The Generalized Higher Criticism for Testing SNP-sets in Genetic Association Studies

Ian Barnett, Rajarshi Mukherjee & Xihong Lin
Harvard University
ibarnett@hsph.harvard.edu

June 24, 2014
Background

- Genome-wide association studies (GWAS): millions of common (minor allele frequency $> 0.05$) SNPs genotyped.
- Gene-level/pathway-level analysis can provide power to detect these types of effects by combining information over the SNPs.
- Goal: Develop powerful, computationally efficient, statistical methodology for SNP-sets that have the power to detect joint SNP effects.
Model

- \( n \) subjects, \( q \) covariates, \( p \) genetic variants.
- \( Y_\cdot_i \) is phenotype for \( i \)th individual
- \( X_{\cdot i} \) contains \( q \) covariates for \( i \)th individual
- \( G_{\cdot i} \) contains SNP information (minor allele counts) in a gene/pathway/SNP-set for \( i \)th individual
- \( \alpha \) and \( \beta \) contain regression coefficients.
- \( \mu_i = E(Y_i|G_{\cdot i}, X_{\cdot i}) \)

\[
h(\mu_i) = X_{\cdot i}\alpha + G_{\cdot i}\beta
\]

- \( h(\cdot) \) is the link function.
The marginal score test statistic for the \( j \)th variant is:

\[
Z_j = \mathbf{G}_j^T (\mathbf{Y} - \hat{\mu}_0)
\]

where \( \hat{\mu}_0 \) is the MLE of \( E(\mathbf{Y}|H_0) \). Assume \( Z_j \) is normalized.

Letting \( \mathbf{U}\mathbf{U}^T = \widehat{\text{Cov}}(\mathbf{Z}) = \hat{\Sigma} \), define the transformed (decorrelated) test statistics:

\[
\mathbf{Z}^* = \mathbf{U}^{-1} \mathbf{Z} \xrightarrow{\mathcal{L}} \text{MVN}(\mathbf{0}, \mathbf{I}_p)
\]
## Current popular methods

<table>
<thead>
<tr>
<th>Method</th>
<th><strong>SKAT</strong></th>
<th><strong>MinP</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>Test statistic</td>
<td>$\sum_{j=1}^{p} Z_j^2$</td>
<td>$\max_j{</td>
</tr>
<tr>
<td>Pros</td>
<td>High power when signal sparsity is low. Accurate p-values can be obtained quickly.</td>
<td>High power when signal sparsity is high.</td>
</tr>
<tr>
<td>Cons</td>
<td>Can have very low power when sparsity is high.</td>
<td>Slightly lower power when sparsity is low. Difficult to obtain accurate analytic p-values.</td>
</tr>
</tbody>
</table>
The higher criticism

Let

\[ S(t) = \sum_{j=1}^{p} 1\{|Z_j| \geq t\} \]

- **Assumes** \( \Sigma = I_p \)
- Under \( H_0 \), \( S(t) \sim \text{Binomial}(p, 2\Phi(t)) \) where \( \Phi(t) = 1 - \phi(t) \) is the survival function of the normal distribution.
- The Higher Criticism test statistic is:

\[ HC = \sup_{t>0} \left\{ \frac{S(t) - 2p\Phi(t)}{\sqrt{2p\Phi(t)(1 - 2\Phi(t))}} \right\} \]
The higher criticism

Histogram of the $Z_i$

$\text{argmax}\{HC(t)\}$

Density

$t$
Adjusting for correlation

Recalling that $Z^* = U^{-1}Z$, let

$$S^*(t) = \sum_{j=1}^{p} \mathbf{1}_{\{|Z_j^*| \geq t\}}$$

- Note that under $H_0$, $S^*(t) \sim \text{Binomial}(p, 2\Phi(t))$ regardless for general correlated $\Sigma$.
- The innovated Higher Criticism test statistic is:

$$iHC = \sup_{t > 0} \left\{ \frac{S^*(t) - 2p\Phi(t)}{\sqrt{2p\Phi(t)(1 - 2\Phi(t))}} \right\}$$
Adjusting for correlation

Cancer Genetic Markers of Susceptibility (CGEM) Breast Cancer GWAS: **FGFR2 gene**

Decorrelating causes iHC to lose power.
### Comparison

<table>
<thead>
<tr>
<th>Method</th>
<th>MinP</th>
<th>SKAT</th>
<th>iHC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robust to signal sparsity</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Robust to correlation/LD structure</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Computationally efficient</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Does not require decorrelating test statistics</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>
*We will also consider the omnibus test, OMNI, in our power simulations. It is based on the minimum p-value of the SKAT, MinP, and GHC.

<table>
<thead>
<tr>
<th>Method</th>
<th>MinP</th>
<th>SKAT</th>
<th>iHC</th>
<th>GHC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robust to signal sparsity</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Robust to correlation/LD structure</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Computationally efficient</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Doesn’t require decorrelating test statistics</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Our contribution: the generalized higher criticism (GHC)

Recall

\[ S(t) = \sum_{j=1}^{p} 1_{\{|Z_j| \geq t\}} \]

- Now we allow \( \Sigma \) to have arbitrary correlation structure.
- \( S(t) \) is no longer binomial. Instead we approximate with Beta-binomial, matching on first two moments.
- The Generalized Higher Criticism test statistic is:

\[
GHC = \sup_{t > 0} \left\{ \frac{S(t) - 2p\bar{\Phi}(t)}{\sqrt{\hat{\text{Var}}(S(t))}} \right\}
\]
The variance estimator $\hat{\text{Var}}(S(t))$

**Theorem 1**

Let $r^n = \frac{2}{p(1-p)} \sum_{1 \leq k < l \leq p} (\Sigma_{kl})^n$ and let $H_i(t)$ be the Hermite polynomials: $H_0(t) = 1$, $H_1(t) = t$, $H_2(t) = t^2 - 1$ and so on. Then

$$\text{Cov} \left( S(t_k), S(t_j) \right) = p \left[ 2 \Phi(\text{max}\{t_j, t_k\}) - 4 \Phi(t_j) \Phi(t_k) \right]$$

$$+ 4p(p - 1)\phi(t_j)\phi(t_k) \sum_{i=1}^{\infty} \frac{H_{2i-1}(t_j)H_{2i-1}(t_k)r^{2i}}{(2i)!}$$

Proof follows from Schwartzman and Lin (2009) where they showed:

$$P(Z_k > t_i, Z_l > t_j) = \Phi(t_i)\Phi(t_j) + \phi(t_i)\phi(t_j) \sum_{n=1}^{\infty} \frac{\Sigma_{kl}^n}{n!} H_{n-1}(t_i)H_{n-1}(t_j)$$
Letting $h$ be the observed GHC statistic:

$$p\text{-value} = pr \left( \sup_{t>0} \left\{ \frac{S(t) - 2p\Phi(t)}{\sqrt{\hat{\text{Var}}(S(t))}} \right\} \geq h \right)$$

There exists $0 < t_1 < \cdots < t_p$, such that

$$p\text{-value} = 1 - pr \left( \bigcap_{k=1}^{p} \{ S(t_k) \leq p - k \} \right)$$
- $\rho_1$: correlation within causal variants.
- $\rho_2$: correlation between causal and noncausal variants.
- $\rho_3$: correlation within non-causal variants.
Data analysis

The National Cancer Institute’s Cancer Genetic Markers of Susceptibility (CGEM) breast cancer GWAS. Sample has 1145 cases, 1142 controls with european ancestry.
Thresholding tests (GHC and MinP) and summing tests (SKAT) are good complements.

Combining these classes of tests in a more principled way (than OMNI) is to use the following test statistic:

\[
\sup_{\gamma, t} \left\{ \sum_{j=1}^{p} |Z_j|^{\gamma} I_{\{|Z_j|>t\}} \right\}
\]

- \(\gamma = 2, \ t = 0 \rightarrow \text{SKAT}\)
- \(\gamma = 0 \rightarrow \text{GHC}\)
- We label this test as OPT.
Simulations $p = 20$, exchangeable correlation $\rho$

- For OPT, the supremum is selected from $t \in (0, 0.5, 1, 1.5, 2, 2.5, 3, 3.5, 4)$ and $\gamma \in (0, 0.5, 1, 1.5, 2)$.
- Non-zero $\beta$ decrease with $\rho$.