Rare-Variant Kernel Machine Test

Qi Yan

Department of Pediatrics, Children’s Hospital of Pittsburgh of UPMC

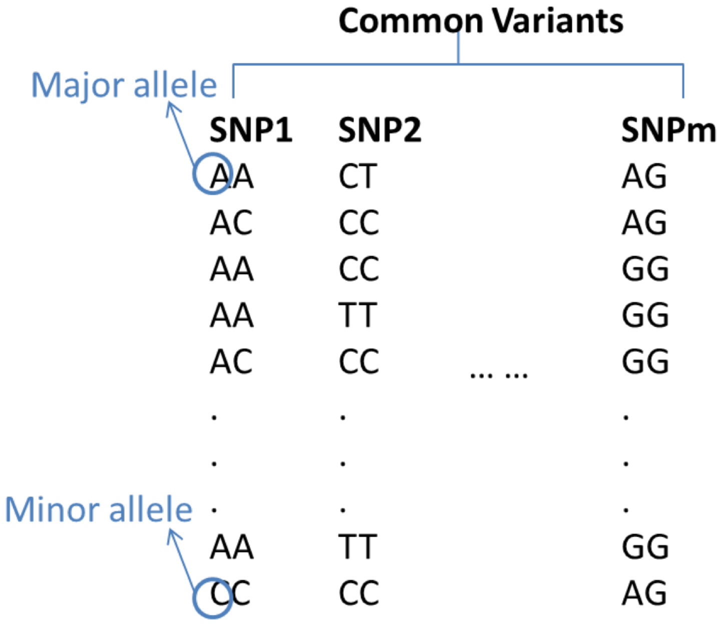
Motivation

- 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).

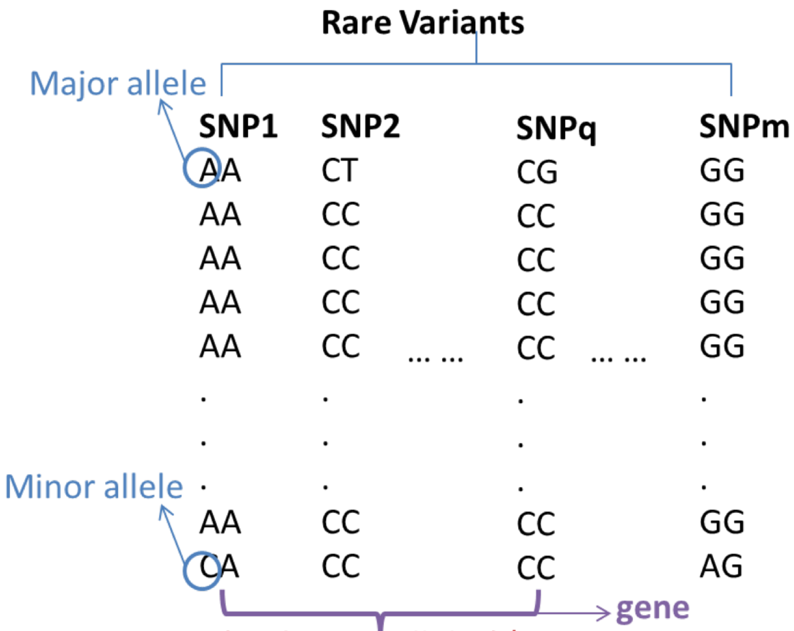


Common Variants

SNP1	SNP2		SNPm
AA	CT		AG
AC	CC		AG
AA	CC		GG
AA	TT		GG
AC	CC	… …	GG
.	.		.
.	.		.
.	.		.
AA	TT		GG
CC	CC		AG

MAF=(# of minor alleles)/2n

MAF>0.05 (common variant)



Rare Variants

SNP1	SNP2		SNPq		SNPm
AA	CT		CG		GG
AA	CC		CC		GG
AA	CC		CC		GG
AA	CC		CC		GG
AA	CC	… …	CC	… …	GG
.	.		.		.
.	.		.		.
.	.		.		.
AA	CC		CC		GG
CA	CC		CC		AG

MAF=(# of minor alleles)/2n

MAF<0.05 (rare variant)

Aims

- Gene-based rare variants test;
- Handle multiple types of traits.

Methods

➢ Sequence Kernel Association Test (SKAT):

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

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

- $\mathbf{X}$ is an $n \times p$ covariate matrix,
- $\boldsymbol{\beta}$ 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 rare genetic variants of interest,
- $\boldsymbol{\gamma}$ is a $q \times 1$ vector for the random effects of the q genetic variants,
- $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector for the random error.

$$\boldsymbol{\gamma} \sim N(0, \tau\mathbf{W})$$

$$\boldsymbol{\varepsilon} \sim N\left(0, \sigma_{\mathrm{E}}^{2}\mathbf{I}\right)$$

where $\mathbf{W}$ is a predefined $q \times q$ diagonal weight matrix for each variant

Thus, the null hypothesis H_0: $\boldsymbol{\gamma} = 0$ 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.*

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{G\gamma} + \boldsymbol{\varepsilon}$$

"linear mixed model"

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

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

$$\text{Q} = \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\widehat{\boldsymbol{\Sigma}}^{-1}\underbrace{\mathbf{GWG}'}\widehat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$$

- Called "**kernel**".
- Linear combination used here. Could be more flexible form.

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

Thus, under H_0: $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon}$ *"linear regression model, no longer mixed model"*

$$\widehat{\boldsymbol{\Sigma}} = \hat{\sigma}_E^2\mathbf{I}$$
$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{y}$$

- *The "full model" of SKAT is a linear mixed model*
- *The "null model" for the score test is a linear model*

Methods

➢ Sequence Kernel Association Test (SKAT):

Under null hypothesis, the variance of residual is

$$\mathrm{var}\big(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\big) = \hat{\sigma}_E^2 - \hat{\sigma}_E^2\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{P}_0.$$

The statistic $\mathrm{Q} = \hat{\sigma}_E^{-4}\big(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\big)'\mathbf{GWG}'\big(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\big)$ is a quadratic form of $\big(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\big)$ and follows a mixture of chi-square distributions under H_0. Thus,

$$\mathrm{Q} \sim \sum_{i=1}^{q} \lambda_i \chi_{1,i}^2$$

where λ_i is the eigenvalues of the matrix $\hat{\sigma}_E^{-4}\mathbf{W}^{\frac{1}{2}}\mathbf{G}'\mathbf{P_0}\mathbf{G}\mathbf{W}^{\frac{1}{2}}$.

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:

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

- **X** is an $n \times p$ covariate matrix,
- $\boldsymbol{\beta}$ 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* genetic variants of interest,
- $\boldsymbol{\gamma}$ is a $q \times 1$ vector for the random effects of the *q* genetic variants,
- $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector for the random error,
- ***u*** is an $n \times 1$ vector for the random effects due to covariates (e.g., relatedness in families, multivariate traits or time for longitudinal data)

Methods

➢ Kernel Machine (KM) Regression for Linear Mixed Model:

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

$$\boldsymbol{\gamma} \sim N(0, \tau\mathbf{W})$$

$$\mathbf{u} \sim N(0, \mathbf{K})$$

$$\boldsymbol{\varepsilon} \sim N\left(0, \sigma_{\mathrm{E}}^2\mathbf{I}\right)$$

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

For a linear mixed model, we use the log-likelihood

$$l = -\frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{y} - \mathbf{X\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X\beta}),$$

where $\boldsymbol{\Sigma} = \mathrm{var}(\mathbf{y}) = \tau\mathbf{GWG}' + \sigma_\delta^2\boldsymbol{\Phi} + \sigma_E^2\mathbf{I}$. In the log-likelihood, the first term $-\frac{1}{2}\log|\boldsymbol{\Sigma}|$ is fixed and independent of trait **y** when replacing **Σ** with its estimator.

Methods

➢ Kernel Machine (KM) Regression for Linear Mixed Model:

Take the first derivative

$$\frac{dl}{d\tau} = -\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{GWG}') + \frac{1}{2}(\mathbf{y}-\mathbf{X\beta})'\boldsymbol{\Sigma}^{-1}\mathbf{GWG}'\boldsymbol{\Sigma}^{-1}(\mathbf{y}-\mathbf{X\beta}),$$

The first term is fixed and independent of **y**. We take twice the second term to be derived as our test statistic Q.

$$\mathrm{Q} = \left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{GWG}'\widehat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)$$

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

Thus, under H_0: $\mathbf{y} = \mathbf{X\beta} + \mathbf{u} + \boldsymbol{\varepsilon}$ *"still a linear mixed model"*

$$\widehat{\boldsymbol{\Sigma}} = \widehat{\mathbf{K}} + \hat{\sigma}_E^2\mathbf{I}$$

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{y}$$

Methods

➢ Kernel Machine (KM) Regression for Linear Mixed Model:

Under null hypothesis, the variance of residual is

$$\mathrm{var}\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)=\mathrm{var}\left(\mathbf{y}-\mathbf{X}\left(\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{y}\right)=\widehat{\boldsymbol{\Sigma}}-\mathbf{X}\left(\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'=\mathbf{P}_0.$$

The statistic Q is a quadratic form of $\left(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)$ and follows a mixture of chi-square distributions under H_0. Thus,

$$\mathrm{Q}\sim\sum_{i=1}^{q}\lambda_i\chi_{1,i}^2$$

where λ_i is the eigenvalues of the matrix $\mathbf{W}^{\frac{1}{2}}\mathbf{G}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{P}_0\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{G}\mathbf{W}^{\frac{1}{2}}$.

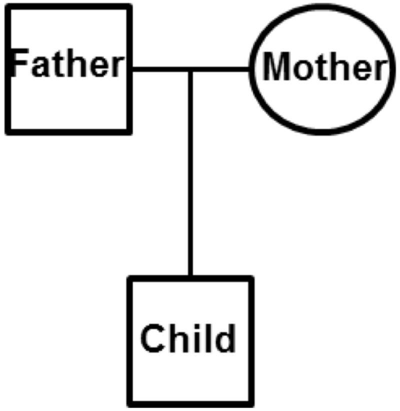
Methods (special case)

➢ Family Sequence Kernel Association Test (famSKAT) for Quantitative Traits for Family Data:

The random variable for familial correlation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\gamma} + \underset{\downarrow}{\color{red}{\mathbf{u}}} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\gamma} \sim N(0, \tau\mathbf{W}) \qquad \boldsymbol{\varepsilon} \sim N\left(0, \sigma_{\mathrm{E}}^2\mathbf{I}\right)$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\gamma} + \boldsymbol{\delta} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\delta} \sim N\left(0, \sigma_\delta^2\boldsymbol{\Phi}\right)$$



$$\boldsymbol{\Phi} = \begin{array}{cccc} \text{Father} & \text{Mother} & \text{Child} & \\ \left[\begin{array}{c} 1 \\ 0 \\ 0.5 \end{array}\right. & \begin{array}{c} 0 \\ 1 \\ 0.5 \end{array} & \left.\begin{array}{c} 0.5 \\ 0.5 \\ 1 \end{array}\right] & \begin{array}{l} \text{Father} \\ \text{Mother} \\ \text{Child} \end{array} \end{array}$$

Under the null hypothesis ($\tau = 0$), $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\delta} + \boldsymbol{\varepsilon}$

Methods (special case)

- **Family Sequence Kernel Association Test (famSKAT) for Quantitative Traits for Family Data:**

We have test statistics:

$$\mathrm{Q} = \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)' \widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{G}\mathbf{W}\mathbf{G}'\widehat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$$

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{y}$$

$$\widehat{\boldsymbol{\Sigma}} = \hat{\sigma}_{\delta}^{2}\boldsymbol{\Phi} + \hat{\sigma}_{E}^{2}\mathbf{I}$$

The statistic Q is a quadratic form of $\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$ and follows a mixture of chi-square distributions

$$\mathrm{Q} \sim \sum_{i=1}^{q} \lambda_i \chi_{1,i}^{2}$$

where λ_i is the eigenvalues of the matrix $\mathbf{W}^{\frac{1}{2}}\mathbf{G}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{P}_{\mathbf{0}}\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{G}\mathbf{W}^{\frac{1}{2}}$.

Methods (special case)

➢ Multivariate Family Kernel Machine (MF-KM) regression for Quantitative Traits for Family Data:

We consider a data set containing *m* individuals and two correlated phenotypes for illustration. The model with correlation among phenotypes and familial correlation is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\gamma} + \mathbf{h} + \boldsymbol{\varepsilon}$$

where **y** is a vector of continuous trait (i.e., **y** = (y_{11}, y_{12}, y_{21}, y_{22}, …, y_{m1}, y_{m2}) where *m* is the number of samples). **h** is the random effect of correlated phenotypes corresponding to the polygenic contribution, and **ε** is the random effect of correlated phenotypes corresponding to the random environmental contribution.

$$\mathbf{h} \sim N\left(0, \quad \boldsymbol{\Phi} \otimes \begin{pmatrix} \sigma_{G1}^2 & \sigma_{G12} \\ \sigma_{G12} & \sigma_{G2}^2 \end{pmatrix}\right) \qquad \boldsymbol{\varepsilon} \sim N\left(0, \quad \mathbf{I} \otimes \begin{pmatrix} \sigma_{E1}^2 & \sigma_{E12} \\ \sigma_{E12} & \sigma_{E2}^2 \end{pmatrix}\right)$$

Methods (special case)

- **Multivariate Family Kernel Machine (MF-KM) regression for Quantitative Traits for Family Data:**

Under the null hypothesis ($\tau = 0$), $\mathbf{y} = \mathbf{X\beta} + \mathbf{h} + \boldsymbol{\varepsilon}$

$$\mathrm{var}(\mathbf{y}) = \mathbf{\Phi} \otimes \begin{pmatrix} \sigma_{G1}^2 & \sigma_{G12} \\ \sigma_{G12} & \sigma_{G2}^2 \end{pmatrix} + \mathbf{I} \otimes \begin{pmatrix} \sigma_{E1}^2 & \sigma_{E12} \\ \sigma_{E12} & \sigma_{E2}^2 \end{pmatrix} = \mathbf{\Sigma}$$

where $\mathbf{\Phi}$ is twice the $m \times m$ kinship matrix obtained either from familial relationship and $\otimes$ is the kronecker product. σ_{G1}^2, σ_{G2}^2, σ_{G12}, σ_{E1}^2, σ_{E2}^2 and σ_{E12} represent the polygenic variances of the first and second traits, the polygenic covariance between the two traits, the environmental variances of the first and second traits, and the environmental covariance between the two traits

Methods (special case)

- **Multivariate Family Kernel Machine (MF-KM) regression for Quantitative Traits for Family Data:**

We have test statistics:

$$\mathrm{Q} = \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{G}\mathbf{W}\mathbf{G}'\widehat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$$

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{y}$$

$$\widehat{\boldsymbol{\Sigma}} = \boldsymbol{\Phi}\otimes\begin{pmatrix} \hat{\sigma}_{G1}^2 & \hat{\sigma}_{G12} \\ \hat{\sigma}_{G12} & \hat{\sigma}_{G2}^2 \end{pmatrix} + \mathbf{I}\otimes\begin{pmatrix} \hat{\sigma}_{E1}^2 & \hat{\sigma}_{E12} \\ \hat{\sigma}_{E12} & \hat{\sigma}_{E2}^2 \end{pmatrix}$$

The statistic Q is a quadratic form of $\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$ and follows a mixture of chi-square distributions

$$\mathrm{Q} \sim \sum_{i=1}^{q} \lambda_i \chi_{1,i}^2$$

where λ_i is the eigenvalues of the matrix $\mathbf{W}^{\frac{1}{2}}\mathbf{G}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{P}_0\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{G}\mathbf{W}^{\frac{1}{2}}$.

Methods (special case)

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

- With longitudinal studies, may treat intercept and (continuous) time (slope) as both fixed and random effects.
- Different covariance structures such as compound symmetry, autoregressive, and Toeplitz, can be easily implemented under this framework.

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. β_0 and β_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 (special case)

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

For one subject, the null model can be rewritten as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i$$

We assume that there are m time points. Thus, $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{im})'$ is an $m \times 1$ vector, $\mathbf{X}_i$ is an $m \times 2$ matrix for intercept and time,$\boldsymbol{\beta} = (\beta_0 \quad \beta_1)$ and $\mathbf{b}_i = (b_{0i} \quad b_{1i})$. For simplicity, we did not include other covariates (which can be easily included) in the model; therefore, $\mathbf{Z}_i$ is the same as $\mathbf{X}_i$, and

$$Var(\mathbf{b}_i) = \begin{pmatrix} \sigma_{int}^2 & \sigma_{cov} \\ \sigma_{cov} & \sigma_{time}^2 \end{pmatrix} \rightarrow Var(\mathbf{y}_i) = \mathbf{Z}_i Var(\mathbf{b}_i)\mathbf{Z}_i' + \sigma_E^2\mathbf{I}_{m\times m}$$

For example,

$$Var(\boldsymbol{y}_i) = \left[\overset{\mathbf{Z}_i}{\begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}} \overset{Var(\mathbf{b}_i)}{\begin{bmatrix} \sigma_{int}^2 & \sigma_{cov} \\ \sigma_{cov} & \sigma_{time}^2 \end{bmatrix}} \overset{\mathbf{Z}_i'}{\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}} \right] + \overset{\sigma_E^2\mathbf{I}_{m\times m}}{\begin{bmatrix} \sigma_E^2 & \\ & \sigma_E^2 \end{bmatrix}}$$

Methods (special case)

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

For the whole data set, the variance term is:

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

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 σ_{int}^2, σ_{time}^2, σ_{cov}, and σ_E^2 can be estimated from the data (e.g., using the R package nlme), and then the L-KM test statistic Q can be constructed in the same way as in the above section.

We have test statistics:

$$\mathrm{Q} = \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\widehat{\mathbf{\Sigma}}^{-1}\mathbf{GWG}'\widehat{\mathbf{\Sigma}}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$$

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\widehat{\mathbf{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\mathbf{\Sigma}}^{-1}\mathbf{y}$$

$$\widehat{\mathbf{\Sigma}} = \mathbf{I}\otimes\mathbf{Z}_i \widehat{Var(\mathbf{b}_i)}\mathbf{Z}_i{}' + \hat{\sigma}_E^2\mathbf{I}$$

Methods

➢ Kernel Machine (KM) Regression for Generalized 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 binary trait $\boldsymbol{y}$ follows a linear mixed model:

$$\text{logit}\big(P(\mathbf{y}=1)\big)=\text{logit}(\boldsymbol{\mu})=\eta=\mathbf{X}\boldsymbol{\beta}+\mathbf{G}\boldsymbol{\gamma}+\mathbf{u}$$

where $\boldsymbol{\mu}$ is the mean of $\mathbf{y}$.

$$\boldsymbol{\gamma}\sim N(0,\tau\mathbf{W})$$

$$\mathbf{u}\sim N(0,\mathbf{K})$$

where $\mathbf{W}$ is a predefined $q \times q$ diagonal weight matrix for each variant, and $\mathbf{K}$ is an $n \times n$ covariance matrix

Methods

➢ Kernel Machine (KM) Regression for Generalized Linear Mixed Model:

For a generalized linear mixed model, we use the quasi-likelihood

$$ql = -\frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}(\mathbf{y}^* - \mathbf{X\beta})'\mathbf{\Sigma}^{-1}(\mathbf{y}^* - \mathbf{X\beta}),$$

- $\mathbf{y}^*$ is the working trait vector that is equal to $\mathbf{H}^{-1}(\mathbf{y} - \boldsymbol{\mu}) + \eta$
- $\mathbf{\Sigma} = \mathbf{R}^{-1} + \tau\mathbf{GWG}' + \mathbf{K}$,
- $\mathbf{H}$ is a diagonal matrix with diagonal elements equal to $\frac{\partial\mu_i}{\partial\eta_i} = {\partial\frac{e^{\eta_i}}{1+e^{\eta_i}}}\Big/{\partial\eta_i} = \frac{e^{\eta_i}}{1+e^{\eta_i}}\left(1 - \frac{e^{\eta_i}}{1+e^{\eta_i}}\right) = \mu_i(1-\mu_i)$,
- $\mathbf{R}$ is also a diagonal matrix with diagonal elements equal to $\left\{\mathrm{var}(\mathrm{y}_i|\gamma_i, \delta_i)/\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2\right\}^{-1} = \mu_i(1-\mu_i)$.
- The first term $-\frac{1}{2}\log|\mathbf{\Sigma}|$ is fixed and independent of phenotype $\mathbf{y}$ when replacing $\mathbf{\Sigma}$ with its estimator.

Methods

Kernel Machine (KM) Regression for Generalized Linear Mixed Model:

Take the first derivative

$$\frac{dql}{d\tau} = -\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{GWG}') + \frac{1}{2}(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}\mathbf{GWG}'\boldsymbol{\Sigma}^{-1}(\mathbf{y}^* - \mathbf{X}\boldsymbol{\beta})$$

Following the same rationale as described in above section, we take twice the second term to be derived as our test statistic Q.

$$\mathrm{Q} = \left(\mathbf{y}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\right)'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{GWG}'\hat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{y}^* - \mathbf{X}\hat{\boldsymbol{\beta}}\right).$$

Under H_0: $\mathrm{logit}\left(P(\mathbf{y} = 1)\right) = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \quad \overset{\text{equivalent}}{\Longleftrightarrow} \quad \mathbf{y}^* = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\varepsilon}^* \qquad \boldsymbol{\varepsilon}^* \sim N(0, \mathbf{R}^{-1})$

"a generalized linear mixed model"

$\mathbf{y}^*$ is the working trait vector

When final $\mathbf{y}^*$:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\boldsymbol{\Sigma}}^{-1}\mathbf{y}^*; \quad \hat{\boldsymbol{\Sigma}} = \hat{\mathbf{K}} + \hat{\mathbf{R}}^{-1}$$

Methods

➢ Kernel Machine (KM) Regression for Generalized Linear Mixed Model:

Under null hypothesis, the variance of residual is

$$\mathrm{var}\left(\mathbf{y}^{*}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)=\mathrm{var}\left(\mathbf{y}^{*}-\mathbf{X}\left(\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{y}^{*}\right)=\widehat{\boldsymbol{\Sigma}}-\mathbf{X}\left(\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'=\mathbf{P}_0.$$

The statistic Q is a quadratic form of $\left(\mathbf{y}^{*}-\mathbf{X}\widehat{\boldsymbol{\beta}}\right)$ and follows a mixture of chi-square distributions under H_0. Thus,

$$\mathrm{Q}\sim\sum_{i=1}^{q}\lambda_i\chi^2_{1,i}$$

where λ_i is the eigenvalues of the matrix $\mathbf{W}^{\frac{1}{2}}\mathbf{G}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{P}_0\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{G}\mathbf{W}^{\frac{1}{2}}$.

Methods (special case)

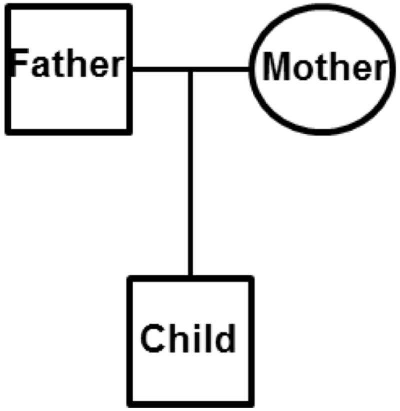
- **Family Sequence Kernel Association Test (F-SKAT) for Binary Traits for Family Data:**

The random variable for familial correlation

$$\mathrm{logit}\big(P(\mathbf{y}=1)\big) = \mathbf{X\beta} + \mathbf{G\gamma} + \mathbf{u} \qquad \boldsymbol{\gamma} \sim N(0, \tau\mathbf{W})$$

$$\downarrow$$

$$\mathrm{logit}\big(P(\mathbf{y}=1)\big) = \mathbf{X\beta} + \mathbf{G\gamma} + \boldsymbol{\delta} \qquad \boldsymbol{\delta} \sim N\big(0, \sigma_\delta^2 \boldsymbol{\Phi}\big)$$



$$\boldsymbol{\Phi} = \begin{array}{c|ccc|l} & \text{Father} & \text{Mother} & \text{Child} & \\ \hline & 1 & 0 & 0.5 & \text{Father} \\ & 0 & 1 & 0.5 & \text{Mother} \\ & 0.5 & 0.5 & 1 & \text{Child} \end{array}$$

Under the null hypothesis ($\tau = 0$), $\mathrm{logit}\big(P(\mathbf{y}=1)\big) = \mathbf{X\beta} + \boldsymbol{\delta}$

Methods (special case)

➢ Family Sequence Kernel Association Test (F-SKAT) for Binary Traits for Family Data:

We have test statistics:

$$\mathrm{Q} = \left(\mathbf{y}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{G}\mathbf{W}\mathbf{G}'\widehat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{y}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}\right).$$

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{y}^*$$

$$\widehat{\boldsymbol{\Sigma}} = \hat{\sigma}_\delta^2\boldsymbol{\Phi} + \widehat{\mathbf{R}}^{-1}$$

The statistic Q is a quadratic form of $\left(\mathbf{y}^* - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$ and follows a mixture of chi-square distributions

$$\mathrm{Q} \sim \sum_{i=1}^{q} \lambda_i \chi_{1,i}^2$$

where λ_i is the eigenvalues of the matrix $\mathbf{W}^{\frac{1}{2}}\mathbf{G}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{P}_0\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{G}\mathbf{W}^{\frac{1}{2}}$.

- ➢ This can be simplified to SKAT for binary population data

Reference

- Michael C. Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke, and Xihong Lin Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. Am J Hum Genet, 2011, 89(1):82-93.
- Han Chen, James B. Meigs, and Josée Dupuis. "Sequence Kernel Association Test for Quantitative Traits in Family Samples". Genet Epidemiol. 2013 Feb; 37(2): 196–204.
- Qi Yan, Daniel E. Weeks, Hemant K. Tiwari, Nengjun Yi, Kui Zhang, Guimin Gao, Wan-Yu Lin, Xiang-Yang Lou, Wei Chen, and Nianjun Liu. "Rare-Variant Kernel Machine Test for Longitudinal Data from Population and Family Samples". Human Heredity, 80:126-138.
- Qi Yan, Daniel E. Weeks, Juan C. Celedon, Hemant K. Tiwari, Bingshan Li, Xiaojing Wang, Wan-Yu Lin, Wei Chen, and Nianjun Liu. "Associating Multivariate Quantitative Phenotypes with Genetic Variants in Family Samples with a Novel Kernel Machine Regression Method". Genetics, 2015 Dec;201(4):1329-39.
- Qi Yan, Hemant K. Tiwari, Nengjun Yi, Guimin Gao, Wan-Yu Lin, Xiang-Yang Lou, and Nianjun Liu. "Sequence Kernel Association Test for Dichotomous Traits in Family Sample under Generalized Linear Mixed Model". Human Heredity, 2016, 79(2):60-68.