Lecture 7: Interaction Analysis

Timothy Thornton and Michael Wu

Summer Institute in Statistical Genetics 2018
Lecture Outline

Beyond main SNP effects

- Introduction to Concept of Statistical Interaction
- Standard Gene-Environment Interaction Testing
- Some More Sophisticated GxE Tests
- Even Fancier Methods – High order Interactions
“Interaction”

▶ “Interaction” means different things to different people:
  ▶ Biological
  ▶ Mechanistic
  ▶ Additive
  ▶ Synergism and Antagonisms
  ▶ Statistical (Primarily “Multiplicative”)
  ▶ Others — a lot of general vagueness

▶ Statistical (multiplicative) interactions: effect modification (one variable changes the effect of the other on outcome); deviation from additivity
**Statistical Interaction**

Multiplicative interactions: combined effect exceeds the additive effects of individual variables

**Example**
Gene-Environment Interactions ($G \times E$)

Complex diseases are caused by interplay of genes & environment. Identification of $G \times E$ aids in disease prevention.
What is environment (E)?

- “Environment” is just as loaded as “Interaction”
- NIEHS (NIH): Basically, chemical exposures or objective measures (e.g. metabolites) – not primary smoking but second hand is OK
- Anything that is not genetics (G): BMI, race, education, gender, diet, etc.
- Treatment?
- Another SNP (Gene-gene interaction, epistasis): main difference between this and GxE is issue of scale (number of pair-wise tests)
- Operationally: often doesn’t matter, but particular scenarios can change assumptions (e.g. independence between E and G)
Marginal Analysis of GxE Interactions

- Idea: Assess statistical interaction between a single exposure of interest and each SNP
- Testing Approaches:
  - Two-way interaction in regression model (standard)
  - Alternative designs
  - Testing joint G and GxE effects
  - Others.
- Multiple comparisons correction: FDR or Bonferroni
Standard 2-way interaction analysis:

Model (quantitative trait):

\[ y_i = \beta_0 + \beta_g G_i + \beta_e E_i + \beta_{ix} G_i E_i + \varepsilon_i \]

Then to test for interaction effect:

\[ H_0 : \beta_{ix} = 0. \]

If \( H_0 \) is true, then \( G \) and \( E \) can have effects (in the presence of each other), but their effects do not modify each other:

\[
\begin{align*}
G_i = 0 \rightarrow E[y_i] &= \beta_0 + \beta_e E_i \\
G_i = 1 \rightarrow E[y_i] &= \beta_0 + \beta_g 1 + \beta_e E_i
\end{align*}
\]

If \( H_0 \) is false (reject null), then total effect of \( G \) and \( E \) differs depending on other variable:

\[
\begin{align*}
G_i = 0 \rightarrow E[y_i] &= \beta_0 + \beta_e E_i \\
G_i = 1 \rightarrow E[y_i] &= \beta_0 + \beta_g + (\beta_e + \beta_{ix}) E_i
\end{align*}
\]
Standard 2-way interaction analysis:

Operationally

Regress $y$ on $G$, $E$ and product of $G$ and $E$. Then can test $H_0 : \beta_{i\times} = 0$ using any 1-df test.

Things to be careful...

- Scale: particularly for continuous $y$
- Interaction testing is harder because the null model still has genetics in it. Under $H_0$

$$y_i = \beta_0 + \beta_g G_i + \beta_e E_i + \varepsilon_i$$

If this model is not correctly specified or captured, then there can be considerable inflation of type I error.
Power of GxE Tests is Low

Power is bad for GxE analysis: Needs many times as many subjects to test for interaction that is equally powerful.

Power as function of sample size: $\alpha = 0.05$ level, disease pop. risk of 0.01%, SNP with MAF of 0.25, environment with prevalence of 20%, both main SNP and interaction effect are 1.25 (OR).
Alternative Strategies?

- Exploit additional assumptions
- Case-only analysis
- Multi-SNP by E Testing (extension of gene/pathway analysis, but harder)
- Intelligently selecting which SNPs to test
- Many more fancy things constantly being developed
Joint Test of G + GxE

Main Idea
Instead of testing just $H_0 : \beta_{ix} = 0$, we test $H_0 : \beta_g = \beta_{ix} = 0$ via 2-df test. **Primarily useful for gene discovery:** significance does not explicitly inform interaction analysis.

References
Consider the data - a binary response $Y$, a binary environmental variable $E$ and a binary gene $G$:

<table>
<thead>
<tr>
<th></th>
<th>$Y = 1$</th>
<th>$Y = 0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E = 1$</td>
<td>$G = 1$</td>
<td>$E = 1$</td>
</tr>
<tr>
<td></td>
<td>112</td>
<td>100</td>
</tr>
<tr>
<td>$E = 0$</td>
<td>112</td>
<td>100</td>
</tr>
<tr>
<td>$G = 0$</td>
<td>64</td>
<td>100</td>
</tr>
<tr>
<td>$G = 0$</td>
<td>112</td>
<td>100</td>
</tr>
</tbody>
</table>
Joint G and GxE Testing

\[
\text{logit}(Pr(Y = 1|G, E)) = \alpha_0 + \alpha_1 G + \alpha_2 E
\]

P-value for \(H_0: \alpha_1 = 0\) is 0.070. Not significant!

\[
\text{logit}(Pr(Y = 1|G, E)) = \beta_0 + \beta_1 G + \beta_2 E + \beta_3 GE
\]

P-value for \(H_0: \beta_3 = 0\) is 0.051. Not significant!

But....

P-value for \(H_0: \beta_1 = \beta_3 = 0\) is 0.029. Significant!
Case-Only Analysis

- Suppose we have case-control study.
- *Case-Only Analysis* involves analyzing *only* the cases.
- **Key Assumption:** Genotype MUST be independent of environment
  - Almost necessarily true for randomized treatment E
  - Often true for traditional exposures (e.g. toxicants, pollution), but can be weird confounding issues
  - Need to be careful for some E like BMI, alcohol use, smoking, etc.
  - Generally: need to consider this situationally and with care
- Assuming the above, then case-only analysis proceeds by looking at the odds-ratio relating environment to genotype.
Case-Only Analysis

\[
\begin{array}{cc|cc}
G = 0 & G = 1 \\
E = 0 & E = 1 & E = 0 & E = 1 \\
Y = 0 & p_{01} & p_{02} & p_{03} & p_{04} \\
Y = 1 & p_{11} & p_{12} & p_{13} & p_{14} \\
\end{array}
\]

For multiplicative interaction:

\[
\logit P(Y = 1 | G, E) = \beta_0 + \beta_g G + \beta_e E + \beta_{ix} G \times E
\]

\[
\exp(\beta_{ix}) = \frac{OR_{11}}{OR_{10} OR_{01}} = \frac{p_{11} p_{14}}{p_{01} p_{04}} = \frac{p_{12} p_{13}}{p_{02} p_{03}}
\]

GxE odds ratio in cases

GxE odds ratio in controls = 1 under G-E independence!!!
Case-Only Analysis

Instead, we model dependency between genotype and environment:

\[
\frac{P(G = 1|E, Y = 1)}{P(G = 0|E, Y = 1)} = \frac{P(Y = 1|G = 1, E)/P(Y = 0|G = 1, E)}{P(Y = 1|G = 0, E)/P(Y = 0|G = 0, E)} \frac{P(Y = 0, G = 1, E)}{P(Y = 0, G = 0, E)}
\]

\[
= \frac{\exp(\beta_0 + \beta_g + \beta_e E + \beta_{ix} E)}{\exp(\beta_0 + \beta_e E)} \frac{P(G = 1|Y = 0, E)}{P(G = 0|Y = 0, E)}
\]

\[
= \exp(\beta_g + \beta_{ix} E) \frac{P(G = 1)}{P(G = 0)}
\]

with last line holding due to G-E independence (in controls).
Power of Case-Only Analysis

Case-only analysis can lead to improved power, but be careful of assumptions.
Multi-SNP by E Interactions

- Instead of looking at one-SNP at a time, can we again conduct analysis at multi-SNP level?
- Idea:
  1. Group SNPs in gene/pathway/region
  2. Test joint interaction between all SNPs and an environmental variable
- Many approaches for main SNP effects are intuitively applicable, but fail!
  - Interaction term $= G \times E$ is correlated with both $E$ and $G$; this makes permutation methods more challenging
  - We have to correctly capture null model
Multi-SNP by E Interactions

Consider the following generalized linear model:

\[ g(\mu_i) = \mathbf{X}_i^T \alpha_1 + \alpha_2 E_i + \mathbf{G}_i^T \alpha_3 + E_i \mathbf{G}_i^T \beta \]

- **Outcome:** \( Y_i \), has distribution from exponential family and \( \mu_i = E(Y_i|\tilde{\mathbf{X}}_i) \).
- **q non-genetic covariates:** \( \mathbf{X}_i \).
- **environmental factor:** \( E_i \).
- **group of p variants:** \( \mathbf{G}_i = (G_{i1}, \cdots, G_{ip})^T \).
- **p G \times E interaction terms:** \( \mathbf{S}_i = (E_i G_{i1}, \cdots, E_i G_{ip})^T \).

We are interested in testing if there is any \( G \times E \):

\[ H_0 : \beta = 0. \]
Averaging/Collapsing Tests for Interactions

Idea: let $G^*$ be a (weighted) average of genotypes within a gene/region/pathway.

To test for main effects:

$$H_{1m} : g(\mu_i) = \alpha_1^* + \alpha_2^* E_i + \alpha_3^* G_i^*$$
$$H_{0m} : \alpha_3^* = 0$$

Can we use it to test for interactions?

$$H_{1x} : g(\mu_i) = \alpha_1^* + \alpha_2^* E_i + \alpha_3^* G_i^* + \beta^* E_i G_i^*$$
$$H_{0x} : \beta^* = 0$$
Bias analysis for Collapsing $G \times E$ tests

**Intuition**

Null model has to be correctly specified for valid inference. Collapsing $G \times E$ tests may not give valid inference as main effects of the SNVs may not be sufficiently accounted for.

**Continuous Outcome:** No, even if $G \perp E$.

- $G$ and $E$ are independent:
  Model for mean of $Y$ is valid;
  Model for variance of $Y$ is **not** valid.

- $G$ and $E$ not independent:
  Model for mean of $Y$ is **not** valid;
  Model for variance of $Y$ is **not** valid.
Bias analysis for Collapsing $G \times E$ tests

Binary Outcome: Yes if disease is rare and $G \perp E$.

- $G$ and $E$ are independent:
  Model for mean of $Y$ is valid;
  Model for variance of $Y$ is valid approximately.

- $G$ and $E$ not independent:
  Model for mean of $Y$ is \textit{not} valid;
  Model for variance of $Y$ is valid approximately.
GESAT: Model

To test if there is any $G \times E$ ($H_0 : \beta = 0$):

\[
H_0 : \text{logit} [P(Y_i = 1|E_i, X_i, G_i)] = X_i^T \alpha_1 + \alpha_2 E_i + G_i^T \alpha_3
\]

\[
H_A : \text{logit} [P(Y_i = 1|E_i, X_i, G_i)] = X_i^T \alpha_1 + G_i^T (\alpha_3 + E_i \beta) + \alpha_2 E_i
\]

In principle, we can do the same thing as with SKAT, but ...

Difficulties

Need to fit null model:

- Need to estimate main effect of variants
- Lots of variants
- LD and rarity make fitting difficult

Modifications are necessary.

GESAT: Extension of SKAT (global test) for GxE
GESAT: Test Statistic

- Assume \((\beta_1, \cdots, \beta_p)^T\) are random and independent with mean zero and common variance \(\tau\).
- Testing \(H_0\) reduces to testing \(H_0 : \tau = 0\).
- Following Lin (1997), the score test statistic is
  \[ T = (Y - \hat{\mu})^T SS^T (Y - \hat{\mu}) = [Y - \mu(\hat{\alpha})]^T SS^T [Y - \mu(\hat{\alpha})]. \]
- \(\hat{\mu} = \mu(\hat{\alpha})\) is estimated under the null model,
  \[ g(\mu_i | X_i, E_i, G_i) = X_i^T \alpha_1 + \alpha_2 E_i + G_i^T \alpha_3 = \tilde{X}_i^T \alpha. \]
- Use ridge regression to estimate \(\alpha\), impose a penalty only on \(\alpha_3\).
- Under \(H_0\), \(T \sim \sum_{v=1}^P d_v \chi_{1}^2\) approximately.
- Invert characteristic function to get p-value (Davies, 1980).
Which SNPs to Test?

- Genome-wide analysis: screen association between all SNPs and outcome
- Candidate genes or pathways (functional groups)
- SNPs with significant main effects
- More sophisticated algorithms: data adaptive procedures that use two-stage screening

Which set to use can influence multiple testing adjustments. Not always clear how many tests to adjust for if considering main effects too.
Additional Work

- Already a lot of work assuming independence
- Can model E better: multi-E analysis
  - Not always clear which E to use: smoking can be yes/no, never/ever, pack-years, cotinine etc.
  - Mixtures of toxicants: many toxicants or exposures happen in conjunction
- Monotonicity constraints
- Omnibus strategies
- Weighted hypothesis testing
- Innovative screening strategies
Higher order interactions

Given that 2-order interactions are already hard to fine, why are we interested in higher order interactions?

- power,
- computational, and
- interpretation,

we should only be interested in higher order interactions when we focus attention on a few targeted regions (e.g. genes), selected because of

- studies (carried out on other data sets),
- biology,
- ...

It is not a surprise that...

- The power is small.
- As such we may want to see these methods as “hypothesis generating” - i.e. we may identify a limited number of interactions that we can follow up on in new studies.
Models

- SNPs as 3 level categorical variables:

<table>
<thead>
<tr>
<th></th>
<th>low</th>
<th>low</th>
<th>low</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>high</td>
<td>low</td>
<td>high</td>
</tr>
<tr>
<td></td>
<td>low</td>
<td>high</td>
<td>high</td>
</tr>
</tbody>
</table>

- Decision tree models.

- Boolean rules like:

  You are at increased risk if you have at least one mutant for SNP1 or two mutants for SNP2.

- Classical interaction model

\[
g[E(Y|G)] = \beta_0 + \beta_1 G_1 + \beta_2 G_2 + \beta_3 G_3 + \beta_4 G_1 G_2 \\
    + \beta_5 G_1 G_3 + \beta_6 G_2 G_3 + \beta_7 G_1 G_2 G_3,
\]
Models

**MDR**  SNPs as 3 level categorical variables:

<table>
<thead>
<tr>
<th></th>
<th>low</th>
<th>low</th>
<th>low</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td>low</td>
<td>high</td>
<td></td>
</tr>
<tr>
<td>low</td>
<td>high</td>
<td>high</td>
<td></td>
</tr>
</tbody>
</table>

**CART**  Decision tree models.

**Logic Regression**  Boolean rules like:

*You are at increased risk if you have at least one mutant for SNP1 or two mutants for SNP2.*

- Classical interaction model

\[
g[E(Y|G)] = \beta_0 + \beta_1 G_1 + \beta_2 G_2 + \beta_3 G_3 + \beta_4 G_1 G_2 + \beta_5 G_1 G_2 G_3 + \beta_6 G_2 G_3 + \beta_7 G_1 G_2 G_3.
\]
Multifactor Dimensionality Reduction

[Hahn et al. (2003) Bioinformatics 19:376–82]
modification of

- Complex interactions are hard to detect because of sparse data via standard parametric models
- Inaccurate parameter estimates and large standard errors with relatively small sample sizes.
- Reduce the dimensionality and identify SNP combinations that lead to high risk of disease.

Hunting for:

<table>
<thead>
<tr>
<th></th>
<th>low</th>
<th>low</th>
<th>low</th>
</tr>
</thead>
<tbody>
<tr>
<td>high</td>
<td>low</td>
<td>high</td>
<td></td>
</tr>
<tr>
<td>low</td>
<td>high</td>
<td>high</td>
<td></td>
</tr>
</tbody>
</table>
MDR

STEP 1: Select Polymorphisms

STEP 2: Calculate Case-Control Ratios for Each Multilocus Genotype

STEP 3: Identify High-Risk Multilocus Genotypes

STEP 4: Cross Validation

High-Risk
Low-Risk
Empty Cell
MDR

For a particular model with $M$ SNPs (or environmental factors):

- 10-fold Cross-validation
  1. Consider each “cell” (if factors are SNPs, there are $3^M$).
  2. On 9/10th of the data decide whether a cell is “high” or “low” risk (for a case-control study the typical cut-off in each cell would be the case/control ratio in the study).
  3. Evaluate the prediction on the remaining 1/10th of the data.
  4. Check how many of the MDR models are the same. Not entirely clear how this is done - if each cell should be consistent, this would work against models that have (m)any cells that are close to 50/50.

- Repeat this a number of times - to achieve stability of the cross-validation. If you have enough computing power, always a good idea.

- Select the model with the lowest prediction error, provided the consistency is better than by chance.
Sporadic breast cancer

200 women with sporadic primary invasive breast cancer with age-matched hospital based controls, 10 estrogen metabolism SNPs
Issues

- While making things binary helps, computation can explode if the number of SNPs in the study is substantial.

- The selected models do not adhere to the usual parsimony that we like in statistics: if a model with, say, 4 factors is $\epsilon$ better than a model with 3 factors, MDR will pick 4 factors. Usually we would prefer 3. Conceivably this could be changed fairly easy. The MDR implementation of cross-validation makes this worse, however (next slide).

- The models are very hard to interpret.

- To me, it would make more sense to identify a smaller number of cells with “extreme high” or “extreme low” risk.
Bias in their implementation of Cross Validation

- Consider the number of models with $M$ SNPs out of a total $T$.

<table>
<thead>
<tr>
<th>$M$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1</td>
<td>10</td>
<td>45</td>
<td>120</td>
<td>210</td>
<td>252</td>
<td>210</td>
<td>120</td>
<td>45</td>
<td>...</td>
</tr>
<tr>
<td>25</td>
<td>1</td>
<td>30</td>
<td>435</td>
<td>4060</td>
<td>27405</td>
<td>142506</td>
<td>593775</td>
<td>2035800</td>
<td>5852925</td>
<td>...</td>
</tr>
</tbody>
</table>

- Imagine what happens if there is no signal, and every model is equally likely, which size would we most likely end up with...

- The consistency reduces this problem a little, but not by much. Think about the situation where there is one SNP with a strong effect...
Take home message well beyond MDR

When using cross-validation for model selection, if the number of models of size $M$ is different for different $M$, you can use cross-validation to find the best model of each size, but you cannot use it to find the best size. You need another test dataset for that!

Even more generally: beware of fancy methods, particularly anything for interaction analysis!!
A sobering note

There likely have been more papers written about methods to identify $G \times E$ and $G \times G$ interactions, than the number of interactions that have successfully been identified.