A novel exact method for significance of higher criticism via Steck's determinant

Jeffrey C. Miecznikowski[a,*], Jiefei Wang[a], Daniel P. Gaile[a], David L. Tritchler[a,b]

[a]*Department of Biostatistics, SUNY University at Buffalo, 3435 Main St, Buffalo, NY 14214*
[b]*Division of Biostatistics, University of Toronto*

Abstract

In this note we provide a novel straightforward approach to calculating the significance for higher criticism statistics using a general result due to Steck (1971) coined Steck's determinant. This result allows users to directly assess higher criticism significance without the need for simulation or asymptotic results.

Keywords: Higher criticism, Steck's determinant, multiple testing

1. Introduction

Higher criticism (HC) is commonly employed in large scale inference to assess whether there is a rare but weak biological signal present in a dataset. Higher criticism biological applications may include gene expression microarray analysis, genome wide association studies (GWAS), and DNA copy number variation. There are several flavors of HC statistics beginning with John Tukey's original proposal discussed in Tukey (1989, 1994). He proposes

*Corresponding Author
Email addresses: jcm38@buffalo.edu (Jeffrey C. Miecznikowski), jwang96@buffalo.edu (Jiefei Wang), dpgaile@buffalo.edu (Daniel P. Gaile), dlt6@buffalo.edu (David L. Tritchler)

the statistic

$$HC_{N,0.05} = \sqrt{N}\left(\text{Fraction of tests significant at } 5\% - 0.05\right)/\sqrt{0.05 \times 0.95} \tag{1}$$

where N is the total number of tests. The equation in (1) is based on a 0.05 significance level and can be generalized to an arbitrary α as

$$HC_{N,\alpha} = \sqrt{N}\left(\text{Fraction of tests significant at } \alpha - \alpha\right)/\sqrt{\alpha \times (1-\alpha)}. \tag{2}$$

If the overall set of tests is significant then we expect $HC_{N,\alpha}$ to be large for some α and hence the significance of the overall set of tests can be captured via

$$HC_N^* = \max_{\{0 \le \alpha \le \alpha_0\}} HC_{N,\alpha}, \tag{3}$$

where $\alpha_0 \in (0,1)$ is a tuning parameter.

Consider the set of N uncorrelated tests and let the individual p-values be denoted by p_i and the ordered p-values (ascending) be denoted by $p_{(i)}$. Then the higher criticism statistic in (3) can be written as:

$$HC_N^* = \max_{\{1 \le i \le \alpha_0 N\}} HC_{N,i}, \tag{4}$$

where

$$HC_{N,i} \equiv \sqrt{N}\frac{(i/N) - p_{(i)}}{\sqrt{p_{(i)}(1-p_{(i)})}}. \tag{5}$$

In the next section we develop a framework to use higher criticism in a hypothesis test setting.

2. HC in a Testing Framework

Suppose we have test statistics X_i with $i \in \{1, 2, \ldots, N\}$ for individual tests of, say, genes. With higher criticism, we are interested in testing whether all test statistics follow the null distribution e.g., $N(0, 1)$, (or, say, an empirical null distribution) versus the alternative that a small fraction are distributed as something else. That is, we want to test the null hypothesis:

$$H_0^{(N)} : X_i \overset{iid}{\sim} N(0, 1), \quad 1 \leq i \leq N, \tag{6}$$

against an alternative,

$$H_1^{(N)} : \text{at least one } X_i \text{ not } N(0, 1). \tag{7}$$

To use higher criticism as a level α test of (6), we must find a critical value $h(N, \alpha)$ so that

$$Pr_{H_0^{(N)}} \left(HC_N^* > h(N, \alpha) \right) \leq \alpha, \tag{8}$$

or similarly for an observed HC statistic, h, we can calculate a valid p-value, p_h as

$$p_h = Pr_{H_0^{(N)}} \left(HC_N^* > h \right), \tag{9}$$

where h is the observed statistic and level α significance is assigned when $p_h < \alpha$.

In the next section we develop a novel exact method to calculate $h(N, \alpha)$ and/or p_h using Steck's determinant.

3. Steck's Determinant and Higher Criticism

If we let $U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(N)}$ denote the order statistics from an independent and identically distributed sample of size N from a uniform $U(0, 1)$ distribution. Steck (1971)

proved that:

$$Pr(l_i \leq U_{(i)} \leq m_i, i = 1, 2, \ldots, N) = \det(S), \tag{10}$$

for any arbitrary set of $\{l_i\}$ and $\{m_i\}$ such that $l_i \leq l_{i+1}$ and $m_i \leq m_{i+1}$. S is a matrix with elements $S_{ij} = \binom{j}{j-i+1}(m_i - l_j)_+^{j-i+1}$ or 0 according as $j - i + 1$ is non negative or negative across $i = 1, 2, \ldots, N$ and $(x)_+ = max(0, x)$. This result has been used before in a multiple testing framework, see Hutson (2002).

From work in Li et al. (2015) for $1 \leq k_0 < k_1 < N$ let

$$Z_N = \max_{k_0 \leq k \leq k_1} \sqrt{N}\{k/N - U_{(k)}\}/\{U_{(k)}(1 - U_{(k)})\}^{1/2}. \tag{11}$$

Let $C(x) = C(x, \epsilon) = \{x + [\epsilon^2 - \epsilon(\epsilon^2 + 4(1 - x)x)^{1/2}]/2\}/(1 + \epsilon^2)$ and observe that $Z_N \geq b$ if and only if $U_{(k)} \leq C(k/N, b/N^{1/2})$ for at least one $k_0 \leq k \leq k_1$.

From combining (4), (5), (9), (10), and (11) and assuming our p-values are independent and $U(0, 1)$ under the null hypothesis, we have

$$
\begin{aligned}
p_h &= Pr_{H_0^{(N)}}\left(HC_N^* > h\right) && \text{(12a)}\\
&= Pr_{H_0^{(N)}}\left(\max_{\{1\le i\le \alpha_0 N\}} HC_{N,i} > h\right) && \text{(12b)}\\
&= Pr_{H_0^{(N)}}\left(\max_{\{1\le i\le \alpha_0 N\}} \sqrt{N}\frac{(i/N)-p_{(i)}}{\sqrt{p_{(i)}(1-p_{(i)})}} > h\right) && \text{(12c)}\\
&= Pr\left(\max_{\{1\le i\le \alpha_0 N\}} \sqrt{N}\frac{(i/N)-U_{(i)}}{\sqrt{U_{(i)}(1-U_{(i)})}} > h\right) && \text{(12d)}\\
&= Pr\left(Z_N > h\right) \text{ with } Z_N \text{ as in (11) and } k_0 = 1 \text{ and } k_1 = \alpha_0 N && \text{(12e)}\\
&= Pr\left(\left\{U_{(k)} \le C\left(k/N, h/N^{1/2}\right) \text{ for at least one } k \text{ where } 1 \le k \le \alpha_0 N\right\}\right) && \text{(12f)}\\
&= 1 - Pr\left(\left\{U_{(k)} > C\left(k/N, h/N^{1/2}\right) \text{ for all } 1 \le k \le \alpha_0 N\right\}\right) && \text{(12g)}\\
&= 1 - Pr\left(U_{(k)} > l_k, k = 1, 2, \ldots, \alpha_0 N\right) && \text{(12h)}\\
&= 1 - \det(S) && \text{(12i)}
\end{aligned}
$$

where S is a matrix with $S_{ij} = \binom{j}{j-i+1}(1-l_j)_+^{j-i+1}$ and $l_j = C(j/N, h/N^{1/2})$. To derive (12h) from (12g), we need to show $\{l_k\}$ is an increasing sequence (see proof in appendix). In the next section we explore computational methods to compute the determinant of S.

4. Computational Methods to Evaluate Steck's Determinant

In `R` there are several methods to calculate the determinant of S; one can simply use the `det` function or QR decomposition (QR factorization) of S via the `qr` function The above methods work well when the dimension of S is relatively small ($N < 20$). However, they will be inaccurate when N is larger. The determinant of S since it estimates a p-value should be bounded between 0 and 1. However using the `det` or `qr` functions in `R` one can easily get an

arbitrarily large determinant when N is large. The breakdown occurs since the computation of a determinant of an N-dimensional matrix requires the multiplication of N numbers by the Leibniz formula (Hazewinkel, 1987). If the S matrix contains a small rounding error then that error will propagate through the multiplication resulting in an arbitrarily large determinant. Therefore calculating the S matrix in a double precision format such as in `R` will not be accurate for large N.

In order to compute an accurate determinant of S, the rounding error in computing elements of S should be controlled or eliminated. MATLAB has a symbolic math toolbox with a symbolic data format that has the capability to store and compute numbers with infinite precision (MATLAB, 2010). Therefore with symbolic computation of S in MATLAB the rounding error can be eliminated and the determinant will be accurate. However there is a memory computing cost involved with a symbolic framework such that for N larger than 800, the memory usage is overwhelmed in many computing environments.

To overcome the memory issue we note that S is an upper Hessenberg matrix. That is, S is 1 on the first subdiagonal and has zeroes for all entries below the first subdiagonal. That

is, S follows the form

$$S = \begin{bmatrix} S_{1,1} & S_{1,2} & S_{1,3} & ... & S_{1,N-1} & S_{1,N} \\ 1 & S_{2,2} & S_{2,3} & ... & S_{2,N-1} & S_{2,N} \\ 0 & 1 & S_{3,3} & ... & S_{3,N-1} & S_{3,N} \\ 0 & 0 & 1 & ... & S_{4,N-1} & S_{4,N} \\ ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & 1 & S_{N,N} \end{bmatrix} \tag{13}$$

We can transform S via elementary operations (see Chapter 4 in Healy (2000)) to a matrix with a straightforward determinant. Firstly we zero-out the first row of S, $S_{1\cdot}$, except for a single non-zero element $S^*_{1,n}$ by adding multiples of the other rows of S (see elementary operations in Chapter 4 in Healy (2000)). That is, let $S^*_{1\cdot}$ be the first row of S^* given by $S^*_{1\cdot} = [0, 0, 0, ..., 0, S^*_{1,n}]$ and S^* is:

$$S^* = \begin{bmatrix} 0 & 0 & 0 & ... & 0 & S^*_{1,N} \\ 1 & S_{2,2} & S_{2,3} & ... & S_{2,N-1} & S_{2,N} \\ 0 & 1 & S_{3,3} & ... & S_{3,N-1} & S_{3,N} \\ 0 & 0 & 1 & ... & S_{4,N-1} & S_{4,N} \\ ... & ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & 1 & S_{N,N} \end{bmatrix} \tag{14}$$

Note $S^* = S$ except for the first row. For example, the zero in $S^*_{1,1}$ is obtained via $S_{1\cdot} - S_{11} \times S_{2\cdot}$. Similarly we zero out the other elements $S_{1\cdot}$. Importantly, we only need 2 rows in memory for each elementary operation thus greatly reducing the memory required in

MATLAB. Also, note since S^* was obtained from S via elementary operations of adding a multiple of one row to another row in S, $\det(S) = \det(S^*)$.

Second, we rearrange the columns of the S^* matrix. Let S^{**} be the matrix obtained after permuting the first and last columns of S^*. Therefore S^{**} is given by

$$S^{**} = \begin{bmatrix} S^*_{1,N} & 0 & 0 & ... & 0 & 0 \\ S_{2,N} & 1 & S_{2,2} & ... & S_{2,N-2} & S_{2,N-1} \\ S_{3,N} & 0 & 1 & ... & S_{3,N-2} & S_{3,N-1} \\ S_{4,N} & 0 & 0 & ... & S_{3,N-2} & S_{3,N-1} \\ ... & ... & ... & ... & ... & ... \\ S_{N,N} & 0 & 0 & ... & 0 & 1 \end{bmatrix} \tag{15}$$

Since permuting columns does not change the determinant we have

$$\det(S) = \det(S^*) = \det(S^{**}) = S^*_{1,N}. \tag{16}$$

In the following section we perform a simulation to compare our exact approach with an asymptotic approach presented in Donoho and Jin (2004).

5. Simulation

We perform a simulation to compare our proposed exact critical value based on Steck's determinant with an asymptotic critical value. The asymptotic approach for $N \to \infty$ is derived in Jaeschke (1979); Donoho and Jin (2004). In the asymptotic approach, the critical

value $h(N, \alpha)$ is given by

$$h(N, \alpha) = b_N^{-1} \left(c_n - \log\log \left(\frac{1}{1-\alpha} \right) \right), \tag{17}$$

with $b_N = \sqrt{2 \log\log(N)}$ and $c_N = 2\log\log(N) + (1/2)(\log\log\log(N) - \log(4\pi))$.

In our simulation for a fixed N and α, we compute the p-values for a range of HC statistics via (16) with symbolic representation in MATLAB. Then we choose the statistic with a p-value nearest to α as the α level critical value. We also compute the asymptotically derived critical value in (17) and compare the exact and asymptotic critical values. As (17) is an asymptotic result we expect that our exact method will be superior in terms of performance with small to moderate N. Figure 1 confirms the asymptotic results are inaccurate for small N but become more accurate as N and α increase. Further, the asymptotic critical values are smaller than the exact critical values inflating the actual significance. Figure 1 also includes the mean Monte Carlo (MC) estimates with 95% confidence bands based on 1000 MC resamplings. As expected, the MC estimates are in line with the exact method. Note a single MC critical value estimate for a given N and α is obtained by repeatedly drawing N observations from a uniform distribution on (0,1) to act as the p_i, calculating HC_N^* for each vector of p-values and then empirically determining the αth percentile of HC_N^*s. The 95% confidence bands are obtained by repeating the MC procedure 1000 times and taking the $5th$ and $95th$ percentiles.

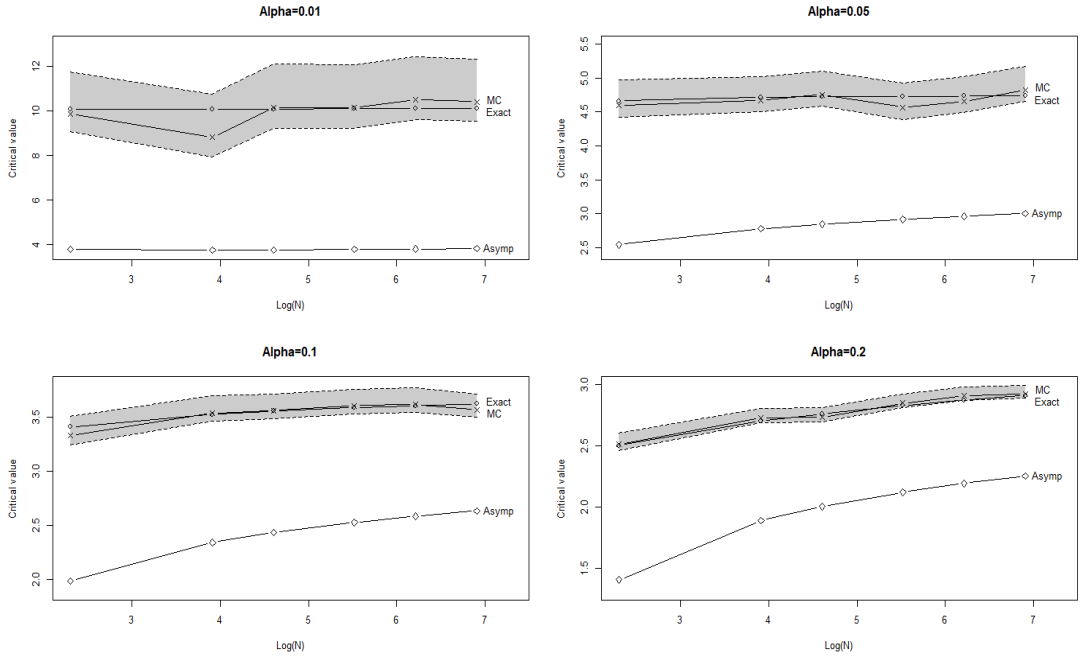

Figure 1: The plot of exact, asymptotic (Asymp) and Monte Carlo (MC) critical values for different choices of N and α. The shaded region provides the estimated 95% confidence intervals for the MC estimates.

6. Application

For an example of our approach we use data from García-Arenzana et al. (2014) where the authors tested associations of 25 dietary variables with mammographic density, a known risk factor for breast cancer. The p-values for association of each dietary variable with mammographic density are presented in Table 2.

For this example, HC^* is 6.17 and the p-value via our exact approach is 0.028 indicating higher criticism at level of 0.05. Note with $N = 26$ the asymptotic p-value is 0.0002, inflating the actual significance. This study is further discussed in light of other multiple testing concerns in McDonald (2009).

Dietary variable	p-value
Total calories	0.001
Olive oil	0.008
Whole milk	0.039
White meat	0.041
Proteins	0.042
Nuts	0.06
Cereals and pasta	0.074
White fish	0.205
Butter	0.212
Vegetables	0.216
Skimmed milk	0.222
Red meat	0.251
Fruit	0.269
Eggs	0.275
Blue fish	0.34
Legumes	0.341
Carbohydrates	0.384
Potatoes	0.569
Bread	0.594
Fats	0.696
Sweets	0.762
Dairy products	0.94
Semi-skimmed milk	0.942
Total meat	0.975
Processed meat	0.986

Table 1: The p-values for the breast cancer risk study in García-Arenzana et al. (2014).

7. Discussion

As noted earlier S is an (upper) Hessenberg matrix and from results in Kaygısız and Sahin (2012) there is a recursive formula to compute the determinant for Hessenberg matrices. We plan to explore the computational advantage, if any, to using the recursive formula to compute the determinant of S. From our simulations the approach of calculating the determinant via (16) in MATLAB with symbolic representation works well for $N < 1500$ while larger N will take substantial computation time. We note that for large N the asymptotic approximation in (17) works well. Also, underlying our method is the assumption that the individual p-values for each test under the null are independent and identically distributed as $U(0,1)$ random variables. We are currently investigating the robustness of our technique to deviations from this assumption.

8. Conclusion

In this note we propose a simple straightforward novel method to compute the significance of higher criticism useful in large scale inference. This avoids the need for asymptotic approximations and works well for N up to 5000.

9. Appendix

Here we provide a proof to show $l_k \leq l_{k+1}$ for all $k = 1, 2, ..., \alpha_0 N$ where $l_k = C(k/N, b/N^{1/2})$ with b, N as positive constants. Showing that $l_k \leq l_{k+1}$ is equivalent to showing that $C(x, \epsilon)$ is a non-decreasing function of x when $x \in (0, 1)$. Let $D(x, \epsilon) = \{x + [\epsilon^2 - \epsilon(\epsilon^2 + 4(1-x)x)^{1/2}]/2\}$, then $C(x, \epsilon) = \{x + [\epsilon^2 - \epsilon(\epsilon^2 + 4(1-x)x)^{1/2}]/2\}/(1+\epsilon^2) = D(x, \epsilon)/(1+\epsilon^2)$. It is clear that

if $D(x,\epsilon)$ is a non-decreasing function of x, then $C(x,\epsilon)$ is a non-decreasing function of x. We examine the first derivative of $D(x,\epsilon)$:

$$\frac{\partial D(x,\epsilon)}{\partial x} = 1 - \frac{\epsilon(1-2x)}{(\epsilon^2+4x-4x^2)^{1/2}}. \tag{18}$$

Setting $\frac{\partial D(x,\epsilon)}{\partial x} = 0$ yields

$$\frac{\epsilon(1-2x)}{(\epsilon^2+4x-4x^2)^{1/2}} = 1 \tag{19a}$$

$$(\epsilon^2+4x-4x^2)^{1/2} = \epsilon(1-2x) \tag{19b}$$

$$\epsilon^2+4x-4x^2 = \epsilon^2(1-2x)^2 \tag{19c}$$

$$\epsilon^2+4x-4x^2 = \epsilon^2-4x\epsilon^2+4x^2\epsilon^2 \tag{19d}$$

$$x(1+\epsilon^2-x(1+\epsilon^2)) = 0 \tag{19e}$$

Solving equation (19e), we have $x = 0$ or 1. However, if we plug $x = \{0,1\}$ into equation (18), only $x = 0$ yields a partial derivative equal 0. Note $x = 1$ has $\frac{\partial D(1,\epsilon)}{\partial x} = 1 + \epsilon/\sqrt{\epsilon^2} \neq 0$. Therefore $D(x,\epsilon)$ is a monotone function on the interval $(0,+\infty)$. To prove that $D(x,\epsilon)$ is an increasing function, one can simply evaluate the derivative at, say, $x = 1/2$ and observe that the partial derivative is positive. Thus $C(x,\epsilon)$ is an increasing function of x on the interval $(0,+\infty)$ and so $l_k \leq l_{k+1}$ for all $k = 1,2,...,\alpha_0 N$.

References

Donoho, D., Jin, J., 2004. Higher criticism for detecting sparse heterogeneous mixtures. Annals of Statistics, 962–994.

García-Arenzana, N., Navarrete-Muñoz, E. M., Lope, V., Moreo, P., Vidal, C., Laso-Pablos, S., Ascunce, N., Casanova-Gómez, F., Sánchez-Contador, C., Santamariña, C., et al., 2014. Calorie intake, olive oil consumption and mammographic density among spanish women. International Journal of Cancer 134 (8), 1916–1925.

Hazewinkel, M., 1987. Encyclopaedia of mathematics.

Healy, M. J., 2000. Matrices for statistics. Oxford University Press.

Hutson, A. D., 2002. Exact bootstrap confidence bands for the quantile function via Steck's determinant. Journal of Computational and Graphical Statistics 11 (2), 471–482.

Jaeschke, D., 1979. The asymptotic distribution of the supremum of the standardized empirical distribution function on subintervals. The Annals of Statistics, 108–115.

Kaygısız, K., Sahin, A., 2012. Determinant and permanent of Hessenberg matrix and Fibonacci type numbers. Gen. Math. Notes 9 (2), 32–41.

Li, J., Siegmund, D., et al., 2015. Higher criticism: p-values and criticism. The Annals of Statistics 43 (3), 1323–1350.

MATLAB, 2010. version 7.10.0 (R2010a). The MathWorks Inc., Natick, Massachusetts.

McDonald, J. H., 2009. Handbook of Biological Statistics. Vol. 2. Sparky House Publishing Baltimore, MD.

Steck, G., 1971. Rectangle probabilities for uniform order statistics and the probability that

the empirical distribution function lies between two distribution functions. The Annals of Mathematical Statistics 42 (1), 1–11.

Tukey, J., 1989. Higher criticism for individual significances in several tables or parts of tables. Princeton University, Princeton (Internal working paper).

Tukey, J. W., 1994. The collected works of John W. Tukey. Vol. 1. Taylor & Francis.