the same argumentation as in [1]. In our setting, for any $\theta \in (0,1)$, the process $U_n(\theta) := \frac{\pi_n}{p_n(\theta, x^n)}$ forms a non-negative martingale with respect to $\mathcal{F}_n$. In addition, we have $\mathbf{E}_\theta[U_n(\theta)] = 1$, since $\pi_1 := \frac{1}{2}$. As in [1], introduce $v_\theta := \min\left\{ n \middle| U_n(\theta) \geq \alpha_1^{-1} \right\}$. By Doobs inequality, we have $\mathbf{P}_\theta[v_\theta \leq n] = \mathbf{P}_\theta[\max_{r=1,\ldots,n} U_r(\theta) \geq \alpha_1^{-1}] \leq \alpha_1$ for all n and so $\mathbf{P}_\theta[v_\theta < \infty] \leq \alpha_1$. For $\theta \in D_1$ with $\theta \neq 0$, we have $\tau_1(\alpha_1) \geq v_\theta$. Therefore, we get

$$\mathbf{P}_\theta[\tau_1(\alpha_1) < \infty] \leq \mathbf{P}_\theta[v_\theta < \infty] \leq \alpha_1.$$

For $\theta = 0$, obviously $\mathbf{P}_\theta[\tau_1(\alpha_1) < \infty] = 0$. Same argumentation for $\theta \in D_2$ and $\mathbf{P}_\theta[\tau_2(\alpha_2) < \infty]$. □

We define θ_n^{MLE} as the ordinary maximum likelihood estimator for θ, $\theta_n^{\text{MLE}} := \frac{1}{n}\sum_{r=1}^{n} x_k$. Furthermore, $V_\epsilon(\theta) := \left\{ \theta' \in (0,1) \;\middle|\; ||\theta - \theta'|| < \epsilon \right\}$ and
$\sigma_\epsilon(\theta) := \min\left\{ n \middle| \hat{\theta}_{k-1} \in V_\epsilon(\theta) \text{ and } \theta_k^{\text{MLE}} \in V_\epsilon(\theta) \text{ for all } k \geq n \right\}$.

Lemma 2. *For all* $\theta \in (0,1)$ *and* $\epsilon > 0$ *such that* $V_\epsilon(\theta) \subset (0,1)$,

$$\mathbf{P}_\theta\left[\sigma_\epsilon(\theta) > n\right] \leq K_{\theta,\epsilon} e^{-n\rho^-(\theta,\epsilon)},$$

for all n, *where* $\rho^-(\theta,\epsilon) := \min[\rho(\theta+\epsilon,\theta), \rho(\theta-\epsilon,\theta)]$.

Proof. Since we have an explicit form for the estimators $\hat{\theta}_{n-1}$ and θ_n^{MLE}, the proof is straightforward. Fix $\theta \in (0,1)$ and $\epsilon > 0$ such that $V_\epsilon(\theta) \subset (0,1)$. Define $X_n := \sum_{k=1}^{n} x_k$. Start with

$$\mathbf{P}_\theta\left[\hat{\theta}_{n-1} \geq \theta + \epsilon\right] \leq e^{-tn(\theta+\epsilon)+t\frac{1}{2}} \mathbf{E}_\theta\left[e^{tX_{n-1}}\right]$$

for all $t > 0$.

From the classical proof of the Chernoff-Hoeffding bound using moment-generating functions, we know that there is a $t^* > 0$, which depends on θ and ϵ, such that

$$e^{-t^* n(\theta+\epsilon)} \mathbf{E}_\theta \left[e^{t^* X_{n-1}} \right] \leq e^{-(n-1)\rho(\theta+\epsilon,\theta)} e^{-t^*(\theta+\epsilon)}.$$

In addition,

$$\mathbf{P}_\theta \left[\hat{\theta}_{n-1} \leq \theta - \epsilon \right] \leq e^{tn(\theta-\epsilon)-t\frac{1}{2}} \mathbf{E}_\theta \left[e^{-tX_{n-1}} \right],$$

for $t > 0$. Analogous argumentation shows the estimate for both estimators.

The statement of the Lemma follows from

$$\mathbf{P}_\theta \left[\sigma_\epsilon(\theta) > n \right] \leq \mathbf{P}_\theta \left[\cup_{r=n}^{\infty} \{\hat{\theta}_{r-1} \notin V_\epsilon(\theta)\} \quad \cup \quad \cup_{r=n}^{\infty} \{\theta_r^{\mathrm{MLE}} \notin V_\epsilon(\theta)\} \right]$$

$$\leq \sum_{r=n}^{\infty} \mathbf{P}_\theta \left[\hat{\theta}_{r-1} \notin V_\epsilon(\theta) \right] + \sum_{r=n}^{\infty} \mathbf{P}_\theta \left[\theta_r^{\mathrm{MLE}} \notin V_\epsilon(\theta) \right].$$

. □

Define $p_\epsilon^-(\theta, x_r) := \begin{cases} \theta - \epsilon & \text{if } x_r = 1 \\ 1 - \theta - \epsilon & \text{if } x_r = 0, \end{cases}$ for an appropriate $\epsilon > 0$, such that $\theta - \epsilon > 0$ and $1 - \theta - \epsilon > 0$.

Lemma 3. *Let $\theta \in (0,1)$, $\theta_0 \in (0,1)$ and $\delta > 0$. If $\epsilon > 0$ is chosen small enough such that*

$$\mathbf{E}_\theta \left[\log p_\epsilon^-(\theta, x_r) - \log p(\theta_0, x_r) \right] \geq \rho(\theta, \theta_0) - \delta, \tag{0.1}$$

then

$$\mathbf{P}_\theta \left[\sum_{k=1}^{n} \log p_\epsilon^-(\theta, x_k) - \log p(\theta_0, x_k) \leq n[\rho(\theta, \theta_0) - 2\delta] \right] \leq e^{-nb_{\theta,\epsilon}}$$

for all n, where $b_{\theta,\epsilon} > 0$.

Proof. The estimate $e^{-nb_{\theta,\epsilon}}$ depends on θ and ϵ, in opposite to the estimate in [1]. We proceed as in [1], but we use the explicit form of the Bernoulli distribution. Let $t > 0$,

define

$$d(\theta, \theta_0, \epsilon, \delta, x_r) := \log p_\epsilon^-(\theta, x_r) - \log p(\theta_0, x_r) - \rho(\theta, \theta_0) + 2\delta,$$

and easily compute

$$\mathbf{E}_\theta\left[e^{-td(\theta,\theta_0,\epsilon,\delta,x_r)}\right] = \theta e^{-t\log(\theta-\epsilon)+t\log(\theta_0)+t\rho(\theta,\theta_0)-2t\delta}$$

$$+(1-\theta)e^{-t\log(1-\theta-\epsilon)+t\log(1-\theta_0)+t\rho(\theta,\theta_0)-2t\delta}.$$

From here we can use the argumentation as in the proof of Lemma 5.1 in [1]. □

Lemma 4. *Let $\epsilon > 0$ and define*

$$L_\epsilon^1(\theta) := \sum_{r=1}^{\sigma_\epsilon(\theta)} \log(2r) + \sum_{r=1}^{\sigma_\epsilon(\theta)} \max[|\log(p_1)|, |\log(1-p_1)|]$$

and

$$L_\epsilon^2(\theta) := \sum_{r=1}^{\sigma_\epsilon(\theta)} \log(2r) + \sum_{r=1}^{\sigma_\epsilon(\theta)} \max[|\log(p_2)|, |\log(1-p_2)|].$$

Then,

$$\mathbf{E}_\theta\left[L_\epsilon^1(\theta)\right] < \infty \text{ for } \theta \in D_2 - \{1\}$$

and

$$\mathbf{E}_\theta\left[L_\epsilon^2(\theta)\right] < \infty \text{ for } \theta \in D_1 - \{0\}.$$

Proof. For $\theta \in D_2 - \{1\}$, we estimate

$$\mathbf{E}_\theta\left[L_\epsilon^1(\theta)\right] \leq C_{p_1} \sum_{r=1}^{\infty} \log(2r)\mathbf{P}_\theta\left[\sigma_\epsilon(\theta) > r\right] < \infty,$$

by the ratio test for infinite series, Lemma 2 and the properties of the geometric sum. Same argumentation for $L_\epsilon^2(\theta)$ if $\theta \in D_1 - \{0\}$. □

Proof of the Theorem. We start with 2.)

Let $\theta \in D_2 - \{1\}$ and $0 < \delta < \rho(\theta, p_1)$. Choose $\epsilon > 0$ such that $V_\epsilon(\theta) \in (p_2, 1)$ and such that

$$\mathbf{E}_\theta \left[\log p_\epsilon^-(\theta, x_r) - \log p(p_1, x_r) \right] \geq \rho(\theta, p_1) - \delta. \tag{0.2}$$

Then, we can estimate

$$\sum_{r=1}^{n} \log p(\hat{\theta}_{r-1}, x_r) - \log \sup_{\theta \in D_1} \prod_{r=1}^{n} p(\theta, x_r) \tag{0.3}$$

$$\geq \sum_{r=1}^{n} \log p_\epsilon^-(\theta, x_r) - \log \sup_{\theta \in D_1} \prod_{r=1}^{n} p(\theta, x_r) + \sum_{r=1}^{\sigma_\epsilon(\theta)} \log p(\hat{\theta}_{r-1}, x_r)$$

$$\geq \sum_{r=1}^{n} \log p_\epsilon^-(\theta, x_r) - \sum_{r=1}^{n} \log p(p_1, x_r) + \sum_{r=1}^{\sigma_\epsilon(\theta)} \log p(\hat{\theta}_{r-1}, x_r) + \sum_{r=1}^{\sigma_\epsilon(\theta)} \log p(p_1, x_r)$$

$$\geq \sum_{r=1}^{n} \log p_\epsilon^-(\theta, x_r) - \sum_{r=1}^{n} \log p(p_1, x_r) - L_\epsilon^1(\theta).$$

Define

$$T_1(\theta, \epsilon, \alpha_1) := \min \left\{ n \middle| \sum_{r=1}^{m} \log p_\epsilon^-(\theta, x_r) - \sum_{r=1}^{m} \log p(p_1, x_r) \geq \log \alpha_1^{-1} + L_\epsilon^1(\theta) \text{ for all } m \geq n \right\}.$$

From this point, with the same argumentation as in the proof of Lemma 5.6 in [1], using Lemma 3, it follows

$$\mathbf{E}_\theta[T_1(\theta, \epsilon, \alpha_1)] \leq \frac{|\log(\alpha_1)|}{\rho(\theta, p1)}(1 + o(1))$$

as $\alpha \to 0$. The same can be shown for the analogously defined $T_2(\theta, \epsilon, \alpha_2)$. This together with Eq.(0.3) concludes the claim regarding the expected number of permutations for $\theta \in D_1 - \{0\}$ and $\theta \in D_2 - \{1\}$. For $\theta \in (p_1, p_2)$ the claim follows since the ratio between

α_1 and α_2 is assumed to be fixed and equal to $\frac{t_1}{t_2}$ as $\alpha \to 0$.

In the scenario $\theta = 1$, we have a deterministic setting with $x_r = 1$ for all r. Then, N is determined by

$$N = \min\left\{ n \in \mathrm{N} \middle| \prod_{r=1}^{n} \left(1 - \frac{1}{2r}\right) \geq \alpha_1^{-1} p_1^n \right\}.$$

Furthermore,

$$N \geq \min\left\{ n \in \mathrm{N} \middle| \prod_{r=1}^{n} \left(1 - \frac{1}{r}\right) \geq 2\alpha_1^{-1} p_1^n \right\} = \min\left\{ n \in \mathrm{N} \middle| \frac{1}{n} \geq 2\alpha_1^{-1} p_1^n \right\}.$$

If we analyze $\min\left\{ n \in \mathrm{N} \middle| \frac{1}{n} \geq 2\alpha_1^{-1} p_1^n \right\}$ using the Lambert W function and the corresponding asymptotic expansions, we can conclude

$$N \leq \frac{\log|\alpha_1|}{\log \frac{1}{p_1}}(1 + o(1))$$

as $\alpha \to 0$. Exactly the same argumentation shows the desired statement for $\theta = 0$.

1.) We showed for all $\theta \in [0, 1]$ that $\mathbf{E}_\theta\left[N < \infty\right]$. This implies $\mathbf{P}_\theta\left[N < \infty\right] = 1$. Therefore, for $\theta \in D_1$,

$$\mathbf{P}_\theta\left[\delta = 1\right] \geq \mathbf{P}_\theta\left[N < \tau_1(\alpha_1)\right] \geq \mathbf{P}_\theta\left[N < \infty, \tau_1(\alpha_1) = \infty\right]$$

$$\geq \mathbf{P}_\theta\left[\tau_1(\alpha_1) = \infty\right] \geq 1 - \alpha_1,$$

leading to

$$\mathbf{P}_\theta\left[\delta = 2\right] \leq \alpha_1.$$

Analogously, $\mathbf{P}_\theta\left[\delta = 1\right] \leq \alpha_2$ if $\theta \in D_2$.

3.) From Lemma 3.2 in [2], we obtain exactly the same bounds as stated in 2.) for $\theta \in (0, 1)$ as lower bounds for $N^{'}$ if $(N^{'}, \delta^{'}) \in K(\alpha, t_1, t_2)$. Thus, only the cases $\theta = 1$ and $\theta = 0$ are missing. Consider $\theta = 1$ and assume there is a decision test in $K(\alpha, t_1, t_2)$ such

that $N' \leq (1-\epsilon)\frac{\log|\alpha_1|}{\log\frac{1}{p_1}}$ for any $\epsilon > 0$. Then, we have

$$\mathbf{P}_\theta[\delta = 2] \geq p_1^{(1-\epsilon)\frac{\log|\alpha_1|}{\log\frac{1}{p_1}}} \geq \alpha_1^{1-\epsilon} > \alpha_1,$$

for $\theta = p_1 \in D_1$, a contradiction. The same argument works in the case $\theta = 0$ and this concludes the proof of the statement 3.) □

References

[1] I.V. Pavlov (1990). A sequential procedure for testing composite hypotheses with application to the Kiefer–Weiss problem. *Theory Probab. Appl., 35:2, 280–292* .

[2] A.G. Tartakovsky (2014). Nearly Optimal Sequential Tests of Composite Hypotheses Revisited. *Proc. Steklov Inst. Math., 287:1, 268-288.*