Rare-Variant Kernel Machine Test for Longitudinal Data for Population and Family Samples

Qi Yan

Department of Pediatrics, Children’s Hospital of Pittsburgh of UPMC
March 14th, 2016
Motivation

- **Phenotypes:**
  - In many genetic studies, phenotypes are measured at multiple time points for each subject. It is expected that a method that is able to take into account all time points jointly in an association test could improve the power;
  - Family based designs have been widely used. Appropriately handling familial correlation can retain Type I error rate;

- **Genotypes:**
  - MAF: Minor Allele Frequency
  - Common variants (MAF≥0.05): single marker test;
  - Rare variants (MAF<0.05): test at gene level (e.g. SKAT).
Motivation

- **Genotypes:**
  - Common variants (e.g. MAF≥0.05): single marker test;
  - Rare variants (e.g. MAF<0.05): test at gene level (e.g. SKAT).
Aims

• Association test between quantitative phenotypes and genes;

• Gene-based rare variants test;

• Multiple time points for each subject are tested simultaneously;

• Family structure is either (1) not considered or (2) considered;
Let there be $n$ subjects with $q$ genetic variants. The $n \times 1$ vector of the quantitative trait $y$ follows a linear mixed model:

$$y = X\beta + G\gamma + \epsilon$$

- $X$ is an $n \times p$ covariate matrix,
- $\beta$ is a $p \times 1$ vector containing parameters for the fixed effects (an intercept and $p - 1$ covariates),
- $G$ is an $n \times q$ genotype matrix for the $q$ rare genetic variants of interest,
- $\gamma$ is a $q \times 1$ vector for the random effects of the $q$ genetic variants,
- $\epsilon$ is an $n \times 1$ vector for the random error.

$$\gamma \sim N(0, \tau W)$$

$$\epsilon \sim N(0, \sigma_E^2 I)$$

where $W$ is a predefined $q \times q$ diagonal weight matrix for each variant.

Thus, the null hypothesis $H_0: \gamma = 0$ is equivalent to $H_0: \tau = 0$, which can be tested with a variance component score test in the mixed model.
**Methods**

**Sequence Kernel Association Test (SKAT):**

Q: What makes mixed model different from linear regression model?
A: random variables in addition to random error.

\[ y = X\beta + Gy + \varepsilon \quad \text{“linear mixed model”} \]

\[ \text{Var}(y) = \tau GWG' + \sigma_E^2 I \]

SKAT test statistic following a mixture of Chi-square distribution is:

\[ Q = (y - X\hat{\beta})' \hat{\Sigma}^{-1}GWG'\hat{\Sigma}^{-1}(y - X\hat{\beta}) \]

where the parameters are estimated under \( H_0 \) (i.e., \( H_0: \tau = 0 \))

Thus, under \( H_0: \quad y = X\beta + \varepsilon \quad \text{“linear regression model, no longer mixed model”} \]

\[ \hat{\Sigma} = \hat{\sigma}_E^2 I \]

\[ \hat{\beta} = (X'\hat{\Sigma}^{-1}X)^{-1}X'\hat{\Sigma}^{-1}y \]

- Called “kernel”.
- Linear combination used here. Could be more flexible form.

- The “full model” of SKAT is a linear mixed model
- The “null model” for the score test is a linear model
Methods

Kernel Machine (KM) Regression for Linear Mixed Model:

With additional random effects (besides the genetic effects):

Let there be $n$ subjects with $q$ genetic variants. The $n \times 1$ vector of the quantitative trait $y$ follows a linear mixed model:

$$ y = \mathbf{X}\beta + \mathbf{G}\gamma + \mathbf{u} + \mathbf{e} $$

- $\mathbf{X}$ is an $n \times p$ covariate matrix,
- $\beta$ is a $p \times 1$ vector containing parameters for the fixed effects (an intercept and $p - 1$ covariates),
- $\mathbf{G}$ is an $n \times q$ genotype matrix for the $q$ genetic variants of interest,
- $\gamma$ is a $q \times 1$ vector for the random effects of the $q$ genetic variants,
- $\mathbf{e}$ is an $n \times 1$ vector for the random error,
- $\mathbf{u}$ is an $n \times 1$ vector for the random effects due to covariates (e.g., time for longitudinal data or relatedness in families)
Methods

Kernel Machine (KM) Regression for Linear Mixed Model:

\[ y = X\beta + G\gamma + u + \varepsilon \]
\[ \gamma \sim N(0, \tau W) \]
\[ u \sim N(0, K) \]
\[ \varepsilon \sim N(0, \sigma^2_E I) \]

where \( W \) is a predefined \( q \times q \) diagonal weight matrix for each variant, and \( K \) is an \( n \times n \) covariance matrix.

The test statistic following a mixture of Chi-square distribution is:

\[ Q = (y - X\hat{\beta})' \hat{\Sigma}^{-1} GWG' \hat{\Sigma}^{-1} (y - X\hat{\beta}) \]

where the parameters are estimated under \( H_0 \) (i.e., \( H_0: \tau = 0 \))

Thus, under \( H_0: \quad y = X\beta + u + \varepsilon \quad \text{“still a linear mixed model”} \]

\[ \hat{\Sigma} = \hat{K} + \hat{\sigma}^2_E I \]
\[ \hat{\beta} = (X'\hat{\Sigma}^{-1}X)^{-1} X'\hat{\Sigma}^{-1} y \]
Under the null hypothesis ($\tau = 0$), the random intercept and time model for the $i$-th subject at time point $j$ is

$$y_{ij} = \beta_0 + t_{ij}\beta_1 + b_{0i} + t_{ij}b_{1i} + \varepsilon_{ij}$$

where $t_{ij}$ indicates time. $\beta_0$ and $\beta_1$ are the fixed effects of intercept and time, while $b_{0i}$ and $b_{1i}$ are the random effects of intercept and time for the $i$-th subject.
Methods

Longitudinal Kernel Machine (L-KM) regression for Quantitative Traits for Population Data:

For one subject, the model can be rewritten as

\[ y_i = X_i \beta + Z_i b_i + \varepsilon_i \]

We assume that there are \( m \) time points. Thus, \( y_i = (y_{i1}, y_{i2}, \ldots, y_{im})' \) is an \( m \times 1 \) vector, \( X_i \) is an \( m \times 2 \) matrix for intercept and time, \( \beta = (\beta_0 \ \beta_1) \) and \( b_i = (b_{0i} \ b_{1i}) \).

For simplicity, we did not include other covariates (which can be easily included) in the model; therefore, \( Z_i \) is the same as \( X_i \), and

\[
\text{Var}(b_i) = \begin{pmatrix} \sigma_{\text{int}}^2 & \sigma_{\text{cov}} \\ \sigma_{\text{cov}} & \sigma_{\text{time}}^2 \end{pmatrix} \Rightarrow \text{Var}(y_i) = Z_i \text{Var}(b_i) Z_i' + \sigma_E^2 I_{m \times m}
\]

For example,

\[
\text{Var}(y_i) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} \sigma_{\text{int}}^2 & \sigma_{\text{cov}} \\ \sigma_{\text{cov}} & \sigma_{\text{time}}^2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} + \begin{bmatrix} \sigma_E^2 \\ \sigma_E^2 \end{bmatrix}
\]
Methods

Longitudinal Kernel Machine (L-KM) regression for Quantitative Traits for Population Data:

For the whole data set, the variance term is:

\[
\text{Var}(y) = \mathbf{I} \otimes \mathbf{Z}_i \text{Var}(\mathbf{b}_i) \mathbf{Z}_i' + \sigma_E^2 \mathbf{I} = \Sigma
\]

where \( y \) is an \( n \cdot m \times 1 \) vector, and \( \otimes \) is the kronecker product to produce a diagonal block matrix. The variance terms \( \sigma_{\text{int}}^2, \sigma_{\text{time}}^2, \sigma_{\text{cov}}, \) and \( \sigma_E^2 \) can be estimated from the data (e.g., using the R package \textit{nlme}), and then the L-KM test statistic \( Q \) can be constructed in the same way as in the above section.
Methods

- Longitudinal Family Kernel Machine (LF-KM) regression for Quantitative Traits for Family Data:

For pedigree data, familial correlation can be added to the model as an additional random variable. Under the null hypothesis, $H_0: \tau = 0$, for the $i$-th subject in the $k$-th family at time point $j$, the random intercept and time model becomes:

$$ y_{ijk} = \beta_0 + t_{ijk}\beta_1 + b_{0ik} + t_{ijk}b_{1ik} + \delta_{ik} + \varepsilon_{ijk} $$

where $\beta_0$ and $\beta_1$ are the fixed effects of intercept and time, while $b_{0ik}$ and $b_{1ik}$ are the random effects of intercept and time. $\delta_{ik}$ is the random effect for familial correlation.
For one subject with $m$ time point observations, the model can be rewritten in vector form as:

$$y_{ik} = X_{ik}\beta + Z_{ik}b_{ik} + \delta_{ik} + \varepsilon_{ik}$$

Again, we assume $m$ time points and no other covariates; thus, $y_{ik}$ is an $m \times 1$ vector, $X_{ik}$ and $Z_{ik}$ are the same $m \times 2$ matrix for intercept and time.
Methods

Longitudinal Family Kernel Machine (LF-KM) regression for Quantitative Traits for Family Data:

For illustration, we consider the model for a trio family:

$$y_k = X_k \beta + Z_k b_k + \delta_k + \epsilon_k$$

Same as above: $Var(Z_k b_k) = I_{3 \times 3} \otimes Z_{ik} Var(b_i)Z_{ik}' = I_{3 \times 3} \otimes Z_{ik} \begin{pmatrix} \sigma_{int}^2 & \sigma_{cov} \\ \sigma_{cov} & \sigma_{time}^2 \end{pmatrix}Z_{ik}'$

New: $Var(\delta_k) = \sigma_G^2 \cdot J_k \Phi_k J_k' = \sigma_G^2 \cdot \begin{bmatrix} 1_{m \times 1} & 0_{m \times 1} & 0_{m \times 1} \\ 0_{m \times 1} & 1_{m \times 1} & 0_{m \times 1} \\ 0_{m \times 1} & 0_{m \times 1} & 1_{m \times 1} \end{bmatrix} \Phi_k \begin{bmatrix} 1_{m \times 1} & 0_{m \times 1} & 0_{m \times 1} \\ 0_{m \times 1} & 1_{m \times 1} & 0_{m \times 1} \\ 0_{m \times 1} & 0_{m \times 1} & 1_{m \times 1} \end{bmatrix}'$

$$\Phi_k = \begin{bmatrix} 1 & 0 & 0.5 \\ 0 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}$$

\text{father} \quad \text{mother} \quad \text{child}

where $y_k$ is a $3m \times 1$ vector, and $\Phi_k$ is twice the kinship matrix for a trio family:

Total: $Var(y_k) = Var(Z_k b_k) + Var(\delta_k) + \sigma_E^2 I_{3m \times 3m}$
Longitudinal Family Kernel Machine (LF-KM) regression for Quantitative Traits for Family Data:

For the whole data set with multiple families, we assume \( n \) individuals from the families. The variance term is:

\[
Var(y) = I \otimes Z_{ik} Var(b_{ik})Z_{ik}' + \sigma_G^2 \cdot J \Phi J' + \sigma_E^2 I = \Sigma
\]

where \( \sigma_{int}^2, \sigma_{time}^2, \sigma_{cov}, \) and \( \sigma_E^2 \) represent the same variance/covariance terms as in the population-based model. \( \sigma_G^2 \) represents the variance term for the random effects of familial correlation. \( \Phi \) is twice the \( n \times n \) kinship matrix obtained from the data. All the variance terms can be estimated (e.g., using the R package pedigreemm), and then the LF-KM test statistic \( Q \) can be constructed as above.
Simultaneous Studies

• Genotypes:
  - Population dataset = 1,000 × 30 rare variants;
  - Trio family dataset = 300 trios × 30 rare variants;
  - Three generation family dataset = 100 families × 30 rare variants;
  - Total = 100 genotype datasets.

• Phenotypes:
  - Type I error rate: 1000 sets of phenotypes for each genotype dataset (independent);
  - Power: 1000 sets of phenotypes for each genotype dataset (Causal variants(+/−) = 30%/0%; 20%/10%).
Results

➢ Simulation of the Type I Error Rate:

Population

Trio

Family

Three generations

(A)

(B)
Statistical Power Comparison:

Population

(A) $+/- = 30%/0\%$

(B) $+/- = 20%/10\%$

Power vs. Alpha level graphs for different methods and conditions.
Family

Trio: +/- = 30%/0%

Power

Alpha level

(A)

Trio: +/- = 20%/10%

Power

Alpha level

(B)

Three generations: +/- = 30%/0%

Power

Alpha level

(C)

Three generations: +/- = 20%/10%

Power

Alpha level

(D)
GAW18 Data Analysis Results:

- 855 subjects from 20 families were used in the analysis and each subject has up to 4 exam points;
- Assigned rare variants to a gene if they are located within a 5kb flank;
- 11,096 genes were used in the analysis;
- Used the LF-KM statistic to analyze the association of genetic variants with diastolic and systolic blood pressure that are considered heritable traits.
-log10(P-values) of the association between 11,096 genes and diastolic blood pressure

(A)

-Blog10(P-values) of the association between 11,096 genes and systolic blood pressure

(B)
Summary

- Implement L-KM for testing the association of rare variants in population samples, which simultaneously considers multiple measurements as well as LF-KM for testing the association of rare variants in family samples.

- L-KM retains the correct Type I error rate, and achieves the best power performance in population samples; LF-KM retains the correct Type I error rate, and achieves the best power performance in family samples.

- Observe potential important genes associated with blood pressure.

- The software is available (http://www.pitt.edu/~qiy17/Softwares.html).
For your final project

In the GAW18 Data Analysis:

\[ y = \beta_0 + \beta_1 \text{Time} + \beta_2 \text{Age} + \beta_3 \text{Gender} + \gamma_1 G_1 + \gamma_2 G_2 + \cdots + \gamma_q G_q + b_0 + b_1 \text{Time} + \delta + \epsilon \]

Thus, under \( H_0: \tau = 0 \)

\[ y = \beta_0 + \beta_1 \text{Time} + \beta_2 \text{Age} + \beta_3 \text{Gender} + b_0 + b_1 \text{Time} + \delta + \epsilon \]

In your final project:

• Instead of doing “gene-based” analysis on rare variants, consider analysis on (each) single common variant
• Question: treat this single common variant as fixed or random?
For your final project

**Objective:**
- Identify the genetic variants significantly associated with DBP.

**You may choose to:**
- Consider both longitudinal and familial structures
- Consider longitudinal structure (only use unrelated samples)
- Consider familial structure (use the value at the last time point, or at the averaged value)
Reference

• Michael C. Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin (2011) Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. Am J Hum Genet 89(1):82-93