Analytical \( p \)-value calculation for the higher criticism test in finite-\( d \) problems

BY IAN J. BARNETT AND XIHONG LIN

Department of Biostatistics, Harvard University, Boston, Massachusetts 02115, U.S.A.
ibarnett@hsph.harvard.edu xlin@hsph.harvard.edu

SUMMARY

The higher criticism test is effective for testing a joint null hypothesis against a sparse alternative, e.g., for testing the effect of a gene or genetic pathway that consists of \( d \) genetic markers. Accurate \( p \)-value calculations for the higher criticism test based on the asymptotic distribution require a very large \( d \), which is not the case for the number of genetic variants in a gene or a pathway. In this paper we propose an analytical method for accurately computing the \( p \)-value of the higher criticism test for finite-\( d \) problems. Unlike previous treatments, this method does not rely on asymptotics in \( d \) or on simulation, and is exact for arbitrary \( d \) when the test statistics are normally distributed. The method is particularly computationally advantageous when \( d \) is not large. We illustrate the proposed method with a case-control genome-wide association study of lung cancer and compare its power with competing methods through simulations.

Some key words: Empirical process; Genome-wide association study; Higher criticism; Multiple hypothesis testing; Signal detection.

1. Introduction

Higher criticism tests a joint null hypothesis against the alternative hypothesis that signals in a set are sparse. This situation is commonly encountered in genetic association studies, where it is often of interest to jointly test the effects of genetic variants within a gene, network or pathway on a disease or trait (Tzeng & Zhang, 2007; Wu et al., 2011). The higher criticism test adaptively aggregates independent marginal test statistics, and has been shown to be an asymptotically powerful test of the joint null hypothesis when signals are sparse and above the detection boundary (Donoho & Jin, 2004; Arias-Castro et al., 2011). Here, asymptotics refers to the number of test statistics, \( d \), tending to infinity. The higher criticism test statistic is the supremum of a standardized empirical process under the null hypothesis and follows a Gumbel distribution asymptotically, but its convergence is very slow and its \( p \)-value cannot be reliably computed analytically based on asymptotic theory unless \( d \) is very large (Jaeschke, 1979).

Many practical situations that could benefit from the higher criticism test do not have a very large \( d \). For example, in genome-wide association studies, one is often interested in testing for the effect of genetic variants in a gene or a pathway. The number of genetic variants in a gene or pathway is often not large; it is in the dozens for the vast majority of genes across the genome. To test for the joint null hypothesis of no gene or pathway effect, the asymptotic theory-based \( p \)-values using the higher criticism are not applicable because of the very slow convergence rate to the asymptotic distribution. Simulation of the null distribution is computationally burdensome in genome-wide association studies, as tens of thousands of genes of different sizes need to be tested, and a control for multiple comparisons results in very stringent significance levels. For example, obtaining \( p \)-values accurate to the genome-wide significance level of \( 10^{-7} \) for testing \( 10^4 \) genes requires at least \( 10^{11} \) test statistics simulated under the null hypothesis.

In this paper we present an analytical method of accurate \( p \)-value calculation for the higher criticism test in signal detection settings where \( d \) is not large. The proposed method relies neither on asymptotics in \( d \) nor on simulation of the null distribution. We show that the proposed method is exact for an arbitrary \( d \) for normally distributed marginal test statistics, and is computationally fast in the non-large-\( d \) settings commonly encountered in genome-wide association studies. We evaluate the finite-sample performance...
of the proposed method using simulation and demonstrate its effectiveness on data from a case-control lung cancer genome-wide association study.

2. HIGHER CRITICISM AND ITS ASYMPTOTIC DISTRIBUTION

Consider $d$ normally distributed independent test statistics $Z = (Z_1, \ldots, Z_d)^T$ with means $\mu = (\mu_1, \ldots, \mu_d)^T$ and unit variance. We are interested in testing the joint null hypothesis $H_0 : \mu = 0$ against the alternative that $\mu$ is a sparse vector with the number of nonzero entries being $d_0 = d^{1-\beta}$, where $\beta \in (1/2, 1)$ (Donoho & Jin, 2004). Letting $\Phi(t)$ be the survival function of the standard normal distribution and $S(t) = \sum_{j=1}^{d} I_{|Z_j| \geq t}$, the higher criticism test statistic is

$$HC = \sup_{t>0} \left( \frac{S(t) - 2d \Phi(t)}{[2d \Phi(t)(1 - 2\Phi(t))]^{1/2}} \right).$$

Under the null hypothesis, $HC$ follows a Gumbel distribution as $d \to \infty$. For large $d$, gains in power can be made by searching for the supremum over a restricted range of $t$ (Donoho & Jin, 2004). For $0 < \epsilon < \delta < 1$, if the supremum in (1) is taken over the interval $\Phi^{-1}(1-\delta/2) < t < \Phi^{-1}(1-\epsilon/2)$, then, as shown in Jaeschke (1979), we can write

$$\text{pr}(HC \leq x) \approx \exp[-\exp(-x(2 \log \rho)^{1/2} - 2 \log \pi + 2 \log_2 \rho + 2 \log \rho)].$$

where $\rho = 2^{-1} \log[(1-\epsilon)/\epsilon(1-\delta)]$. Jaeschke (1979) showed that $HC$ converges in distribution at a very slow rate of $O((\log d)^{-1/2})$, so the asymptotic approximation (2) is inaccurate for $d$ as large as $10^6$. Hence, accurate higher criticism $p$-values at stringent significance levels for gene- or pathway-level analysis in genome-wide association studies must be computed without making use of asymptotic theory.

In genetic association studies, the individual marker test statistics $Z_j$ within a gene or genetic pathway are often correlated, with covariance $\Sigma$, say, which can be estimated from the genotypes of a study sample. Letting $UU^T = \Sigma$ be the Cholesky decomposition, under the joint null hypothesis the transformed statistics $U^{-1}Z$ are uncorrelated standard normal random variables, and so the higher criticism can be applied directly to these transformed test statistics (Hall & Jin, 2010). This is appropriate only when the sample size is larger than $d$, which is often the case in gene- or pathway-level analysis in genome-wide association studies.

3. ESTIMATION OF $p$-VALUES FOR THE HIGHER CRITICISM TEST IN FINITE-$d$ SETTINGS

The higher criticism test rejects the joint null hypothesis for large values of $HC$. In this section, we show that finding the supremum does not require an exhaustive search over all $t > 0$. Let $HC(t) = \{S(t) - 2d \times \Phi(t)[2d \Phi(t)(1 - 2\Phi(t))]^{-1/2}$. Then $HC(t)$ is a piecewise-increasing function with a local maximum at each observed $|Z_j|$. Hence, calculating the supremum in the higher criticism test statistic entails finding a maximum over only $d$ quantities. Specifically, let $h$ be the observed $HC$ statistic in (1). Letting $c(t \mid h) = h[2d \Phi(t)(1 - 2\Phi(t))]^{1/2} + 2d \Phi(t)$, the $p$-value corresponding to a given observed higher criticism statistic $h$ is

$$\text{pr} \left\{ \sup_{t>0} HC(t) \geq h \right\} = 1 - \text{pr} \left[ \bigcap_{t>0} \left\{ S(t) < c(t \mid h) \right\} \right].$$

At first glance, evaluating (3) seems to require determining the probability of an intersection of an uncountable number of events, one for each $t > 0$. Without having asymptotics in $d$, we can instead use the fact that $S(t)$ is binomially distributed and can take only values $0, \ldots, d$. This will reduce the intersection in (3) to an intersection over a finite number of events, as defined by the partition given in Lemma 1.

**Lemma 1.** There exists a partition of the positive real line, $0 = t_0 < \cdots < t_{d+1} = \infty$, such that $c(t \mid h) > d$ for $t_0 < t < t_1$ and $d - k < c(t \mid h) \leq d - k + 1$ for $t_k \leq t < t_{k+1}$ for each $k = 1, \ldots, d$.

The proof of Lemma 1 is given in the Appendix. Lemma 1 asserts that, as a function of $t$, $c(t \mid h)$ equals $d$ when $t = 0$ and then increases to a global maximum before decreasing towards an asymptote at $0$; see Fig. 1. The form of $c(t \mid h)$ is the same for all $h > 0$, so in each case the partition in Lemma 1 exists.
We can ignore the case where \( h = 0 \), because the \( p \)-value then equals 1. For \( h > 0 \), using Lemma 1, the following theorem simplifies (3) to the joint probability of a finite intersection, which is computationally feasible to obtain.

**Theorem 1.** If \( 0 = t_0 < \cdots < t_{d+1} = \infty \) is the partition given by Lemma 1, then

\[
\Pr \left( \bigcap_{t>0} \{ S(t) < c(t \mid h) \} \right) = \Pr \left( \bigcap_{k=1}^d \{ S(t_k) \leq d - k \} \right).
\]

According to Theorem 1, for the partition given in Lemma 1, (3) simplifies to

\[
1 - \Pr \left( \bigcap_{k=1}^d \{ S(t_k) \leq d - k \} \right).
\]

For given \( d \) and \( h \), this partition from Lemma 1 is obtained by solving, for each \( k = 1, \ldots, d \), the equation \( c(t_k \mid h) = d - k + 1 \) for \( t_k \). The result is

\[
t_k = \Phi^{-1} \left[ 1 - \frac{2(d - k + 1) + h^2 - h[h^2 + 4(d - k + 1) - 4(d - k + 1)^2/(h^2 + d)]^{1/2}}{4h^2/d} \right],
\]

which defines the partition.

Evaluating (4) directly is difficult because the \( d \) events in the intersection are not independent. Instead, by the chain rule for conditioning, the \( p \)-value can be written as

\[
1 - \Pr \left( \bigcap_{k=1}^d \{ S(t_k) \leq d - k \} \right) = 1 - \prod_{k=1}^d \Pr \left( S(t_k) \leq d - k \mid \bigcap_{l=1}^{k-1} \{ S(t_l) \leq d - l \} \right).
\]

Empirical processes have the Markov property (Gaenssler, 1983, p. 3); so, conditional on \( S(t_{k-1}) = m_{k-1}, \ldots, S(t_1) = m_1, S(t_k) \) is binomial with denominator \( m_{k-1} \) and probability \( \tilde{\Phi}(t_k)/\tilde{\Phi}(t_{k-1}) \). Information about \( S(t_{k-2}), \ldots, S(t_1) \) has no bearing on the distribution of \( S(t_k) \) if \( S(t_{k-1}) \) is known. We can use this fact to compute the terms in the product of (6) by further conditioning on \( S(t_{k-1}) \) in the \( k \)th term. Letting \( q_{k,a} = \Pr[S(t_k) = a \mid S(t_{k-1}) \leq d - k + 1, \ldots, S(t_1) \leq d - 1] \), after some calculations one can show that

\[
q_{k,a} = \sum_{m=0}^{d-k+1} \Pr[S(t_k) = a \mid S(t_{k-1}) = m] \frac{q_{k-1,m}}{\sum_{l=0}^{d-k+1} q_{k-1,l}}
\]

\[
= \sum_{m=0}^{d-k+1} I_{a \leq m} \left( \frac{m}{a} \right) \tilde{\Phi}(t_k)/\tilde{\Phi}(t_{k-1})^a (1 - \tilde{\Phi}(t_k)/\tilde{\Phi}(t_{k-1}))^{m-a} \frac{q_{k-1,m}}{\sum_{l=0}^{d-k+1} q_{k-1,l}}.
\]
From (7), in order to compute \( q_{k,a} \), only knowledge of \( q_{k-1,m} \) for \( m = 0, \ldots, d - k + 1 \) is required. Because \( q_{1,a} = \Pr[S(t_1) = a] \) is a binomial probability, \( q_{1,a} \) offers a base case for calculating the \( p \)-value by computing each \( q_{k,a} \) for \( k = 1, \ldots, d \) and \( a = 0, \ldots, d - k \).

The main result of this paper, Theorem 2, integrates the \( q_{k,a} \) into the exact analytical \( p \)-value calculation of higher criticism for an arbitrary \( d \).

**Theorem 2.** For the partition in (5), the probabilities \( q_{k,a} \) from (7), and the observed higher criticism statistic \( h \), the \( p \)-value for the higher criticism test statistic (1) is

\[
\Pr(\text{HC} \geq h) = 1 - \prod_{k=1}^{d} \sum_{a=0}^{d-k} q_{k,a}.
\]

**Proof.** The result follows immediately from the definition of \( q_{k,a} \), combined with Theorem 1, equation (3) and equation (6). \( \square \)

**Remark 1.** Obtaining the higher criticism \( p \)-value analytically in finite samples is a three-step process. First, the observed test statistic \( h \) is computed by finding the supremum in (1), i.e., the maximum value attained over all observed test statistics \(|Z_i|\). Upon computing \( h \), the partition in (5) is computed. Lastly, the \( q_{k,a} \) are calculated using this partition. As there are \( d(d + 1)/2 \) different \( q_{k,a} \) terms, each requiring a sum of order \( d \) to be calculated, the total computation time for this last step is \( O(d^3) \).

**Remark 2.** The \( p \)-value calculation has been implemented in the statistical computing software R (R Development Core Team, 2014), in the package GHC. The precision of this method, as well as the inaccuracy of the asymptotic \( p \)-values, is confirmed by the results shown in Table 1. The computation time in seconds for a given \( d \) on a 2-30 GHz laptop with 4 GB memory can be well approximated by the polynomial \((3.69 \times 10^{-5})p - (6.98 \times 10^{-6})p^2 + (3.63 \times 10^{-6})p^3\). For \( d = 10, 50, 200 \) and 1000 this corresponds to 0.007 s, 0.45 s, 28.9 s and 4 hours, respectively. Table 1 also presents the empirical Type I error rates calculated using the asymptotic distribution in (2), which are subject to considerable bias.

### 4. Power simulations

We compare the power of the higher criticism test and competing methods through simulation in the context of genetic association studies. An \( n \times d \) genotype matrix \( G \) is generated such that the rows are independent and the columns are autocorrelated with correlation parameter \( \rho \); here \( n \) is the sample size. Marginally, each \( G_{ij} \sim \text{Bi}(2, 0.3) \). Let \( \epsilon \) be the \( n \)-vector of independent standard normal noise and \( \beta \) the \( d \)-vector of regression coefficients; then the phenotypes are generated according to \( y = G\beta + \epsilon \). The test for the association between the \( j \)th genetic variant and \( y \) is \( Z_j = G_j'(y - \bar{y})/(\hat{\sigma}^2 G_j' G_j)^{1/2} \), where \( G_j \) is the \( j \)th column of \( G \) and \( \hat{\sigma}^2 \) is the sample variance.

Power is calculated for a region of size \( d = 40 \), with 10% and 5% sparsity and with autocorrelation \( \rho = 0, 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35 \). In each setting, power estimates were obtained from 1000 iterations. For each iteration, causal variants were selected at random. If the \( j \)th variant is causal, it has \( \beta_j = 0.11 \) for the 5% sparsity case and \( \beta_j = 0.08 \) for the 10% sparsity case. If the \( j \)th variant is not causal, then \( \beta_j = 0 \). In the cases where \( \rho > 0 \), the test statistics must first be decorrelated, as in Hall & Jin (2010),
so that higher criticism is applied instead to \( U^{-1/2}Z \). The sequence kernel association test of Wu et al. (2011) is provided as a comparison alongside the standard likelihood ratio test, and the results are shown in Fig. 2.

The higher criticism test has the highest power in the higher-sparsity situations, while the likelihood ratio test and the sequence kernel association test have higher power for lower sparsity. The sequence kernel association test seems to benefit greatly from increased correlation, whereas the higher criticism test loses power as \( \rho \) increases, because the transformation \( U^{-1/2}Z \) deflates some of the signal for larger \( \rho \), leading to lower power. Overall, it seems that all three methods can be viable. Higher criticism is a good complement to the other methods, and it performs better in the presence of weak correlation among the variants and sparse signals.

5. Data analysis

We apply our \( p \)-value calculation method for the higher criticism test to a case-control lung cancer genome-wide association study conducted at the Massachusetts General Hospital, which aims to identify genes associated with the risk of non-small-cell lung cancer. The study consists of a total of 984 cases and 970 controls. We use the higher criticism test to test for the association of lung cancer risk and each gene, which consists of multiple genetic variants. We analyse 14 396 genes throughout the genome, with \( d = 20 \) genetic variants per gene on average; 92% of the genes have \( d < 50 \) genetic variants, with the largest \( d \) being 1665.

For each gene, we calculate the marginal test statistics \( Z_j (j = 1, \ldots, d) \) for genetic variant \( j \) by fitting a logistic regression model of case-control status on the genetic variant \( j \) while controlling for age, sex, smoking status, and the principal components for population stratification (Price et al., 2006). As in § 3, the marginal test statistics corresponding to the genetic variants within the same gene are decorrelated using the approach of Hall & Jin (2010).

Three of the most significant genes in this analysis are CHRNA5, MYH10 and CLPTM1L, which have been independently found to be associated with lung cancer in other studies (Spitz et al., 2008; Zienolddiny et al., 2009; Wang et al., 2012). For the CHRNA5, MYH10 and CLPTM1L genes, the higher criticism \( p \)-values are \( 6.37 \times 10^{-4}, 6.42 \times 10^{-4} \) and \( 1.29 \times 10^{-2} \), respectively; the \( p \)-values of the sequence kernel...
association test are $8.19 \times 10^{-4}$, 0.41 and 1.54 $\times 10^{-5}$, respectively; and the $p$-values of the likelihood ratio test are $4.92 \times 10^{-3}$, 1.96 $\times 10^{-3}$ and $7.05 \times 10^{-4}$, respectively.

As observed in the power simulations of § 3, the higher criticism test and the sequence kernel association test complement each other for these most significant genes. In order to correct for multiple comparisons, the higher criticism test can be used for signal identification in a hierarchical fashion on the 14,396 $p$-values (Donoho & Jin, 2008). A threshold is set at the point where the test statistics attain the supremum for the higher criticism, and all genes with test statistics more extreme than that threshold are declared to be disease-associated. This procedure controls for the false nondiscovery rate (Ahdesmki & Strimmer, 2009). In our case, this procedure selected the four most significant genes.

As a clear example of how the asymptotic distribution of the higher criticism can be wildly inaccurate in finite-$d$ settings, the asymptotic $p$-values for CHRNA5, MYH10 and CLPTM1L were found to be $5.97 \times 10^{-72}$, $3.83 \times 10^{-64}$ and $3.27 \times 10^{-12}$, respectively. These inaccuracies are amplified in the tails of the distribution, which is why the asymptotic $p$-values differ by so many orders of magnitude from the exact $p$-values obtained from Theorem 2.

6. DISCUSSION

The proposed analytical method for calculating the $p$-value of the higher criticism test is exact for arbitrary finite $d$ for normally distributed test statistics. It is computationally fast for the non-large-$d$ settings commonly encountered in gene-level analysis in genome-wide association studies. For nonnormally distributed outcomes, such as binary outcomes in case-control studies, the accuracy of the proposed calculations depends on the accuracy of the normality approximation for individual marginal test statistics. For large samples, which often occur in genome-wide association studies, the normality assumption on individual test statistics holds quite well and the proposed $p$-value calculations for an arbitrary $d$ have high accuracy.

While the vast majority of genes in the genome will have only dozens of genetic variants, there may be a few large genes with values of $d$ in the thousands. For such large genes, simulating the null distribution could be less of a computational burden than the analytical $p$-value calculation given by Theorem 2. Hence, a mixture of the two techniques might yield a faster overall analysis of genome-wide association data. For example, Theorem 2 could be used to calculate $p$-values except when the gene has $d > 500$, in which case simulation of the null distribution would be used to obtain the $p$-value. In the presence of correlation, the decorrelation transformation of Hall & Jin (2010) could dampen the nonnull signals when the correlation between some of the marginal test statistics is moderate or strong. It would be of interest to develop an alternative higher criticism method to account for the correlation more effectively so as to improve the test power.

ACKNOWLEDGEMENT

This work was supported by grants from the U.S. National Cancer Institute and the National Institute of General Medicine.

APPENDIX

Proof of Lemma 1

Recall that $c(t \mid h) = h[2d \Phi(t)(1 - 2\Phi(t))]^{1/2} + 2d \Phi(t)$. By letting $t \to 0$, we see that $h \geq 0$ and the $p$-value in this case is 1. Considering a fixed $h > 0$, we evaluate the behaviour of $c(t \mid h)$. We have $c(0 \mid h) = d$, $\lim_{t \to -\infty} c(t \mid h) = 0$, and the first derivative is

$$
c'(t \mid h) = d\phi(t) \left( \frac{h(4\Phi(t) - 1)}{[2d\Phi(t)(1 - 2\Phi(t))]^{1/2}} - 2 \right).
$$

Letting $t_{\max} = \Phi^{-1}[1 - (1 + d(h^2 + d)^{-1})/4]$, we can see that $c(t \mid h)$ is increasing on $0 < t < t_{\max}$, achieves its maximum at $t_{\max}$, and decreases on $t_{\max} < t < \infty$, approaching 0. This, along with $c(0 \mid h) = d$
and $c(t \mid h)$ being continuous, gives the result.

**Proof of Theorem 1**

Rather than take the intersection over all $t > 0$, we need only consider $t \geq t_1$. Because $c(t \mid h) > d$ for $0 < t < t_1$, and since $S(t) \leq d$, $S(t)$ must be less than $c(t \mid h)$ for every $t \in (0, t_1)$ with probability 1. Therefore

$$
\Pr \left( \bigcap_{t \geq t_1} \{ S(t) < c(t \mid h) \} \right) = \Pr \left( \bigcap_{t > 0} \{ S(t) < c(t \mid h) \} \right).
$$

As $S(t)$ is an integer-valued nonincreasing function of $t$, we have that for each $k \in \{1, \ldots, d\}$,

$$
\bigcap_{t < t_k \leq t_{k+1}} \{ S(t) < c(t \mid h) \} = \{ S(t_k) \leq d - k \}.
$$

By using this and breaking the intersection into its partition

$$
\bigcap_{t \geq t_1} \{ S(t) < c(t \mid h) \} = \bigcup_{k=1}^{d} \bigcap_{t_k < t \leq t_{k+1}} \{ S(t) < c(t \mid h) \},
$$

the results follow.

**REFERENCES**


[Received September 2013. Revised May 2014]