Fast and accurate construction of confidence intervals for heritability

Regev Schweiger

Joint work with Eran Halperin, Saharon Rosset, Shachar Kaufman and Eleazar Eskin
What is heritability?

- **Heritability** - the proportion of variance of a trait explained by a genetic component

Past: Twin studies - what's the chance of having a disease if my twin has it? Or a continuous trait?

Present: Large cohort studies - What is the effect of genetic similarity on trait similarity?

Prioritizing research on certain diseases, genes, etc.

Some examples (from SNPedia):

<table>
<thead>
<tr>
<th>Disease</th>
<th>Heritability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alzheimer's</td>
<td>58-79%</td>
</tr>
<tr>
<td>Crohn's Disease</td>
<td>53%</td>
</tr>
<tr>
<td>Height</td>
<td>55-81%</td>
</tr>
<tr>
<td>Obesity</td>
<td>70%</td>
</tr>
</tbody>
</table>
What is heritability?

- **Heritability** - the proportion of variance of a trait explained by a genetic component

- **Past: Twin studies** - what’s the chance of having a disease if my twin has it? Or a continuous trait?

- **Present: Large cohort studies** - What is the effect of genetic similarity on trait similarity?

- Prioritizing research on certain diseases, genes, etc.

- Some examples (from SNPedia):
  - Disease | Heritability
  - Alzheimer’s Disease | 58-79%
  - Crohn’s Disease | 53%
  - Height | 55-81%
  - Obesity | 70%
What is heritability?

- **Heritability** - the proportion of variance of a trait explained by a genetic component

- **Past: Twin studies** - what’s the chance of having a disease if my twin has it? Or a continuous trait?

- **Present: Large cohort studies** - What is the effect of genetic similarity on trait similarity?

<table>
<thead>
<tr>
<th>Disease</th>
<th>Heritability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alzheimer's Disease</td>
<td>58-79%</td>
</tr>
<tr>
<td>Crohn's Disease</td>
<td>53%</td>
</tr>
<tr>
<td>Height</td>
<td>55-81%</td>
</tr>
<tr>
<td>Obesity</td>
<td>70%</td>
</tr>
</tbody>
</table>
What is heritability?

- **Heritability** - the proportion of variance of a trait explained by a genetic component

- **Past: Twin studies** - what’s the chance of having a disease if my twin has it? Or a continuous trait?

- **Present: Large cohort studies** - What is the effect of genetic similarity on trait similarity?

- Prioritizing research on certain diseases, genes, etc.

<table>
<thead>
<tr>
<th>Disease</th>
<th>Heritability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alzheimer's</td>
<td>58-79%</td>
</tr>
<tr>
<td>Crohn's Disease</td>
<td>53%</td>
</tr>
<tr>
<td>Height</td>
<td>55-81%</td>
</tr>
<tr>
<td>Obesity</td>
<td>70%</td>
</tr>
</tbody>
</table>
What is heritability?

- **Heritability** - the proportion of variance of a trait explained by a genetic component

- **Past: Twin studies** - what’s the chance of having a disease if my twin has it? Or a continuous trait?

- **Present: Large cohort studies** - What is the effect of genetic similarity on trait similarity?

- Prioritizing research on certain diseases, genes, etc.

- Some examples (from SNPedia):

<table>
<thead>
<tr>
<th>Disease</th>
<th>Heritability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alzheimer’s Disease</td>
<td>58-79%</td>
</tr>
<tr>
<td>Crohn’s Disease</td>
<td>53%</td>
</tr>
<tr>
<td>Height</td>
<td>55-81%</td>
</tr>
<tr>
<td>Obesity</td>
<td>70%</td>
</tr>
</tbody>
</table>
Heritability using the linear mixed model

- A phenotype $y$ is assumed to follow:

$$y = X\beta + Zs + e$$

where:

- $y$ is an $n \times 1$ vector of phenotype measurements for each individual
- $X$ is an $n \times p$ matrix of $p$ covariates (possibly including an intercept vector $1_n$, as well as other covariates such as sex, age, etc.)
- $Z$ is the $n \times m$ standardized genotype matrix, i.e., columns have zero mean and unit variance
- $\beta$ is a $p \times 1$ vector of fixed effects
- $s$ is a $m \times 1$ vector of random effects
- $e$ is an $n \times 1$ vector of errors
Heritability using the linear mixed model

- A phenotype $y$ is assumed to follow:

$$y = X\beta + Zs + e$$

where:

- $y$ is a $n \times 1$ vector of phenotype measurements for each individual
Heritability using the linear mixed model

- A phenotype $y$ is assumed to follow:

$$y = X\beta + Zs + e$$

where:

- $y$ is a $n \times 1$ vector of phenotype measurements for each individual
- $X$ is a $n \times p$ matrix of $p$ covariates (possibly including an intercept vector $1_n$, as well as other covariates such as sex, age, etc.)
Heritability using the linear mixed model

- A phenotype $y$ is assumed to follow:

$$y = X\beta + Zs + e$$

where:

- $y$ is a $n \times 1$ vector of phenotype measurements for each individual
- $X$ is a $n \times p$ matrix of $p$ covariates (possibly including an intercept vector $1_n$, as well as other covariates such as sex, age, etc.)
- $Z$ is the $n \times m$ standardized genotype matrix, i.e., columns have zero mean and unit variance
Heritability using the linear mixed model

- A phenotype $y$ is assumed to follow:

\[ y = X\beta + Zs + e \]

where:

- $y$ is a $n \times 1$ vector of phenotype measurements for each individual
- $X$ is a $n \times p$ matrix of $p$ covariates (possibly including an intercept vector $\mathbf{1}_n$, as well as other covariates such as sex, age, etc.)
- $Z$ is the $n \times m$ standardized genotype matrix, i.e., columns have zero mean and unit variance
- $\beta$ is a $p \times 1$ vector of fixed effects
- $s$ is a $m \times 1$ vector of random effects
- $e$ is a $n \times 1$ vector of errors
Heritability using the linear mixed model

A phenotype $y$ is assumed to follow:

$$y = X\beta + Zs + e$$

where:

- $y$ is a $n \times 1$ vector of phenotype measurements for each individual
- $X$ is a $n \times p$ matrix of $p$ covariates (possibly including an intercept vector $1_n$, as well as other covariates such as sex, age, etc.)
- $Z$ is the $n \times m$ standardized genotype matrix, i.e., columns have zero mean and unit variance
- $\beta$ is a $p \times 1$ vector of fixed effects
- $s$ is a $m \times 1$ vector of random effects
Heritability using the linear mixed model

- A phenotype $y$ is assumed to follow:

$$y = X\beta + Zs + e$$

where:

- $y$ is a $n \times 1$ vector of phenotype measurements for each individual
- $X$ is a $n \times p$ matrix of $p$ covariates (possibly including an intercept vector $1_n$, as well as other covariates such as sex, age, etc.)
- $Z$ is the $n \times m$ standardized genotype matrix, i.e., columns have zero mean and unit variance
- $\beta$ is a $p \times 1$ vector of fixed effects
- $s$ is a $m \times 1$ vector of random effects
- $e$ is a $n \times 1$ vector of errors
Heritability using the linear mixed model

- A phenotype \( y \) is assumed to follow:

\[
y = X\beta + Zs + e
\]

- Assume \( s \) and \( e \) are statistically independent

\[
s \sim N(0, m \sigma_g^2 I_m)
\]

\[
e \sim N(0, \sigma_e^2 I_n)
\]

- Define \( K = ZZ^T \), a kinship matrix capturing the genetic relatedness between \( n \) individuals, then:

\[
y \sim N(X\beta, \sigma_g^2 K + \sigma_e^2 I_n)
\]

- \( \sigma_g^2, \sigma_e^2 \) are genetic and environmental variance components

- The fixed effects \( \beta \) and the coefficients \( \sigma_g^2 \) and \( \sigma_e^2 \) are the parameters of the model.
Heritability using the linear mixed model

- A phenotype $y$ is assumed to follow:

$$y = X\beta + Zs + e$$

- Assume $s$ and $e$ are statistically independent.

$\sigma^2_g$, $\sigma^2_e$ are genetic and environmental variance components.

The fixed effects $\beta$ and the coefficients $\sigma^2_g$ and $\sigma^2_e$ are the parameters of the model.
Heritability using the linear mixed model

- A phenotype $y$ is assumed to follow:

$$y = X\beta + Zs + e$$

- Assume $s$ and $e$ are statistically independent
- $s \sim \mathcal{N}(0_m, \frac{1}{m} \sigma^2_g I_m)$
Heritability using the linear mixed model

- A phenotype $y$ is assumed to follow:

$$ y = X\beta + Zs + e $$

- Assume $s$ and $e$ are statistically independent

- $s \sim \mathcal{N}(0, \frac{1}{m}\sigma_g^2 I_m)$

- $e \sim \mathcal{N}(0, \sigma_e^2 I_n)$. 

$K = Z ZZ^T$, a kinship matrix capturing the genetic relatedness between $n$ individuals, then:

$$ y \sim \mathcal{N}(X\beta, \sigma_g^2 K + \sigma_e^2 I_n) $$

$\sigma_g^2, \sigma_e^2$ are genetic and environmental variance components

The fixed effects $\beta$ and the coefficients $\sigma_g^2$ and $\sigma_e^2$ are the parameters of the model.
Heritability using the linear mixed model

- A phenotype $y$ is assumed to follow:
  \[ y = X\beta + Zs + e \]

- Assume $s$ and $e$ are statistically independent

- $s \sim \mathcal{N} \left( 0_m, \frac{1}{m} \sigma_g^2 I_m \right)$
- $e \sim \mathcal{N} \left( 0_n, \sigma_e^2 I_n \right)$. 

$K = \frac{1}{m} ZZ^T$, a kinship matrix capturing the genetic relatedness between $n$ individuals, then:

$y \sim \mathcal{N} \left( X\beta, \sigma_g^2 K + \sigma_e^2 I_n \right)$

- $\sigma_g^2, \sigma_e^2$ are genetic and environmental variance components
- The fixed effects $\beta$ and the coefficients $\sigma_g^2$ and $\sigma_e^2$ are the parameters of the model.
Heritability using the linear mixed model

- A phenotype $y$ is assumed to follow:
  \[ y = X\beta + Zs + e \]

- Assume $s$ and $e$ are statistically independent
  - $s \sim \mathcal{N}(0_m, \frac{1}{m}\sigma_g^2 I_m)$
  - $e \sim \mathcal{N}(0_n, \sigma_e^2 I_n)$.

- Define $K = \frac{1}{m}ZZ^T$, a *kinship matrix* capturing the genetic relatedness between $n$ individuals, then:
  \[ y \sim \mathcal{N}(X\beta, \sigma_g^2 K + \sigma_e^2 I_n) \]

- $\sigma_g^2$, $\sigma_e^2$ are genetic and environmental variance components
- The fixed effects $\beta$ and the coefficients $\sigma_g^2$ and $\sigma_e^2$ are the parameters of the model.
Heritability using the linear mixed model

- This is a variance components model:

$$ y \sim \mathcal{N} \left( X\beta, \sigma_g^2 K + \sigma_e^2 I_n \right) $$

- Commonly estimated with restricted maximum likelihood (REML) - denoted here

- Caveats:
  - The kinship matrix $K$ is assumed known
  - Continuous phenotype (no case-control for now)
Heritability using the linear mixed model

- This is a variance components model:

\[ y \sim \mathcal{N}(X\beta, \sigma_g^2 K + \sigma_e^2 I_n) \]

- The (narrow-sense, SNP-) heritability is defined as:

\[ h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \]
Heritability using the linear mixed model

- This is a variance components model:

\[ y \sim \mathcal{N}(X\beta, \sigma_g^2 K + \sigma_e^2 I_n) \]

- The (narrow-sense, SNP-) heritability is defined as:

\[ h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \]

- Commonly estimated with *restricted maximum likelihood* (REML) - denoted here \( \hat{h}^2 = \hat{\sigma}_g^2 / (\hat{\sigma}_g^2 + \hat{\sigma}_e^2) \)
Heritability using the linear mixed model

- This is a variance components model:

\[ y \sim \mathcal{N} \left( X\beta, \sigma^2_g K + \sigma^2_e I_n \right) \]

- The (narrow-sense, SNP-) heritability is defined as:

\[ h^2 = \frac{\sigma^2_g}{\sigma^2_g + \sigma^2_e} \]

- Commonly estimated with restricted maximum likelihood (REML) - denoted here \( \hat{h}^2 = \hat{\sigma}^2_g / (\hat{\sigma}^2_g + \hat{\sigma}^2_e) \)

- Caveats:
Heritability using the linear mixed model

- This is a variance components model:
  \[ y \sim \mathcal{N} \left( X\beta, \sigma_g^2 K + \sigma_e^2 I_n \right) \]

- The (narrow-sense, SNP-) heritability is defined as:
  \[ h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \]

- Commonly estimated with *restricted maximum likelihood* (REML) - denoted here \( \hat{h}^2 = \hat{\sigma}_g^2 / (\hat{\sigma}_g^2 + \hat{\sigma}_e^2) \)

- Caveats:
  - The kinship matrix \( K \) is assumed *known*
Heritability using the linear mixed model

- This is a variance components model:

\[ y \sim \mathcal{N}(X\beta, \sigma_g^2 K + \sigma_e^2 I_n) \]

- The (narrow-sense, SNP-) heritability is defined as:

\[ h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \]

- Commonly estimated with restricted maximum likelihood (REML) - denoted here \( \hat{h}^2 = \frac{\hat{\sigma}_g^2}{\hat{\sigma}_g^2 + \hat{\sigma}_e^2} \)

- Caveats:
  - The kinship matrix \( K \) is assumed \textit{known}
  - Continuous phenotype (no case-control for now)
Invariance of heritability estimator

The distribution of $\hat{h}^2$ depends only on $h^2$. We can limit ourselves to the case $\sigma_g^2 + \sigma_e^2 = 1$ and $\beta = 0$. To see that:

- Define $\sigma_p^2 = \sigma_g^2 + \sigma_e^2$, $h^2 = \sigma_g^2 / \sigma_p^2$, $V_{h^2} = h^2 K + (1 - h^2)I_n$ (so $\text{Cov}[y] = \sigma_g^2 K + \sigma_e^2 I_n = \sigma_p^2 V_{h^2}$)
Invariance of heritability estimator

The distribution of $\hat{h^2}$ depends only on $h^2$. We can limit ourselves to the case $\sigma_g^2 + \sigma_e^2 = 1$ and $\beta = 0$. To see that:

- Define $\sigma_p^2 = \sigma_g^2 + \sigma_e^2$, $h^2 = \sigma_g^2 / \sigma_p^2$, $V_{h^2} = h^2 K + (1 - h^2)I_n$ (so $\text{Cov}[y] = \sigma_g^2 K + \sigma_e^2 I_n = \sigma_p^2 V_{h^2}$)

$$
\ell_{\text{REML}}(h^2, \sigma_p^2, \beta) \propto -(n - p) \log \sigma_p^2 - \log |V_{h^2}|
- \log |X^T V_{h^2}^{-1} X| - \frac{(y - X\beta)^T V_{h^2}^{-1} (y - X\beta)}{\sigma_p^2},
$$
Invariance of heritability estimator

The distribution of $\hat{h}^2$ depends only on $h^2$. We can limit ourselves to the case $\sigma_g^2 + \sigma_e^2 = 1$ and $\beta = 0$. To see that:

- Define $\sigma_p^2 = \sigma_g^2 + \sigma_e^2$, $h^2 = \sigma_g^2 / \sigma_p^2$, $V_{h^2} = h^2 K + (1 - h^2)I_n$ (so $\text{Cov}[y] = \sigma_g^2 K + \sigma_e^2 I_n = \sigma_p^2 V_{h^2}$)

$\ell_{REML}(h^2, \sigma_p^2, \beta) \propto -(n - p) \log \sigma_p^2 - \log |V_{h^2}|$

$- \log |X^T V_{h^2}^{-1} X| - \frac{(y - X\beta)^T V_{h^2}^{-1} (y - X\beta)}{\sigma_p^2}$

- For a fixed $\hat{h}^2$, the values of $\sigma_p^2$ and $\beta$ that maximize $\ell_{REML}$ can be derived analytically:

$\hat{\beta}(\hat{h}^2) = (X^T V_{\hat{h}^2}^{-1} X)^{-1} X^T V_{\hat{h}^2}^{-1} y$

$\hat{\sigma}_p^2(\hat{h}^2) = \frac{1}{n - p} (y - X\hat{\beta}(\hat{h}^2))^T V_{\hat{h}^2}^{-1} (y - X\hat{\beta}(\hat{h}^2))$. 
Invariance of heritability estimator

-it can be shown that

\[ \hat{h}^2(y) = \hat{h}^2(\lambda y + X\gamma) \]
\[ \lambda^2 \cdot \hat{\sigma}_p^2(y) = \hat{\sigma}_p^2(\lambda y + X\gamma) \]
\[ \lambda \cdot \hat{\beta}(y) + \gamma = \hat{\beta}(\lambda y + X\gamma) \].

-The distribution of \( \hat{h}^2 \) depends only on \( h^2 \).

-We may limit our study to the \( \hat{h}^2 \) estimator alone, in the special case of fixed \( \sigma_p^2 = 1 \) and \( \beta = 0_p \).
Datasets

- Northern Finland Birth Cohort (NFBC): 2,520 individuals and 331,476 genotyped SNPs

- The Genotype-Tissue Expression (GTEx) study: 185 individuals and 3,575,877 SNPs
  - Gene expression profiles for several tissues

- Ludwigshafen Risk and Cardiovascular Health (LURIC): 867 individuals and 687,262 SNPs
  - Lipid measurements
Example 1 (NFBC)

$ gcta64 -reml ...

<table>
<thead>
<tr>
<th>Source</th>
<th>Variance</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>V(G)</td>
<td>0.405484</td>
<td>0.052382</td>
</tr>
<tr>
<td>V(e)</td>
<td>0.622200</td>
<td>0.050043</td>
</tr>
<tr>
<td>Vp</td>
<td>1.027684</td>
<td>0.020033</td>
</tr>
<tr>
<td>V(G)/Vp</td>
<td>0.394561</td>
<td>0.049171</td>
</tr>
<tr>
<td>logL</td>
<td>-2745.720</td>
<td></td>
</tr>
<tr>
<td>logL0</td>
<td>-2781.759</td>
<td></td>
</tr>
<tr>
<td>LRT</td>
<td>72.079</td>
<td></td>
</tr>
<tr>
<td>df</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Pval</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>2520</td>
<td></td>
</tr>
</tbody>
</table>
Example 1 (NFBC)

$ gcta64 -reml ...$

<table>
<thead>
<tr>
<th>Source</th>
<th>Variance</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>V(G)</td>
<td><strong>0.405484</strong></td>
<td>0.052382</td>
</tr>
<tr>
<td>V(e)</td>
<td><strong>0.622200</strong></td>
<td>0.050043</td>
</tr>
<tr>
<td>Vp</td>
<td><strong>1.027684</strong></td>
<td>0.020033</td>
</tr>
<tr>
<td>V(G)/Vp</td>
<td><strong>0.394561</strong></td>
<td>$\hat{h}^2$ 0.049171</td>
</tr>
</tbody>
</table>

logL        | -2745.720     |
logL0       | -2781.759     |
LRT         | 72.079        |
df          | 1             |
Pval        | 0             |
n            | 2520          |
Example 1 (NFBC)

<table>
<thead>
<tr>
<th>Source</th>
<th>Variance</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V(G)/V_p$</td>
<td>0.394561</td>
<td>$\hat{h}^2$ 0.049171</td>
</tr>
</tbody>
</table>
Constructing CIs assuming normality

- Standard errors imply the construction of confidence intervals (CIs) for the true heritability, $h^2$
Constructing CIs assuming normality

- Standard errors imply the construction of confidence intervals (CIs) for the true heritability, $h^2$
- Assume $\hat{h}^2$ is distributed normally, with a known variance (Fisher information matrix or variants)
Constructing CIs assuming normality

- Standard errors imply the construction of confidence intervals (CIs) for the true heritability, $h^2$
- Assume $\hat{h}^2$ is distributed normally, with a known variance (Fisher information matrix or variants)
- E.g., from $\hat{h}^2$ and a standard error estimate, se, construct a 95% CI ($\hat{h}^2 - 1.96 \cdot se$, $\hat{h}^2 + 1.96 \cdot se$)
Example 1 (NFBC)

<table>
<thead>
<tr>
<th>Source</th>
<th>Variance</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>V(G)/Vp</td>
<td>0.394561</td>
<td>0.049171</td>
</tr>
</tbody>
</table>
Example 1 (NFBC)

<table>
<thead>
<tr>
<th>Source</th>
<th>Variance</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V(G)/Vp$</td>
<td><strong>0.394561</strong></td>
<td><strong>0.049171</strong></td>
</tr>
</tbody>
</table>

The distribution is centered at 0.394561 with a 95% interval from 0 to 1.
Example 1 (NFBC)

<table>
<thead>
<tr>
<th>Source</th>
<th>Variance</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V(G)/Vp$</td>
<td>0.394561</td>
<td>0.049171</td>
</tr>
</tbody>
</table>

Diagram with 95% confidence interval for $\hat{h}^2$ at 0.394561.

0.394561

95%
## Example 2 - Over the boundary (GTEx)

<table>
<thead>
<tr>
<th>Source</th>
<th>Variance</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>V(G)/Vp</td>
<td>0.089128</td>
<td>0.174675</td>
</tr>
</tbody>
</table>
Example 2 - Over the boundary (GTEx)

<table>
<thead>
<tr>
<th>Source</th>
<th>Variance</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V(G)/V_p$</td>
<td>0.089128</td>
<td>0.174675</td>
</tr>
</tbody>
</table>

![Diagram](image_url)
### Example 2 - Over the boundary (GTEx)

<table>
<thead>
<tr>
<th>Source</th>
<th>Variance</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V(G)/Vp$</td>
<td><strong>0.089128</strong></td>
<td><strong>0.174675</strong></td>
</tr>
</tbody>
</table>

![Graph showing variance and standard error over the boundary](image-url)
### Example 3 - Zero heritability (GTEx)

<table>
<thead>
<tr>
<th>Source</th>
<th>Variance</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V(G)/V_p$</td>
<td>0.0000000</td>
<td>0.218484</td>
</tr>
</tbody>
</table>
Example 3 - Zero heritability (GTEx)

<table>
<thead>
<tr>
<th>Source</th>
<th>Variance</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V(G)/V_p$</td>
<td>0.000000</td>
<td>0.218484</td>
</tr>
</tbody>
</table>

The graph shows a downward trend representing $\hat{h}^2$. The table lists the source, variance, and standard error (SE) for different sources.
Example 3 - Zero heritability (GTEx)

<table>
<thead>
<tr>
<th>Source</th>
<th>Variance</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V(G)/V_p$</td>
<td>0.000000</td>
<td>0.218484</td>
</tr>
</tbody>
</table>
Deviation from normality

Assumption of normal distribution for $\hat{h}^2$ requires:

- A large number of phenotypes drawn, or independence between individuals
- True parameter value far from parameter boundaries
- (Other regularity conditions)

Some known results:

Deviation from normality

Assumption of normal distribution for $\hat{h}^2$ requires:

- A large number of phenotypes drawn, or independence between individuals
- True parameter value far from parameter boundaries
- (Other regularity conditions)

Some known results:

Empirical distributions - NFBC

Draw 10,000 phenotypes with true heritability $h^2$; estimate $\hat{h}^2$ for each one; construct histogram

\[ h^2 = 0.1 \]

Bins: 0, (0, 0.01), (0.01, 0.02), \ldots, (0.99, 1), 1
Empirical distributions - NFBC

Draw 10,000 phenotypes with true heritability $h^2$; estimate $\hat{h}^2$ for each one; construct histogram

![Graph showing distribution of estimated heritability for $h^2 = 0.1$ and $h^2 = 0.2$.]

Bins: 0, (0, 0.01), (0.01, 0.02), ..., (0.99, 1), 1
Empirical distributions - NFBC

Draw 10,000 phenotypes with true heritability $h^2$; estimate $\hat{h}^2$ for each one; construct histogram

Bins: 0, (0, 0.01), (0.01, 0.02), ..., (0.99, 1), 1
Empirical distributions - GTEx

Draw 10,000 phenotypes with true heritability $h^2$; estimate $\hat{h}^2$ for each one; construct histogram

Bins: 0, (0, 0.01), (0.01, 0.02), ..., (0.99, 1), 1
Empirical distributions - LURIC

Draw 10,000 phenotypes with true heritability $h^2$; estimate $\hat{h}^2$ for each one; construct histogram

Bins: 0, (0, 0.01), (0.01, 0.02), \ldots, (0.99, 1), 1
Boundary probabilities

Probability of $\hat{h}^2 = 0$

- GTEx
- LURIC
- NFBC

True value of $h^2$ vs Probability

[Graph showing probability against true value of $h^2$ for GTEx, LURIC, and NFBC]
Measuring the accuracy of CI coverage

For a 95% CI,
Measuring the accuracy of CI coverage

For a 95% CI,

\[ h^2 \rightarrow y_1, \ldots, y_N \]
Measuring the accuracy of CI coverage

For a 95% CI,

\[ h^2 \rightarrow y_1, \ldots, y_N \]

\[ \rightarrow (\hat{h}^2(y_1), se(y_1)) \]

\[ \vdots \]

\[ (\hat{h}^2(y_N), se(y_N)) \]

If the CI is accurate, this should happen 95% of the times.
Measuring the accuracy of CI coverage

For a 95% CI,

\[ h^2 \rightarrow y_1, \ldots, y_N \]
\[ \rightarrow (\hat{h}^2(y_1), se(y_1)) \]
\[ \ldots \]
\[ (\hat{h}^2(y_N), se(y_N)) \]
\[ \rightarrow h^2 \in \hat{h}^2(y_1) \pm 1.96 \cdot se(y_1) \]
\[ \ldots \]
\[ h^2 \in \hat{h}^2(y_N) \pm 1.96 \cdot se(y_N) \]
Measuring the accuracy of CI coverage

For a 95% CI,

\[ h^2 \rightarrow y_1, \ldots, y_N \]
\[ \rightarrow (\hat{h}^2(y_1), se(y_1)) \]
\[ \ldots \]
\[ (\hat{h}^2(y_N), se(y_N)) \]
\[ \rightarrow h^2 \in \hat{h}^2(y_1) \pm 1.96 \cdot se(y_1) \]
\[ \ldots \]
\[ h^2 \in \hat{h}^2(y_N) \pm 1.96 \cdot se(y_N) \]

If the CI is accurate, this should happen 95% of the times
CIs are inaccurate in real datasets.

NFBC, GCTA’s CI

Coverage Probability vs. True value of $h^2$
CIs are inaccurate in real datasets

![Graph showing coverage probability for NFBC, GCTA's CI]

**Coverage Probability**

- 95% CI
- 90% CI

**True value of \( h^2 \)**

- 80%
- 85%
- 90%
- 95%
- 100%
CIs are inaccurate in real datasets

---

**GTEx, GCTA’s CI**

- Coverage Probability
- True value of $h^2$
CIs are inaccurate in real datasets.

![Graph showing coverage probability for GTEx and GCTA's CI against true value of $h^2$. The graph displays two lines, one for 95% CI and another for 90% CI, with coverage probability ranging from 80% to 100%.](image-url)
CIs are inaccurate in real datasets.
CIs are inaccurate in real datasets.
ALBI - Method overview

Motivated by these inaccuracies, we developed a novel method, Accurate LMM-Based heritability Bootstrap confidence Intervals (ALBI):

1. Rapidly compute the distribution of $\hat{h}^2$, over a grid of $h^2$ values (fast approximate parametric bootstrap)
2. Given the true distributions of $\hat{h}^2$, define accurate CIs

Advantages:
▶ Accurate CIs
▶ No assumption of normality, asymptotics, etc.
▶ Very fast (analytic and computational improvements)
ALBI - Method overview

Motivated by these inaccuracies, we developed a novel method, Accurate LMM-Based heritability Bootstrap confidence Intervals (ALBI):

1. Rapidly compute the distribution of $\hat{h}^2$, over a grid of $h^2$ values (fast approximate parametric bootstrap)
Motivated by these inaccuracies, we developed a novel method, Accurate LMM-Based heritability Bootstrap confidence Intervals (ALBI):

1. Rapidly compute the distribution of $\hat{h}^2$, over a grid of $h^2$ values (fast approximate parametric bootstrap)
2. Given the true distributions of $\hat{h}^2$, define accurate CIs
Motivated by these inaccuracies, we developed a novel method, **Accurate LMM-Based heritability Bootstrap confidence Intervals (ALBI)**:

1. Rapidly compute the distribution of $\hat{h}^2$, over a grid of $h^2$ values (fast approximate parametric bootstrap)
2. Given the true distributions of $\hat{h}^2$, define accurate CIs

Advantages:

- Accurate CIs
- No assumption of normality, asymptotics, etc.
- Very fast (analytic and computational improvements)
Fast parametric bootstrap

- For a fixed $h^2$ (and $\sigma_p^2 = 1, \beta = 0_p$), we can estimate the distribution of $\hat{h}^2$ with a parametric bootstrap method.
Fast parametric bootstrap

- For a fixed $h^2$ (and $\sigma^2_p = 1, \beta = 0_p$), we can estimate the distribution of $\hat{h}^2$ with a parametric bootstrap method.
- We estimate the probability mass function of a random variable taking values in the set
  \[ \{0, (0, 0.01], (0.01, 0.02], \ldots, (0.99, 1), 1\} \]
Fast parametric bootstrap

- For a fixed $h^2$ (and $\sigma_p^2 = 1$, $\beta = 0_p$), we can estimate the distribution of $\hat{h}^2$ with a parametric bootstrap method.
- We estimate the probability mass function of a random variable taking values in the set \{0, (0, 0.01], (0.01, 0.02], \ldots, (0.99, 1), 1\}.
- Steps:
  1. **Random sampling**: Draw $N$ (e.g., 10,000) phenotype vectors $\mathbf{y}_1^*, \ldots, \mathbf{y}_N^*$ from $\mathcal{N}(\mathbf{0}_n, h^2 \mathbf{K} + (1 - h^2) \mathbf{I}_n)$.
Fast parametric bootstrap

- For a fixed $h^2$ (and $\sigma_p^2 = 1, \beta = 0_p$), we can estimate the distribution of $\hat{h}^2$ with a parametric bootstrap method.
- We estimate the probability mass function of a random variable taking values in the set
  \[ \{0, (0, 0.01], (0.01, 0.02], \ldots, (0.99, 1), 1\} \]
- Steps:
  1. **Random sampling**: Draw $N$ (e.g., 10,000) phenotype vectors $y_1^*, \ldots, y_N^*$ from $\mathcal{N}(0_n, h^2K + (1 - h^2)I_n)$
  2. **REML estimation**: Calculate the REML estimates $\hat{h}^2(y_1^*), \ldots, \hat{h}^2(y_N^*)$ for each of these phenotype vectors
Fast parametric bootstrap

- For a fixed $h^2$ (and $\sigma^2_p = 1, \beta = 0_p$), we can estimate the distribution of $\hat{h}^2$ with a parametric bootstrap method.

- We estimate the probability mass function of a random variable taking values in the set
  \[
  \{0, (0, 0.01], (0.01, 0.02], \ldots, (0.99, 1), 1\}
  \]

- Steps:
  1. **Random sampling**: Draw $N$ (e.g., 10,000) phenotype vectors $y^*_1, \ldots, y^*_N$ from $\mathcal{N}(0_n, h^2 K + (1 - h^2)I_n)$
  2. **REML estimation**: Calculate the REML estimates $\hat{h}^2(y^*_1), \ldots, \hat{h}^2(y^*_N)$ for each of these phenotype vectors
  3. **Density estimation**: Count the proportion of estimates $\hat{h}^2(y^*_i)$ that fall in each bin or on the boundary 0 or 1. Use these fractions as an estimate of the density of $\hat{h}^2$ for this value of $h^2$
Fast parametric bootstrap: 1. Random sampling

- Drawing a vector \( y \sim \mathcal{N}(0_n, V_{h^2}) \) may be done by drawing a vector \( \tilde{y} \sim \mathcal{N}(0, I_n) \), and calculating \( y = V_{h^2}^{1/2} \tilde{y} \).
Fast parametric bootstrap: 1. Random sampling

- Drawing a vector $\mathbf{y} \sim \mathcal{N}(0_n, \mathbf{V}_{h^2})$ may be done by drawing a vector $\tilde{\mathbf{y}} \sim \mathcal{N}(0, \mathbf{I}_n)$, and calculating $\mathbf{y} = \mathbf{V}^{1/2} \tilde{\mathbf{y}}$.
- Any statement about $\mathbf{y}$ can then be restated in terms of $\tilde{\mathbf{y}}$, or further in terms of a vector $\mathbf{u} = \mathbf{U}^T \tilde{\mathbf{y}} = \mathbf{U}^T \mathbf{V}_{h^2}^{-1/2} \mathbf{y}$, where $\mathbf{K} = \mathbf{U} \mathbf{D} \mathbf{U}^T$.
Fast parametric bootstrap: 1. Random sampling

- Drawing a vector $\mathbf{y} \sim \mathcal{N}(\mathbf{0}_n, \mathbf{V}_{h^2})$ may be done by drawing a vector $\tilde{\mathbf{y}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, and calculating $\mathbf{y} = \mathbf{V}_{h^2}^{1/2} \tilde{\mathbf{y}}$.

- Any statement about $\mathbf{y}$ can then be restated in terms of $\tilde{\mathbf{y}}$, or further in terms of a vector $\mathbf{u} = \mathbf{U}^T \tilde{\mathbf{y}} = \mathbf{U}^T \mathbf{V}_{h^2}^{-1/2} \mathbf{y}$, where $K = \mathbf{U} \mathbf{D} \mathbf{U}^T$.

- Since $\mathbf{U}$ is orthonormal, $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$. 
Fast parametric bootstrap: 1. Random sampling

- Drawing a vector $\mathbf{y} \sim \mathcal{N}(\mathbf{0}_n, \mathbf{V}_{h^2})$ may be done by drawing a vector $\tilde{\mathbf{y}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, and calculating $\mathbf{y} = \mathbf{V}_{h^2}^{1/2} \tilde{\mathbf{y}}$. 
- Any statement about $\mathbf{y}$ can then be restated in terms of $\tilde{\mathbf{y}}$, or further in terms of a vector $\mathbf{u} = \mathbf{U}^T \tilde{\mathbf{y}} = \mathbf{U}^T \mathbf{V}_{h^2}^{-1/2} \mathbf{y}$, where $\mathbf{K} = \mathbf{UDU}^T$.
- Since $\mathbf{U}$ is orthonormal, $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$.
- Therefore, instead of drawing multiple phenotype vectors $\mathbf{y}_1^*, \ldots, \mathbf{y}_N^*$, we draw $\mathbf{u}_1, \ldots, \mathbf{u}_N \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$, and rephrase later stages in terms of these $\mathbf{u}$-s.
Fast parametric bootstrap: 1. Random sampling

- Drawing a vector \( y \sim \mathcal{N}(0_n, V_{h^2}) \) may be done by drawing a vector \( \tilde{y} \sim \mathcal{N}(0, I_n) \), and calculating \( y = V_{h^2}^{1/2} \tilde{y} \).

- Any statement about \( y \) can then be restated in terms of \( \tilde{y} \), or further in terms of a vector \( u = U^T \tilde{y} = U^T V_{h^2}^{-1/2} y \), where \( K = UDU^T \).

- Since \( U \) is orthonormal, \( u \sim \mathcal{N}(0, I_n) \).

- Therefore, instead of drawing multiple phenotype vectors \( y_1^*, \ldots, y_N^* \), we draw \( u_1, \ldots, u_N \sim \mathcal{N}(0, I_n) \), and rephrase later stages in terms of these \( u \)-s.

- Using \( u \) instead of \( \tilde{y} \) simplifies further calculations, and additionally avoids expensive matrix multiplications.
Fast parametric bootstrap: 2. REML estimation

- Instead of finding the global maximum of $\ell_{REML}$ directly:

  - Search for local maxima instead of the global maximum
  - Use the derivative $\frac{\partial \ell_{REML}}{\partial h^2}$ instead of $\ell_{REML}$ itself

- Multiplicity happens only rarely, and even an arbitrary decision between local maxima does not noticeably hurt CI accuracy.

- To check if $\hat{h}^2 = H^2$ is a maximum, just check $\frac{\partial \ell_{REML}}{\partial h^2}(H^2) = 0$

- To check if $\hat{h}^2 = 0$ (resp., $1$) is a maximum, just check $\frac{\partial \ell_{REML}}{\partial h^2}(0) \leq 0$ (resp., $\frac{\partial \ell_{REML}}{\partial h^2}(1) \geq 0$)

- We are more interested in the question of the estimate $\hat{h}^2$ being inside an interval: If $\frac{\partial \ell_{REML}}{\partial h^2}(c_1) > 0$ and $\frac{\partial \ell_{REML}}{\partial h^2}(c_2) < 0$, then there exists at least one $\hat{h}^2 \in (c_1, c_2)$ which is a local maximum (necessary condition but sufficient if grid is fine).
Fast parametric bootstrap: 2. REML estimation

- Instead of finding the global maximum of $\ell_{REML}$ directly:
  1. Search for local maxima instead of the global maximum

- Multiplicity happens only rarely, and even an arbitrary decision between local maxima does not noticeably hurt CI accuracy.

- To check if $\hat{h}^2 = H^2$ is a maximum, just check $\frac{\partial \ell_{REML}}{\partial h^2}(H^2) = 0$

- To check if $\hat{h}^2 = 0$ (resp., 1) is a maximum, just check $\frac{\partial \ell_{REML}}{\partial h^2}(0) \leq 0$ (resp., $\frac{\partial \ell_{REML}}{\partial h^2}(1) \geq 0$)

- We are more interested in the question of the estimate $\hat{h}^2$ being inside an interval: If $\frac{\partial \ell_{REML}}{\partial h^2}(c_1) > 0$ and $\frac{\partial \ell_{REML}}{\partial h^2}(c_2) < 0$, then there exists at least one $\hat{h}^2 \in (c_1, c_2)$ which is a local maximum (necessary condition but sufficient if grid is fine).
Fast parametric bootstrap: 2. REML estimation

- Instead of finding the global maximum of $\ell_{REML}$ directly:
  1. Search for local maxima instead of the global maximum
  2. Use the derivative $\frac{\partial \ell_{REML}}{\partial h^2}$ instead of $\ell_{REML}$ itself

Multiplicity happens only rarely, and even an arbitrary decision between local maxima does not noticeably hurt CI accuracy.

To check if $\hat{h}^2 = H^2$ is a maximum, just check $\frac{\partial \ell_{REML}}{\partial h^2}(H^2) = 0$

To check if $\hat{h}^2 = 0$ (resp., $1$) is a maximum, just check $\frac{\partial \ell_{REML}}{\partial h^2}(0) \leq 0$ (resp., $\frac{\partial \ell_{REML}}{\partial h^2}(1) \geq 0$)

We are more interested in the question of the estimate $\hat{h}^2$ being inside an interval: If $\frac{\partial \ell_{REML}}{\partial h^2}(c_1) > 0$ and $\frac{\partial \ell_{REML}}{\partial h^2}(c_2) < 0$, then there exists at least one $\hat{h}^2 \in (c_1, c_2)$ which is a local maximum (necessary condition but sufficient if grid is fine).
Fast parametric bootstrap: 2. REML estimation

- Instead of finding the global maximum of $\ell_{REML}$ directly:
  1. Search for local maxima instead of the global maximum
  2. Use the derivative $\frac{\partial \ell_{REML}}{\partial h^2}$ instead of $\ell_{REML}$ itself

- Multiplicity happens only rarely, and even an arbitrary decision between local maxima does not noticably hurt CI accuracy.
Fast parametric bootstrap: 2. REML estimation

- Instead of finding the global maximum of $\ell_{REML}$ directly:
  1. Search for local maxima instead of the global maximum
  2. Use the derivative $\frac{\partial \ell_{REML}}{\partial h^2}$ instead of $\ell_{REML}$ itself

- Multiplicity happens only rarely, and even an arbitrary decision between local maxima does not noticeably hurt CI accuracy.

- To check if $\hat{h}^2 = H^2$ is a maximum, just check $\frac{\partial \ell_{REML}}{\partial h^2}(H^2) = 0$
Fast parametric bootstrap: 2. REML estimation

- Instead of finding the global maximum of $\ell_{REML}$ directly:
  1. Search for local maxima instead of the global maximum
  2. Use the derivative $\frac{\partial \ell_{REML}}{\partial h^2}$ instead of $\ell_{REML}$ itself

- Multiplicity happens only rarely, and even an arbitrary decision between local maxima does not noticeably hurt CI accuracy.

- To check if $\hat{h}^2 = H^2$ is a maximum, just check $\frac{\partial \ell_{REML}}{\partial h^2}(H^2) = 0$

- To check if $\hat{h}^2 = 0$ (resp., 1) is a maximum, just check $\frac{\partial \ell_{REML}}{\partial h^2}(0) \leq 0$ (resp., $\frac{\partial \ell_{REML}}{\partial h^2}(1) \geq 0$)
Fast parametric bootstrap: 2. REML estimation

- Instead of finding the global maximum of $\ell_{REML}$ directly:
  1. Search for local maxima instead of the global maximum
  2. Use the derivative $\frac{\partial \ell_{REML}}{\partial h^2}$ instead of $\ell_{REML}$ itself

- Multiplicity happens only rarely, and even an arbitrary decision between local maxima does not noticeably hurt CI accuracy.

- To check if $\hat{h}^2 = H^2$ is a maximum, just check $\frac{\partial \ell_{REML}}{\partial h^2}(H^2) = 0$

- To check if $\hat{h}^2 = 0$ (resp., 1) is a maximum, just check $\frac{\partial \ell_{REML}}{\partial h^2}(0) \leq 0$ (resp., $\frac{\partial \ell_{REML}}{\partial h^2}(1) \geq 0$)

- We are more interested in the question of the estimate $\hat{h}^2$ being inside an interval: If $\frac{\partial \ell_{REML}}{\partial h^2}(c_1) > 0$ and $\frac{\partial \ell_{REML}}{\partial h^2}(c_2) < 0$, then there exists at least one $\hat{h}^2 \in (c_1, c_2)$ which is a local maximum (necessary condition but sufficient if grid is fine).
Fast parametric bootstrap: 2. REML estimation

- We are interested in evaluating \( \frac{\partial \ell_{REML}}{\partial h^2} (H^2; y) \) restated as a function of \( u = U^T V_{h^2}^{-1/2} y \).

- Fortunately, in several common scenarios, it is possible to calculate this derivative efficiently:
Proposition 1

Assume $X = 0$ and let $K = UDU^T$ be the eigen-decomposition of $K$, with the eigenvalues $d_1, \ldots, d_n$. Define

$$
\xi_{i, H^2} = \frac{h^2(d_i - 1) + 1}{H^2(d_i - 1) + 1} \left( \frac{d_i - 1}{H^2(d_i - 1) + 1} - \frac{1}{n} \sum_{j=1}^{n} \frac{d_j - 1}{H^2(d_j - 1) + 1} \right),
$$

for $i = 1, \ldots, n$. Then, a necessary condition for $\hat{h}^2 = H^2$, for $0 < H^2 < 1$, is

$$
\sum_{i=1}^{n} \xi_{i, H^2} u_i^2 = 0
$$

where $u = U^T V^{-1/2} y$, and it can be evaluated in time complexity $O(n)$ given $u$. 
Proof.

\( \hat{h}^2 = H^2 \) (when \( 0 < H^2 < 1 \) \( \Rightarrow \exists (\hat{h}^2, \hat{\sigma}_p^2) \), with \( \hat{h}^2 = H^2 \), for which

\[
\frac{\partial \ell_{REML}}{\partial h^2}(H^2, \hat{\sigma}_p^2) = 0
\]

\[
\frac{\partial \ell_{REML}}{\partial \sigma_p^2}(H^2, \hat{\sigma}_p^2) = 0
\]

\[
\ell_{REML}(\hat{h}^2, \hat{\sigma}_p^2) \propto -n \log \hat{\sigma}_p^2 - \frac{y^T V^{-1}_H y}{\hat{\sigma}_p^2} - \log |V_{H^2}|
\]

\[
\frac{\partial \ell_{REML}}{\partial h^2}(H^2, \hat{\sigma}_p^2) \propto \frac{1}{\hat{\sigma}_p^2} y^T V^{-1}_H (K - I) V^{-1}_H y - \frac{\partial}{\partial h^2}(\log |V_{h^2}|)|_{h^2=H^2} = 0.
\]

When \( X = 0 \), \( \hat{\sigma}_p^2(H^2) = \frac{1}{n} y^T V^{-1}_H y \). Substitute to get:
Fast parametric bootstrap: 2. REML estimation

Proof.

\[ \hat{h}^2 = H^2 \Rightarrow n \cdot \frac{y^T V_{H^2}^{-1}(K - I)V_{H^2}^{-1}y}{y^T V_{H^2}^{-1}y} - \frac{\partial}{\partial h^2} (\log |V_{h^2}|)\bigg|_{h^2=H^2} = 0. \]

\[ \frac{\partial}{\partial h^2} \log |V_{h^2}| = \frac{\partial}{\partial h^2} \log |U \cdot \text{diag} (h^2 d_i + (1 - h^2)) \cdot U^T| \]

\[ = \frac{\partial}{\partial h^2} \log |\text{diag} (h^2 d_i + (1 - h^2))| \]

\[ = \frac{\partial}{\partial h^2} \sum_{i=1}^{n} \log (h^2 d_i + (1 - h^2)) \]

\[ = \sum_{i=1}^{n} \frac{d_i - 1}{h^2(d_i - 1) + 1} \]
Fast parametric bootstrap: 2. REML estimation

Proof.

\[ \hat{h}^2 = H^2 \Rightarrow n \cdot \frac{y^T V_{H^2}^{-1} (K - I) V_{H^2}^{-1} y}{y^T V_{H^2}^{-1} y} - \sum_{i=1}^{n} \frac{d_i - 1}{H^2 (d_i - 1) + 1} = 0. \]

Substitute \( y = V_{h^2}^{1/2} U u \) to get

\[ \Rightarrow n \cdot \frac{u^T U^T V_{h^2}^{1/2} V_{H^2}^{-1} (K - I) V_{H^2}^{-1} V_{h^2}^{1/2} U u}{u^T U^T V_{h^2}^{1/2} V_{H^2}^{-1} V_{h^2}^{1/2} U u} - \sum_{i=1}^{n} \frac{d_i - 1}{H^2 (d_i - 1) + 1} = 0. \]

All of \( V_{h^2}^{1/2} \), \( V_{H^2}^{-1} \) and \( K \) are diagonalizeable by \( U \) (polynomials of \( K \)), so we may replace:
Fast parametric bootstrap: 2. REML estimation

Proof.

\[
U^T V_{h^2}^{1/2} (K - I) V_{H^2}^{-1} V_{h^2}^{1/2} U = \text{diag} \left( \frac{(h^2(d_i - 1) + 1)(d_i - 1)}{(H^2(d_i - 1) + 1)^2} \right)
\]

\[
U^T V_{h^2}^{1/2} V_{H^2}^{-1} V_{h^2}^{1/2} U = \text{diag} \left( \frac{h^2(d_i - 1) + 1}{H^2(d_i - 1) + 1} \right).
\]

Substitute back to get:

\[
\Rightarrow n \cdot \frac{\sum_{i=1}^{n} \frac{(h^2(d_i - 1) + 1)(d_i - 1)}{(H^2(d_i - 1) + 1)^2} \cdot u_i^2}{\sum_{i=1}^{n} \frac{h^2(d_i - 1) + 1}{H^2(d_i - 1) + 1} \cdot u_i^2} - \sum_{i=1}^{n} \frac{d_i - 1}{H^2(d_i - 1) + 1} = 0
\]
Proof.

Finally, rearrange to get:

\[ \hat{h}^2 = H^2 \]

\[ \Rightarrow \sum_{i=1}^{n} \frac{h^2(d_i - 1) + 1}{H^2(d_i - 1) + 1} \left( \frac{d_i - 1}{H^2(d_i - 1) + 1} - \frac{1}{n} \sum_{j=1}^{n} \frac{d_j - 1}{H^2(d_j - 1) + 1} \right) u_i^2 \]

\[ = \sum_{i=1}^{n} \xi_i h^2, H^2 u_i^2 = 0 \]
Proposition 2

Assume the columns of $\mathbf{X}$ are eigenvectors of $\mathbf{K}$, let $p_i = 0$ if the $i$-th eigenvector is in $\mathbf{X}$ and 1 otherwise. Define:

$$
\xi_{i, h^2, H^2} = p_i \cdot \frac{h^2 (d_i - 1) + 1}{H^2 (d_i - 1) + 1} \left( \frac{p_i (d_i - 1)}{H^2 (d_i - 1) + 1} - \frac{1}{n - p} \sum_{j=1}^{n} \frac{p_j (d_j - 1)}{H^2 (d_j - 1) + 1} \right)
$$

Then, a necessary condition for $\hat{h}^2 = H^2$, for $0 < H^2 < 1$, is

$$
\sum_{i=1}^{n} \xi_{i, h^2, H^2} u_i^2 = 0
$$

where $u = \mathbf{U}^T \mathbf{V}_{h^2}^{-1/2} \mathbf{y}$. It can be evaluated in time complexity $O(n)$ given $\mathbf{u}$. 

Fast parametric bootstrap: 2. REML estimation
Fast parametric bootstrap: 2. REML estimation

When are the columns of $X$ eigenvectors of $K$?

- $X = 0$
Fast parametric bootstrap: 2. REML estimation

When are the columns of $X$ eigenvectors of $K$?

- $X = 0$
- Intercept: $X = 1_n$
Fast parametric bootstrap: 2. REML estimation

When are the columns of $X$ eigenvectors of $K$?

- $X = 0$
- Intercept: $X = 1_n$
  - Proof: $Z$ is mean-centered, so $Z^T 1_n = 0_m$. Then, $K = \frac{1}{m} ZZ^T$, so $K \cdot 1_n = 0 \cdot 1_n$. 
Fast parametric bootstrap: 2. REML estimation

When are the columns of \( X \) eigenvectors of \( K \)?

- \( X = 0 \)
- Intercept: \( X = 1_n \)
  - Proof: \( Z \) is mean-centered, so \( Z^T 1_n = 0_m \). Then, \( K = \frac{1}{m} Z Z^T \), so \( K \cdot 1_n = 0 \cdot 1_n \).
- Eigenvectors of \( K \) used as covariates (population structure, batch effects...).
Proposition 3

Assume the columns of $X$ are eigenvectors of $K$. Define $u$ and $\xi_h^{h^2,H^2}$ as before. Then, a necessary condition for $\hat{h}^2 = H^2$ is

$$\sum_{i=1}^{n} \xi_i^{h^2,H^2} u_i^2 = 0 \quad \text{for} \ 0 < H^2 < 1$$

$$\sum_{i=1}^{n} \xi_i^{h^2,0} u_i^2 \leq 0 \quad \text{for} \ H^2 = 0$$

$$\sum_{i=1}^{n} \xi_i^{h^2,1} u_i^2 \geq 0 \quad \text{for} \ H^2 = 1,$$

and a necessity condition for $\hat{h}^2 \in [c_1, c_2]$ is

$$\sum_{i=1}^{n} \xi_i^{h^2,c_1} u_i^2 \geq 0 \quad \text{and} \quad \sum_{i=1}^{n} \xi_i^{h^2,c_2} u_i^2 \leq 0$$

They can be evaluated in time complexity $O(n)$ given $u$. 
Fast parametric bootstrap: 2. REML estimation

**Proposition 4**

Assume a general $n \times p$ covariate matrix $\mathbf{X}$. Then, a necessary condition for $\hat{h}^2 = H^2$ or for $\hat{h}^2 \in [c_1, c_2]$ can be evaluated in time complexity $O(np^2 + p^3)$ given $\mathbf{u}$ and $\mathbf{U}^\mathsf{T}\mathbf{X}$.

Extension of work by Crainiceanu and Ruppert (2004).
(Note: Similar proofs exist for ML.)
Fast parametric bootstrap: 3. Density estimation

To estimate the distribution of $\hat{h}^2$, given $h^2$ (covariates are eigenvectors, for simplicity):

1. Draw $u_1, \ldots, u_N \sim \mathcal{N}(0_n, I_n)$
Fast parametric bootstrap: 3. Density estimation

To estimate the distribution of $\hat{h}^2$, given $h^2$ (covariates are eigenvectors, for simplicity):

1. Draw $u_1, \ldots, u_N \sim \mathcal{N}(0_n, I_n)$

2. Calculate $\xi_{i, H^2}$ over a grid (e.g. $[0, 0.01, \ldots, 0.99, 1]$) of values for $H^2$ (complexity: $O(nG)$)
Fast parametric bootstrap: 3. Density estimation

To estimate the distribution of $\hat{h}^2$, given $h^2$ (covariates are eigenvectors, for simplicity):

1. Draw $u_1, \ldots, u_N \sim \mathcal{N}(0_n, I_n)$

2. Calculate $\xi_i^{h^2,H^2}$ over a grid (e.g. $[0, 0.01, \ldots, 0.99, 1]$) of values for $H^2$ (complexity: $O(nG)$)

3. Estimate $\Pr_{h^2}(\hat{h}^2 = 0)$ as the proportion of $u$ values for which $\sum_{i=1}^n \xi_i^{h^2,0} u_i^2 \leq 0$ (complexity: $O(nN)$)
Fast parametric bootstrap: 3. Density estimation

To estimate the distribution of $\hat{h}^2$, given $h^2$ (covariates are eigenvectors, for simplicity):

1. Draw $u_1, \ldots, u_N \sim \mathcal{N}(0_n, I_n)$

2. Calculate $\xi_{i,h^2,H^2}$ over a grid (e.g. $[0, 0.01, \ldots, 0.99, 1]$) of values for $H^2$ (complexity: $O(nG)$)

3. Estimate $Pr_{h^2}(\hat{h}^2 = 0)$ as the proportion of $u$ values for which
   \[ \sum_{i=1}^{n} \xi_{i,h^2,0} u_i^2 \leq 0 \] (complexity: $O(nN)$)

4. Estimate $Pr_{h^2}(\hat{h}^2 = 1)$ as the proportion of $u$ values for which
   \[ \sum_{i=1}^{n} \xi_{i,h^2,1} u_i^2 \geq 0 \] (complexity: $O(nN)$)
Fast parametric bootstrap: 3. Density estimation

To estimate the distribution of \( \hat{h}^2 \), given \( h^2 \) (covariates are eigenvectors, for simplicity):

1. Draw \( u_1, \ldots, u_N \sim \mathcal{N}(0_n, I_n) \)
2. Calculate \( \xi_i^{h^2, H^2} \) over a grid (e.g. \([0, 0.01, \ldots, 0.99, 1]\)) of values for \( H^2 \) (complexity: \( O(nG) \))
3. Estimate \( \Pr_{h^2}(\hat{h}^2 = 0) \) as the proportion of \( u \) values for which \( \sum_{i=1}^n \xi_i^{h^2, 0} u_i^2 \leq 0 \) (complexity: \( O(nN) \))
4. Estimate \( \Pr_{h^2}(\hat{h}^2 = 1) \) as the proportion of \( u \) values for which \( \sum_{i=1}^n \xi_i^{h^2, 1} u_i^2 \geq 0 \) (complexity: \( O(nN) \))
5. Estimate \( \Pr_{h^2}(\hat{h}^2 \in [c_1, c_2]) \) as the proportion of \( u \) values for which \( \sum_{i=1}^n \xi_i^{h^2, c_1} u_i^2 \geq 0 \) and \( \sum_{i=1}^n \xi_i^{h^2, c_2} u_i^2 \leq 0 \) (complexity: \( O(nN) \))
Fast parametric bootstrap: 3. Density estimation

- Total complexity: $O(nG + nN) = O(n)$ (vs. the naive $O(n^3)$).
- Note that for every $u_i$, one of the above events must hold.
- In order to estimate the probability of a boundary estimate (e.g., $\hat{h}^2 = 0$) or an interval estimate (e.g., $0 < \hat{h}^2 < 0.01$), we need not find the maximum for each $u_i$. 
ALBI: Benchmarks

- Parametric bootstrap for computing all $\hat{h}^2$ distributions:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>GCTA</th>
<th>pylmm</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTEx (185 individuals)</td>
<td>7.8 hours</td>
<td>1.05 hours</td>
</tr>
<tr>
<td>LURIC (867 individuals)</td>
<td>3.79 days</td>
<td>1.5 hours</td>
</tr>
<tr>
<td>NFBC (2520 individuals)</td>
<td>&gt; 30 days</td>
<td>3.8 hours</td>
</tr>
</tbody>
</table>

Distributions of $\hat{h}^2$ for $h^2 = 0, 0.01, \ldots, 1$, with 1000 random bootstrap samples
ALBI: Benchmarks

- Parametric bootstrap for computing all $\hat{h}^2$ distributions:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>GCTA</th>
<th>pylmm</th>
<th>ALBI</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTEx (185 individuals)</td>
<td>7.8 hours</td>
<td>1.05 hours</td>
<td>27 sec</td>
</tr>
<tr>
<td>LURIC (867 individuals)</td>
<td>3.79 days</td>
<td>1.5 hours</td>
<td>1.2 min</td>
</tr>
<tr>
<td>NFBC (2520 individuals)</td>
<td>&gt; 30 days</td>
<td>3.8 hours</td>
<td>5.35 min</td>
</tr>
</tbody>
</table>

Distributions of $\hat{h}^2$ for $h^2 = 0, 0.01, \ldots, 1$, with 1000 random bootstrap samples
ALBI: Benchmarks

- Parametric bootstrap for computing all $\hat{h}^2$ distributions:

<table>
<thead>
<tr>
<th>Dataset</th>
<th>GCTA</th>
<th>pylmm</th>
<th>ALBI</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTEx (185 individuals)</td>
<td>7.8 hours</td>
<td>1.05 hours</td>
<td>27 sec</td>
</tr>
<tr>
<td>LURIC (867 individuals)</td>
<td>3.79 days</td>
<td>1.5 hours</td>
<td>1.2 min</td>
</tr>
<tr>
<td>NFBC (2520 individuals)</td>
<td>&gt; 30 days</td>
<td>3.8 hours</td>
<td>5.35 min</td>
</tr>
</tbody>
</table>

Distributions of $\hat{h}^2$ for $h^2 = 0, 0.01, \ldots, 1$, with 1000 random bootstrap samples

- Performed only once, then building a CI is immediate
ALBI: Constructing accurate CIs

- Utilizes the duality between hypothesis testing and confidence intervals
ALBI: Constructing accurate CIs

- Utilizes the duality between hypothesis testing and confidence intervals
- Given the distribution of $\hat{h}^2$ given the true heritability value $h^2$, for all values of $h^2$,
ALBI: Constructing accurate CIs

- Utilizes the duality between hypothesis testing and confidence intervals
- Given the distribution of $\hat{h}^2$ given the true heritability value $h^2$, for all values of $h^2$,
- For each $h^2$, define an acceptance region $A_{h^2}$ of $\hat{h}^2$ values which occur in probability $1 - \alpha$ (e.g. 95%), where $h^2 \in A_{h^2}$
ALBI: Constructing accurate CIs

- Utilizes the duality between hypothesis testing and confidence intervals
- Given the distribution of $\hat{h}^2$ given the true heritability value $h^2$, for all values of $h^2$,
- For each $h^2$, define an acceptance region $A_{h^2}$ of $\hat{h}^2$ values which occur in probability $1 - \alpha$ (e.g. 95%), where $h^2 \in A_{h^2}$
- The CI for $\hat{h}^2$ is the interval of $h^2$ values for which $\hat{h}^2 \in A_{h^2}$
ALBI: Constructing accurate CIs

- Utilizes the duality between hypothesis testing and confidence intervals
- Given the distribution of $\hat{h}^2$ given the true heritability value $h^2$, for all values of $h^2$,
- For each $h^2$, define an acceptance region $A_{h^2}$ of $\hat{h}^2$ values which occur in probability $1 - \alpha$ (e.g. 95%), where $h^2 \in A_{h^2}$
- The CI for $\hat{h}^2$ is the interval of $h^2$ values for which $\hat{h}^2 \in A_{h^2}$
- Then, $\Pr_{h^2}(h^2 \in C_{\hat{h}^2}) = \Pr_{h^2}(\hat{h}^2 \in A_{h^2}) = 1 - \alpha$. 
ALBI: Constructing accurate CIs

- Utilizes the duality between hypothesis testing and confidence intervals
- Given the distribution of $\hat{h}^2$ given the true heritability value $h^2$, for all values of $h^2$,
- For each $h^2$, define an acceptance region $A_{h^2}$ of $\hat{h}^2$ values which occur in probability $1 - \alpha$ (e.g. 95%), where $h^2 \in A_{h^2}$
- The CI for $\hat{h}^2$ is the interval of $h^2$ values for which $\hat{h}^2 \in A_{h^2}$
- Then, $\text{Pr}_{h^2}(h^2 \in C_{\hat{h}^2}) = \text{Pr}_{h^2}(\hat{h}^2 \in A_{h^2}) = 1 - \alpha$.
- Coverage is $\inf_{h^2} \text{Pr}_{h^2}(h^2 \in C_{\hat{h}^2}) = 1 - \alpha$
CI construction - example

$h^2 = 0.1$

$\hat{h}^2$

$A_{0.1} = [0, 0.3]$
CIV construction - example

\[ h^2 = 0.2 \]

\[ A_{0.2} = [0, 0.4] \]
CI construction - example

\[ h^2 = 0.3 \]

\[ A_{0.3} = [0.1, 0.5] \]
CI construction - example

Requirements from choice of regions:
1. For accurate CIs, $P(h^2) = 1 - \alpha$.
2. $h^2 \in A_{h^2}$.
3. Lower and upper bounds of $A_{h^2}$ must be monotone functions of $h^2$. 

\[ C_{0.4} = [0.2, 0.6] \]
CI construction - example

Requirements from choice of regions:

1. For accurate CIs, \( P_h^2(A_h^2) = 1 - \alpha \).

2. Lower and upper bounds of \( A_h^2 \) must be monotone functions of \( h^2 \).

3. \( C_{0.4} = [0.2, 0.6] \)
Cl construction - example

Requirements from choice of regions:

1. For accurate CIs, \( \Pr_{h^2}(A_{h^2}) = 1 - \alpha \)
2. \( h^2 \in A_{h^2} \)
3. Lower and upper bounds of \( A_{h^2} \) must be monotone functions of \( h^2 \)
Choosing acceptance regions 1

Define:

\[ p_0(h^2) = \Pr_{h^2}(\hat{h}^2 = 0) \]
\[ p_1(h^2) = \Pr_{h^2}(\hat{h}^2 = 1) \]

The quantile function \( c_\beta \) of \( \hat{h}^2 \) is

\[ c_\beta(h^2) = \begin{cases} 
0 & \text{if } \beta \in [0, p_0(h^2)] \\
F_{h^2}^{-1}(\beta) & \text{if } \beta \in (p_0(h^2), 1 - p_1(h^2)) \\
1 & \text{if } \beta \in [1 - p_1(h^2), 1]
\end{cases} \]
Choosing acceptance regions II

Estimator distributions are discontinuous at the boundaries, so possibly $c_\beta(h^2)$ does not obey $\Pr_{h^2}([0, c_\beta(h^2)]) = \beta$:

$$\Pr_{h^2}([0, c_\beta(h^2)]) = \begin{cases} p_0(h^2) & \text{if } \beta \in [0, p_0(h^2)] \\ \beta & \text{if } \beta \in (p_0(h^2), 1 - p_1(h^2)) \\ 1 & \text{if } \beta \in [1 - p_1(h^2), 1]. \end{cases}$$
Choosing acceptance regions - illustration

We will use three types of regions:

\[ A^1_{h^2} = [0, c_{1-\alpha}(h^2)] \]
\[ A^2_{h^2} = [c_{\alpha/2}(h^2), c_{1-\alpha/2}(h^2)] \]
\[ A^3_{h^2} = [c_{\alpha}(h^2), 1] \]

Defined so that \( \Pr_{h^2}(A^i_{h^2}) \geq 1 - \alpha \) (possibly \( > \)).
Choosing acceptance regions: Case 1

- We start with $A^1$-s, and end with $A^3$, but when do we use $A^2$-s?
Choosing acceptance regions: Case 1

- We start with $A_1^1$-s, and end with $A_3^3$, but when do we use $A_2^2$-s?
- For which $h^2$, 
  \[ \Pr_{h^2}(A_{h^2}^2) = \Pr_{h^2}([c_{\alpha/2}(h^2), c_{1-\alpha/2}(h^2)]) = 1 - \alpha? \]
Choosing acceptance regions: Case 1

- We start with $A^1$-s, and end with $A^3$, but when do we use $A^2$-s?
- For which $h^2$,
  \[ \Pr_{h^2}(A^2_{h^2}) = \Pr_{h^2}([c_{\alpha/2}(h^2), c_{1-\alpha/2}(h^2)]) = 1 - \alpha? \]
- We need a small enough left boundary: $p_0(h^2) < \alpha/2$
Choosing acceptance regions: Case 1

- We start with $A^1$-s, and end with $A^3$, but when do we use $A^2$-s?
- For which $h^2$, 
  \[ \Pr_{h^2}(A_{h^2}^2) = \Pr_{h^2}([c_{\alpha/2}(h^2), c_{1-\alpha/2}(h^2)]) = 1 - \alpha? \]
- We need a small enough left boundary: $p_0(h^2) < \alpha/2$
- We need a small enough right boundary: $p_1(h^2) < \alpha/2$
Choosing acceptance regions: Case 1

- We start with $A^1$-s, and end with $A^3$, but when do we use $A^2$-s?
- For which $h^2$,
  \[ \Pr_{h^2}(A^2_{h^2}) = \Pr_{h^2}([c_{\alpha/2}(h^2), c_{1-\alpha/2}(h^2)]) = 1 - \alpha? \]
- We need a small enough left boundary: $p_0(h^2) < \alpha/2$
- We need a small enough right boundary: $p_1(h^2) < \alpha/2$
- \( \Rightarrow p_0^{-1}(\alpha/2) \leq h^2 \leq p_1^{-1}(\alpha/2). \)
Choosing acceptance regions: Case 1

- We start with $A^1$-s, and end with $A^3$, but when do we use $A^2$-s?
- For which $h^2$,
  \[ \Pr_{h^2}(A_{h^2}^2) = \Pr_{h^2}([c_{\alpha/2}(h^2), c_{1-\alpha/2}(h^2)]) = 1 - \alpha? \]
- We need a small enough left boundary: \( p_0(h^2) < \alpha/2 \)
- We need a small enough right boundary: \( p_1(h^2) < \alpha/2 \)
- \( \Rightarrow p_0^{-1}(\alpha/2) \leq h^2 \leq p_1^{-1}(\alpha/2). \)
- So if \( p_0^{-1}(\alpha/2) \leq p_1^{-1}(\alpha/2) \)...
Choosing acceptance regions: Case 1

If \( p_0^{-1}(\alpha/2) \leq p_1^{-1}(\alpha/2) \), define

\[
A_{h^2} = \begin{cases} 
A_{h^2}^1 = [0, c_{1-\alpha}(h^2)] & \text{if } h^2 \in [0, p_0^{-1}(\alpha/2)) \\
A_{h^2}^2 = [c_{\alpha/2}(h^2), c_{1-\alpha/2}(h^2)] & \text{if } h^2 \in [p_0^{-1}(\alpha/2), p_1^{-1}(\alpha/2)] \\
A_{h^2}^3 = [c_{\alpha}(h^2), 1] & \text{if } h^2 \in (p_1^{-1}(\alpha/2), 1] 
\end{cases}
\]
Choosing acceptance regions: Case 2

- If no $A^2$ achieve accuracy (because $p_0^{−1}(\alpha/2) > p_1^{−1}(\alpha/2)$)...
Choosing acceptance regions: Case 2

- If no $A^2$ achieve accuracy (because $p_{0}^{-1}(\alpha/2) > p_{1}^{-1}(\alpha/2))$...
- For which $h^2$, $\Pr_{h^2}(A_{h^2}^1) = \Pr_{h^2}([0, c_{1-\alpha}(h^2)]) = 1 - \alpha$?
Choosing acceptance regions: Case 2

- If no $A^2$ achieve accuracy (because $p_0^{-1}(\alpha/2) > p_1^{-1}(\alpha/2)$)...
- For which $h^2$, $\Pr_{h^2}(A_{h^2}^1) = \Pr_{h^2}([0, c_{1-\alpha}(h^2)]) = 1 - \alpha$?
- We need a small enough right boundary: $p_1(h^2) < \alpha$
Choosing acceptance regions: Case 2

- If no $A^2$ achieve accuracy (because $p_0^{-1}(\alpha/2) > p_1^{-1}(\alpha/2)$)...
- For which $h^2$, $\Pr h^2(A_{h^2}^1) = \Pr h^2([0, c_{1-\alpha}(h^2)]) = 1 - \alpha$?
- We need a small enough right boundary: $p_1(h^2) < \alpha$
- Similarly,
  $\Pr h^2(A_{h^2}^3) = \Pr h^2([c_\alpha(h^2), 1]) = 1 - \alpha \iff p_0(h^2) < \alpha$. 
Choosing acceptance regions: Case 2

- If no $A^2$ achieve accuracy (because $p_0^{-1}(\alpha/2) > p_1^{-1}(\alpha/2))$
- For which $h^2$, $\Pr_{h^2}(A^1_{h^2}) = \Pr_{h^2}([0, c_{1-\alpha}(h^2)]) = 1 - \alpha$?
- We need a small enough right boundary: $p_1(h^2) < \alpha$
- Similarly,
  
  $\Pr_{h^2}(A^3_{h^2}) = \Pr_{h^2}([c_{\alpha}(h^2), 1]) = 1 - \alpha \iff p_0(h^2) < \alpha$.
- Either $A^1_{h^2}$ or $A^3_{h^2}$ cover exactly $1 - \alpha$ if
  
  $p_0^{-1}(\alpha) \leq h^2 \leq p_1^{-1}(\alpha)$. 
Choosing acceptance regions: Case 2

- If no $A^2$ achieve accuracy (because $p_0^{-1}(\alpha/2) > p_1^{-1}(\alpha/2)$)...
- For which $h^2$, $\Pr_{h^2}(A^1_{h^2}) = \Pr_{h^2}([0, c_{1-\alpha}(h^2)]) = 1 - \alpha$?
- We need a small enough right boundary: $p_1(h^2) < \alpha$
- Similarly,
  \[ \Pr_{h^2}(A^3_{h^2}) = \Pr_{h^2}([c_{\alpha}(h^2), 1]) = 1 - \alpha \iff p_0(h^2) < \alpha. \]
- Either $A^1_{h^2}$ or $A^3_{h^2}$ cover exactly $1 - \alpha$ if $p_0^{-1}(\alpha) \leq h^2 \leq p_1^{-1}(\alpha)$.
- So if $p_0^{-1}(\alpha) \leq p_1^{-1}(\alpha)$,
Choosing acceptance regions: Case 2

If $p_0^{-1}(\alpha/2) > p_1^{-1}(\alpha/2)$ but $p_0^{-1}(\alpha) \leq p_1^{-1}(\alpha)$, define

$$
\delta = (p_0^{-1}(\alpha) + p_1^{-1}(\alpha))/2
$$

as the transition point, i.e.,

$$
A_{h^2} = \begin{cases} 
A_{h^2}^1 = [0, c_{1-\alpha}(h^2)] & \text{if } h^2 \in [0, \delta) \\
A_{h^2}^3 = [c_{\alpha}(h^2), 1] & \text{if } h^2 \in [\delta, 1] 
\end{cases}
$$

The graph shows three functions:

- $c_{\alpha/2}(h^2)$
- $c_{\alpha}(h^2)$
- $c_{1-\alpha}(h^2)$
- $c_{1-\alpha/2}(h^2)$

The range of $h^2$ is from 0 to 1, and the range of the functions is from 0 to 1.
Choosing acceptance regions - Case 3

- If $p_0^{-1}(\alpha) > p_1^{-1}(\alpha)$, then for all $h^2$ in this range, there is not an acceptance region with probability $1 - \alpha$ (using the statistic $\hat{h}^2$)
Choosing acceptance regions - Case 3

- If $p_0^{-1}(\alpha) > p_1^{-1}(\alpha)$, then for all $h^2$ in this range, there is not acceptance region with probability $1 - \alpha$ (using the statistic $\hat{h^2}$)

- Randomized confidence intervals: Make the upper bound of $C_0$ and lower bound of $C_1$ random variables with desired properties.
ALBI constructs accurate CIs

NFBC, GCTA’s CI

NFBC, ALBI’s CI

Coverage Probability

True value of $h^2$
ALBI constructs accurate CIs

![Graph showing comparison between GTEx, GCTA's CI and GTEx, ALBI's CI. The x-axis represents the true value of $h^2$, and the y-axis represents the coverage probability. The graphs show the performance of the CIs at different true values of $h^2$. The GTEx, GCTA's CI shows a decrease in coverage probability as $h^2$ increases, while the GTEx, ALBI's CI maintains a higher coverage probability across the range of $h^2$.](graph.png)
ALBI constructs accurate CIs

LURIC, GCTA’s CI

LURIC, ALBI’s CI

Coverage Probability

True value of $h^2$
Comparison of CI width. The ratio between the mean width of CIs derived from GCTA, and the width of ALBI’s CIs, as a function of the true $h^2$, for the studied datasets.
## ALBI vs. GCTA

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Heritability</th>
<th>ALBI 95% CI</th>
<th>GCTA 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG00000223972</td>
<td>0.0</td>
<td>[0, 0.65]</td>
<td>[-0.28, 0.28]</td>
</tr>
<tr>
<td>ENSG00000238009</td>
<td>0.104</td>
<td>[0, 0.8]</td>
<td>[-0.45, 0.66]</td>
</tr>
<tr>
<td>ENSG00000222623</td>
<td>0.364</td>
<td>[0, 1]</td>
<td>[-0.33, 1.06]</td>
</tr>
<tr>
<td>ENSG00000227232</td>
<td>0.463</td>
<td>[0.01, 1]</td>
<td>[-0.31, 1.23]</td>
</tr>
<tr>
<td>ENSG00000223663</td>
<td><strong>0.738</strong></td>
<td>[0.18, 1]</td>
<td><strong>[-0.08, 1.55]</strong></td>
</tr>
<tr>
<td>ENSG00000234619</td>
<td>0.873</td>
<td>[0.25, 1]</td>
<td>[0.04, 1.7]</td>
</tr>
<tr>
<td>ENSG00000240414</td>
<td>1.0</td>
<td>[0.41, 1]</td>
<td>[0.31, 1.69]</td>
</tr>
</tbody>
</table>

**Heritability estimates and CIs for phenotypes.** The estimated heritability values, along with the ALBI and current CIs, for a selection of GTEx gene expression profiles.
Summary

- Current methods for constructing confidence intervals are inaccurate

Inferences based on inaccurate CIs may lead to incorrect conclusions about the heritability of a trait.

ALBI - an efficient method for computing the distribution of the REML estimator of heritability and for constructing accurate confidence intervals.

ALBI is significantly faster than standard parametric bootstrap approaches in computing the true estimator distribution.
Summary

- Current methods for constructing confidence intervals are inaccurate
- Inferences based on inaccurate CIs may lead to incorrect conclusions about the heritability of a trait
Summary

- Current methods for constructing confidence intervals are inaccurate
- Inferences based on inaccurate CIs may lead to incorrect conclusions about the heritability of a trait
- ALBI - an efficient method for computing the distribution of the REML estimator of heritability and for constructing accurate confidence intervals
Summary

- Current methods for constructing confidence intervals are inaccurate
- Inferences based on inaccurate CIs may lead to incorrect conclusions about the heritability of a trait
- ALBI - an efficient method for computing the distribution of the REML estimator of heritability and for constructing accurate confidence intervals
- ALBI is significantly faster than standard parametric bootstrap approaches in computing the true estimator distribution
Future Directions

- Using results about quadratic forms of normal variables
Future Directions

- Using results about quadratic forms of normal variables
- Quantile regression
Future Directions

- Using results about quadratic forms of normal variables
- Quantile regression
- CIs for multiple variance components:

\[ y \sim \mathcal{N} \left( \mathbf{X}\beta, \sigma_{g_1}^2 \mathbf{K}_1 + \ldots + \sigma_{g_r}^2 \mathbf{K}_r + \sigma_e^2 \mathbf{I} \right), \]
Future Directions

- Using results about quadratic forms of normal variables
- Quantile regression
- CIs for multiple variance components:
  \[ y \sim \mathcal{N} \left( X\beta, \sigma^2_{g_1} K_1 + \ldots + \sigma^2_{g_r} K_r + \sigma^2_e I \right), \]
- Multi-trait model
Acknowledgements

- Collaborators: Shachar Kaufman, Reijo Laaksonen, Marcus Kleber, Winfried Marz, Eleazar Eskin, Saharon Rosset, Eran Halperin
- Thanks to: Yael Baran, Eli Levy-Karin, Noah Zaitlen
- Research funded by: The US-Israel Binational Science Foundation (BSF), National Science Foundation (NSF), The Edmond J. Safra Bioinformatics Center, the Colton Foundation and the Blavatnik Research Foundation

The original, ALBI the hamster:
The probability of estimating $\hat{h}^2 = 0$ or 1, for true underlying heritability values $h^2$.
Boundary probabilities

**Probability of boundary heritability estimates.** The probability of estimating $\hat{h}^2 = 0$, as a function of the true underlying heritability values $h^2$, for the studied datasets. The probability of $\hat{h}^2 = 0$ is high, especially for small values.
Bias

\[ \text{bias}(\hat{h}^2) \]

0 0.2 0.4 0.6 0.8 1
−0.2
−0.1
0
0.1
True value of \( h^2 \)

GTEx
LURIC
NFBC

True value of \( h^2 \)
The ratio between the mean standard error derived from GCTA and the empirical standard deviation of the REML estimator $\hat{h}^2$
The probability of $h^2 = 0$ being included in the CI, as a function of the true value of $h^2$, for the GTEx and LURIC datasets. These probabilities are shown for GCTA’s CI and ALBI’s CI, with a confidence level of 95%. It can be seen that CIs derived from the normal approximation tend to include $h^2 = 0$ more than necessary.