A Multiple Imputation Approach to Test a SNP-set Association with A Censored Trait in the Presence of a Familial Dependence

Lajmi Lakhal-Chaieb

Département de Mathématiques et Statistique, Université Laval, Québec

The Biostatistics Seminar Series, Dalla Lana School of Public Health, University of Toronto, 2015-01-27.
Table of contents

1. Introduction
2. Association test
3. Numerical investigations
4. Conclusion
Multi-marker association tests

- Genome-wide association studies (GWAS) have identified hundreds of common genetic variants associated with complex human phenotypes.
Multi-marker association tests

- Genome-wide association studies (GWAS) have identified hundreds of common genetic variants associated with complex human phenotypes.
- The genetic variation explained by these common variants is small for most diseases (Dickson *et al.*, PLoS Biology 2010).
Multi-marker association tests

- Genome-wide association studies (GWAS) have identified hundreds of common genetic variants associated with complex human phenotypes.
- The genetic variation explained by these common variants is small for most diseases (Dickson et al., PLoS Biology 2010).
- With rapid development in sequencing technologies and lowering of sequencing costs, it is now possible to collect reliable information concerning low frequent and rare variants, defined as variants with a minor allele frequency smaller than 5%.
Multi-marker association tests

- Genome-wide association studies (GWAS) have identified hundreds of common genetic variants associated with complex human phenotypes.
- The genetic variation explained by these common variants is small for most diseases (Dickson et al., PLoS Biology 2010).
- With rapid development in sequencing technologies and lowering of sequencing costs, it is now possible to collect reliable information concerning low frequent and rare variants, defined as variants with a minor allele frequency smaller than 5%.
- Rare variants are thought to explain a portion of the missed heritability for complex diseases.
Multi-marker association tests

Due to the low frequency of many genetic variants identified by re-sequencing, statistical methods testing a single genetic variant at a time have low power for detecting associations.
Multi-marker association tests

- Due to the low frequency of many genetic variants identified by re-sequencing, statistical methods testing a single genetic variant at a time have low power for detecting associations.
- Moreover, many phenotypes are related to multiple rare and common variants through complex relationships.
Multi-marker association tests

- Due to the low frequency of many genetic variants identified by re-sequencing, statistical methods testing a single genetic variant at a time have low power for detecting associations.

- Moreover, many phenotypes are related to multiple rare and common variants through complex relationships.

- Consequently, many attempts have been made to develop multi-marker association tests that can test jointly multiple common and/or rare variants (Wang et al., BMC genetics 2013).
Multi-marker association tests

- Due to the low frequency of many genetic variants identified by re-sequencing, statistical methods testing a single genetic variant at a time have low power for detecting associations.
- Moreover, many phenotypes are related to multiple rare and common variants through complex relationships.
- Consequently, many attempts have been made to develop multi-marker association tests that can test jointly multiple common and/or rare variants (Wang et al., BMC genetics 2013).
- The prime interest of such tests is to check whether the group of genetic variants in the chosen region has an influence on the phenotype under investigation.
Multi-marker association tests for common variants use standard regression techniques (LRT, Wald, Score).
Multi-marker association tests

- Multi-marker association tests for common variants use standard regression techniques (LRT, Wald, Score).

- In the presence of rare variants, standard regression strategies are underpowered to detect associations.
Multi-marker association tests

- Multi-marker association tests for common variants use standard regression techniques (LRT, Wald, Score).

- In the presence of rare variants, standard regression strategies are underpowered to detect associations.

- Existing approaches to test rare variants are mainly categorized into three classes:
Multi-marker association tests

- Multi-marker association tests for common variants use standard regression techniques (LRT, Wald, Score).

- In the presence of rare variants, standard regression strategies are underpowered to detect associations.

- Existing approaches to test rare variants are mainly categorized into three classes:
  - Bunden tests (Lee et al. AJHG 2014).
Multi-marker association tests

- Multi-marker association tests for common variants use standard regression techniques (LRT, Wald, Score).

- In the presence of rare variants, standard regression strategies are underpowered to detect associations.

- Existing approaches to test rare variants are mainly categorized into three classes:
  - Bunden tests (Lee et al. AJHG 2014).
  - Variance-components tests (Wu et al. AJHG 2011).
Multi-marker association tests

- Multi-marker association tests for common variants use standard regression techniques (LRT, Wald, Score).

- In the presence of rare variants, standard regression strategies are underpowered to detect associations.

- Existing approaches to test rare variants are mainly categorized into three classes:
  - Bunden tests (Lee et al. AJHG 2014).
  - Variance-components tests (Wu et al. AJHG 2011).
  - Combination of Bunden and Variance-components tests (Ionita-Laza et al. AJHG 2013).
Family-based designs have received a growing interest in genetics as including related individuals allows one to obtain larger samples of better quality and can potentially increase the power of association tests (Kazma and Bailey, Gen. Epi. 2011).
Family-based design

- Family-based designs have received a growing interest in genetics as including related individuals allows one to obtain larger samples of better quality and can potentially increase the power of association tests (Kazma and Bailey, Gen. Epi. 2011).

- However, failure to appropriately take into account the familial correlation may yield an inflated type 1 error and/or significant loss of power.
Family-based design

- Family-based designs have received a growing interest in genetics as including related individuals allows one to obtain larger samples of better quality and can potentially increase the power of association tests (Kazma and Bailey, Gen. Epi. 2011).

- However, failure to appropriately take into account the familial correlation may yield an inflated type 1 error and/or significant loss of power.

- Consequently, several authors developed region-based association tests in the presence of familial correlation (Oualkacha et al. Gen. Epi. 2013).
Most approaches cited above were developed for dichotomous and quantitative traits and very little attention has been paid to time-to-event outcomes in the presence of right-censoring.
Censored trait

- Most approaches cited above were developed for dichotomous and quantitative traits and very little attention has been paid to time-to-event outcomes in the presence of right-censoring.

- Lin et al. (Gen. Epi, 2011) and Chen et al. (Gen. Epi., 2014) considered variance-components tests for survival outcomes.
Most approaches cited above were developed for dichotomous and quantitative traits and very little attention has been paid to time-to-event outcomes in the presence of right-censoring.

Lin *et al.* (Gen. Epi, 2011) and Chen *et al.* (Gen. Epi., 2014) considered variance-components tests for survival outcomes.

However, these few methods for censored traits do not consider family-based designs.
Censored trait

- Most approaches cited above were developed for dichotomous and quantitative traits and very little attention has been paid to time-to-event outcomes in the presence of right-censoring.

- Lin et al. (Gen. Epi, 2011) and Chen et al. (Gen. Epi., 2014) considered variance-components tests for survival outcomes.

- However, these few methods for censored traits do not consider family-based designs.

- In this presentation, we propose a genomic region association test for censored traits in the presence of familial dependencies.
Data and Model

Let $T_i$ be the age-at-onset of the disease under investigation and $G_i = (G_{i1}, \cdots, G_{is})^\top$ and $X_i = (X_{i1}, \cdots, X_{ip})^\top$ be row vectors of $s$ genotypes and $p$ non-genetic covariates, respectively, for individual $i$. 
Data and Model

- Let $T_i$ be be the age-at-onset of the disease under investigation and $G_i = (G_{i1}, \ldots, G_{is})^\top$ and $X_i = (X_{i1}, \ldots, X_{ip})^\top$ be row vectors of $s$ genotypes and $p$ non-genetic covariates, respectively, for individual $i$.

- The genotype data are the counts of the minor allele frequencies.
Data and Model

- Let $T_i$ be the age-at-onset of the disease under investigation and $G_i = (G_{i1}, \cdots, G_{is})^\top$ and $X_i = (X_{i1}, \cdots, X_{ip})^\top$ be row vectors of $s$ genotypes and $p$ non-genetic covariates, respectively, for individual $i$.
- The genotype data are the counts of the minor allele frequencies.
- The assumed proportional hazards model is
  \[
  \lambda(t_i | G_i, X_i) = \lambda_0(t_i) e^{G_i \beta + X_i \xi},
  \]
  where $\lambda$ and $\lambda_0$ are the conditional and the baseline hazard functions, respectively and $\beta$ and $\xi$ are regression coefficients.
Data and Model

- Let $T_i$ be the age-at-onset of the disease under investigation and $G_i = (G_{i1}, \cdots, G_{is})^\top$ and $X_i = (X_{i1}, \cdots, X_{ip})^\top$ be row vectors of $s$ genotypes and $p$ non-genetic covariates, respectively, for individual $i$.

- The genotype data are the counts of the minor allele frequencies.

- The assumed proportional hazards model is

$$
\lambda(t_i|G_i, X_i) = \lambda_0(t_i)e^{G_i\beta + X_i\xi},
$$

where $\lambda$ and $\lambda_0$ are the conditional and the baseline hazard functions, respectively and $\beta$ and $\xi$ are regression coefficients.

- $H_0: \beta_1 = \cdots = \beta_s = 0.$
Data and Model

In the presence of right-censoring, one only observes \( n \) quadruplets \( \{(Y_i, \delta_i, G_i, X_i), i = 1, \cdots, n\} \), where
In the presence of right-censoring, one only observes $n$ quadruplets $\{(Y_i, \delta_i, G_i, X_i), i = 1, \cdots, n\}$, where

- $Y_i = \min(T_i, C_i)$ is the observed failure time.
Data and Model

In the presence of right-censoring, one only observes $n$ quadruplets $\{(Y_i, \delta_i, G_i, X_i), i = 1, \cdots, n\}$, where

- $Y_i = \min(T_i, C_i)$ is the observed failure time.
- $\delta_i = I(T_i < C_i)$ is the censoring indicator.
Data and Model

In the presence of right-censoring, one only observes \( n \) quadruplets \( \{(Y_i, \delta_i, G_i, X_i), i = 1, \cdots, n\} \), where

- \( Y_i = \min(T_i, C_i) \) is the observed failure time.

- \( \delta_i = I(T_i < C_i) \) is the censoring indicator.

- \( C_i \) is the censoring variable, assumed to be independent from \( T_i \).
In the case of unrelated persons, i.e. $n$ independent observations, Lin et al. (Gen. Epi., 2011) and Chen et al. (Gen. Epi., 2014) proposed tests for $H_0 : \beta_1 = \cdots = \beta_s = 0$. Both test statistics are expressed in terms of $Q_0 = M^\top K M$, where
In the case of unrelated persons, i.e. \( n \) independent observations, Lin et al. (Gen. Epi., 2011) and Chen et al. (Gen. Epi., 2014) proposed tests for \( H_0 : \beta_1 = \cdots = \beta_s = 0 \). Both test statistics are expressed in terms of \( Q_0 = M^\top K M \), where

\[ M = (M_1, \cdots, M_n)^\top \]

is a vector of martingales residuals estimated under the null hypothesis.
Association test for unrelated persons

In the case of unrelated persons, i.e. $n$ independent observations, Lin et al. (Gen. Epi., 2011) and Chen et al. (Gen. Epi., 2014) proposed tests for $H_0 : \beta_1 = \cdots = \beta_s = 0$. Both test statistics are expressed in terms of $Q_0 = M^\top KM$, where

- $M = (M_1, \cdots, M_n)^\top$ is a vector of martingales residuals estimated under the null hypothesis.
- $K = GWG^\top$. 
Association test for unrelated persons

In the case of unrelated persons, i.e. $n$ independent observations, Lin et al. (Gen. Epi., 2011) and Chen et al. (Gen. Epi., 2014) proposed tests for $H_0 : \beta_1 = \cdots = \beta_s = 0$. Both test statistics are expressed in terms of $Q_0 = M^\top KM$, where

- $M = (M_1, \cdots, M_n)^\top$ is a vector of martingales residuals estimated under the null hypothesis.
- $K = GWG^\top$.
- $G$ is an $n \times s$ matrix with rows $G_i$. 
Association test for unrelated persons

In the case of unrelated persons, i.e. \( n \) independent observations, Lin et al. (Gen. Epi., 2011) and Chen et al. (Gen. Epi., 2014) proposed tests for \( H_0 : \beta_1 = \cdots = \beta_s = 0 \). Both test statistics are expressed in terms of \( Q_0 = M^\top KM \), where

- \( M = (M_1, \cdots, M_n)^\top \) is a vector of martingales residuals estimated under the null hypothesis.
- \( K = GWG^\top \).
- \( G \) is an \( n \times s \) matrix with rows \( G_i \).
- \( W = \text{diag}\{w_1, \cdots, w_s\} \) is an \( s \times s \) diagonal matrix with the weights to be used for the \( s \) variants.
Association test for unrelated persons

In the case of unrelated persons, i.e. $n$ independent observations, Lin et al. (Gen. Epi., 2011) and Chen et al. (Gen. Epi., 2014) proposed tests for $H_0: \beta_1 = \cdots = \beta_s = 0$. Both test statistics are expressed in terms of $Q_0 = M^\top K M$, where

- $M = (M_1, \cdots, M_n)^\top$ is a vector of martingales residuals estimated under the null hypothesis.
- $K = GWG^\top$.
- $G$ is an $n \times s$ matrix with rows $G_i$.
- $W = \text{diag}\{w_1, \cdots, w_s\}$ is an $s \times s$ diagonal matrix with the weights to be used for the $s$ variants.
- The $n \times n$ matrix $K$ is the weighted linear kernel matrix whose entries $K_{ij} = \sum_{k=1}^s w_k G_{ik} G_{jk}$ capture the similarities in pairs of individuals in the tested region.
Kinship-adjusted association test

In the presence of a family-based design, the \( n \) individuals of the sample are clustered into families but for simplicity we do not introduce a second subscript.
Kinship-adjusted association test

- In the presence of a family-based design, the $n$ individuals of the sample are clustered into families but for simplicity we do not introduce a second subscript.

- By Cheng et al. (Biometrika, 1995), the proportional hazards model can be alternatively written as

$$H(T_i) = -G_i \beta - X_i \xi + \epsilon_i,$$

where $H(\cdot)$ is an unknown monotone increasing function and $\epsilon$ follows the extreme value distribution with CDF

$$F(x) = P(\epsilon \leq x) = \exp\{-\exp(x)\}, \quad -\infty \leq x \leq \infty.$$
Kinship-adjusted association test

To model the familial association, we assume that the joint distribution of $\epsilon = (\epsilon_1, \cdots, \epsilon_n)^\top$ follows a Gaussian copula with correlation matrix $\Gamma$. 
Kinship-adjusted association test

- To model the familial association, we assume that the joint distribution of \( \varepsilon = (\varepsilon_1, \cdots, \varepsilon_n)^\top \) follows a Gaussian copula with correlation matrix \( \Gamma \).

- The diagonal elements of \( \Gamma \) are equal to 1.
To model the familial association, we assume that the joint distribution of $\mathbf{e} = (\mathbf{e}_1, \cdots, \mathbf{e}_n)^\top$ follows a Gaussian copula with correlation matrix $\Gamma$.

The diagonal elements of $\Gamma$ are equal to 1.

The off-diagonal entries are $\Gamma_{ij} = h^2 \varphi_{ij}$, where $h^2$ measures the polygenic heritability and $\varphi_{ij}$ reflects the proportion of the genome that is IBD between the pair of individuals $i$ and $j$. 
Kinship-adjusted association test

- To model the familial association, we assume that the joint distribution of $\mathbf{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^\top$ follows a Gaussian copula with correlation matrix $\Gamma$.

- The diagonal elements of $\Gamma$ are equal to 1.

- The off-diagonal entries are $\Gamma_{ij} = h^2 \varphi_{ij}$, where $h^2$ measures the polygenic heritability and $\varphi_{ij}$ reflects the proportion of the genome that is IBD between the pair of individuals $i$ and $j$.

- It is easy to see that $\Gamma = h^2 \varphi + (1 - h^2)I_n$, where $I_n$ is the $n \times n$ identity matrix.
The likelihood ratio test and the Wald test require the estimation of $\beta$, which may not be possible in the presence of rare variants (Wu et al., AJHG 2011).
Kinship-adjusted association test

- The likelihood ratio test and the Wald test require the estimation of $\beta$, which may not be possible in the presence of rare variants (Wu et al., AJHG 2011).

- On the other hand, the score test involves partial derivatives of the log-likelihood function, which is complicated and computationally demanding with censored observations in the presence of a family-based design.
The likelihood ratio test and the Wald test require the estimation of $\beta$, which may not be possible in the presence of rare variants (Wu et al., AJHG 2011).

On the other hand, the score test involves partial derivatives of the log-likelihood function, which is complicated and computationally demanding with censored observations in the presence of a family-based design.

In this paper, we consider an alternative approach.
Kinship-adjusted association test

Let \( q_i = \Phi^{-1}[F(\varepsilon_i)] \) be the inverse normal score for individual \( i \).
Let $q_i = \Phi^{-1}[F(\varepsilon_i)]$ be the inverse normal score for individual $i$.

Under the proposed mode, $q = (q_1, \ldots, q_n)^\top$ follows a multivariate normal distribution with mean zero and covariance matrix $\Gamma$. 

Kinship-adjusted association test
Let \( q_i = \Phi^{-1}[F(\varepsilon_i)] \) be the inverse normal score for individual \( i \).

Under the proposed mode, \( q = (q_1, \cdots, q_n)^\top \) follows a multivariate normal distribution with mean zero and covariance matrix \( \Gamma \).

It follows that under the null hypothesis, \( q^\top Kq \) is distributed as a linear combination of independent chi-squared random variables \( q^\top Kq \sim \sum_{l=1}^{L} \mu_l \chi^2_{l,1} \), where \( 0 < \mu_1 < \cdots < \mu_L \) are the \( L \) positive eigenvalues of \( \Gamma^{1/2} K \Gamma^{1/2} \).
Kinship-adjusted association test

- Let $q_i = \Phi^{-1}[F(\varepsilon_i)]$ be the inverse normal score for individual $i$.
- Under the proposed mode, $q = (q_1, \ldots, q_n)^\top$ follows a multivariate normal distribution with mean zero and covariance matrix $\Gamma$.
- It follows that under the null hypothesis, $q^\top K q$ is distributed as a linear combination of independent chi-squared random variables $q^\top K q \sim \sum_{l=1}^{L} \mu_l \chi^2_{l,1}$, where $0 < \mu_1 < \cdots < \mu_L$ are the $L$ positive eigenvalues of $\Gamma^{1/2} K \Gamma^{1/2}$.
- Motivated by these results, we propose the use of the test statistics $r^\top K r$, where $r = (r_1, \cdots, r_n)^\top$ is a vector of appropriately defined residuals, computed under the null hypothesis.
Estimation under the null hypothesis

- The parameters involved in our model under the null hypothesis are $H$, $\xi$ and $h^2$. 
The parameters involved in our model under the null hypothesis are $H$, $\xi$ and $h^2$.

We propose to estimate these parameters under the null hypothesis using a two-stage procedure in the spirit of Othus and Li (Stat. in BioSc., 2010).
Estimation under the null hypothesis

- The parameters involved in our model under the null hypothesis are $H$, $\xi$ and $h^2$.

- We propose to estimate these parameters under the null hypothesis using a two-stage procedure in the spirit of Othus and Li (Stat. in BioSc., 2010).

- In the first stage, estimates $\hat{H}$ and $\hat{\xi}$ are obtained using the algorithm of Chen et al. (Biometrika, 2002).
Estimation under the null hypothesis

- The parameters involved in our model under the null hypothesis are $H$, $\xi$ and $h^2$.

- We propose to estimate these parameters under the null hypothesis using a two-stage procedure in the spirit of Othus and Li (Stat. in BioSc., 2010).

- In the first stage, estimates $\hat{H}$ and $\hat{\xi}$ are obtained using the algorithm of Chen et al. (Biometrika, 2002).

- In the second stage, we estimate $h^2$ as follows.
Estimation under the null hypothesis

- We begin by computing $\hat{q}_i = \Phi^{-1}[F(\hat{H}(Y_i) + X_i\hat{\xi})]$. 
Estimation under the null hypothesis

- We begin by computing $\hat{q}_i = \Phi^{-1}[F(\hat{H}(Y_i) + X_i\hat{\xi})]$.

- Actually, $q_i$ is a censored version of $\Phi^{-1}[F(\hat{H}(T_i) + X_i\hat{\xi})]$, which is not observed when $\delta_i = 0$. 
Estimation under the null hypothesis

- We begin by computing \( \hat{q}_i = \Phi^{-1}[F(\hat{H}(Y_i) + X_i\hat{\xi})] \).

- Actually, \( q_i \) is a censored version of \( \Phi^{-1}[F(\hat{H}(T_i) + X_i\hat{\xi})] \), which is not observed when \( \delta_i = 0 \).

- The censoring pattern is preserved as \( \Phi^{-1}[F(\hat{H}(T_i) + X_i\hat{\xi})] \) is a non-decreasing transformation of \( T_i \).
Estimation under the null hypothesis

- We begin by computing $\hat{q}_i = \Phi^{-1}[F(\hat{H}(Y_i) + X_i\hat{\xi})]$.

- Actually, $q_i$ is a censored version of $\Phi^{-1}[F(\hat{H}(T_i) + X_i\hat{\xi})]$, which is not observed when $\delta_i = 0$.

- The censoring pattern is preserved as $\Phi^{-1}[F(\hat{H}(T_i) + X_i\hat{\xi})]$ is a non-decreasing transformation of $T_i$.

- The estimate $\hat{h}^2$ is then obtained by maximizing the likelihood function of the right-censored sample $\{(\hat{q}_i, \delta_i), i = 1, \ldots, n\}$.
We begin by computing $\hat{q}_i = \Phi^{-1}[F(\hat{H}(Y_i) + X_i \hat{\xi})]$.

Actually, $q_i$ is a censored version of $\Phi^{-1}[F(\hat{H}(T_i) + X_i \hat{\xi})]$, which is not observed when $\delta_i = 0$.

The censoring pattern is preserved as $\Phi^{-1}[F(\hat{H}(T_i) + X_i \hat{\xi})]$ is a non-decreasing transformation of $T_i$.

The estimate $\hat{h}^2$ is then obtained by maximizing the likelihood function of the right-censored sample $\{(\hat{q}_i, \delta_i), i = 1, \ldots, n\}$.

Computational details of the likelihood function involving a multivariate normal distribution in the presence of right-censoring are given in Othus and Li (Stat. in BioSc., 2010).
Test statistics and \( p \)-value computation

- Imputation approaches treating censored observation as partially missing data have become very popular in recent years (Lapidus et al., Stat. Med., 2014).
Test statistics and $p$-value computation

- Imputation approaches treating censored observation as partially missing data have become very popular in recent years (Lapidus et al., Stat. Med., 2014).
- In this work, we propose to use an imputation procedure to replace the censored $\hat{q}_i$’s by imputed values in order to obtain a completed vector of residuals, whose joint distribution can be easily approximated.
Test statistics and $p$-value computation

- Imputation approaches treating censored observation as partially missing data have become very popular in recent years (Lapidus et al., Stat. Med., 2014).

- In this work, we propose to use an imputation procedure to replace the censored $\hat{q}_i$’s by imputed values in order to obtain a completed vector of residuals, whose joint distribution can be easily approximated.

- Without loss of generality, we rearrange the indices \{1, \cdots, n\} so that $\delta_1 = \cdots = \delta_{n_1} = 1$ and $\delta_{n_1+1} = \cdots = \delta_n = 0$, where $n_1 = \sum_{i=1}^{n} \delta_i$ and consider the partition $\hat{q} = (\hat{q}^{(1)}, \hat{q}^{(0)})$. 
Test statistics and $p$-value computation

- Imputation approaches treating censored observation as partially missing data have become very popular in recent years (Lapidus et al., Stat. Med., 2014).
- In this work, we propose to use an imputation procedure to replace the censored $\hat{q}_i$’s by imputed values in order to obtain a completed vector of residuals, whose joint distribution can be easily approximated.
- Without loss of generality, we rearrange the indices $\{1, \cdots, n\}$ so that $\delta_1 = \cdots = \delta_{n_1} = 1$ and $\delta_{n_1+1} = \cdots = \delta_n = 0$, where $n_1 = \sum_{i=1}^{n} \delta_i$ and consider the partition $\hat{q} = (\hat{q}^{(1)}, \hat{q}^{(0)})$.
- Similarly, write

$$\hat{\Gamma} = \hat{h}^2 \varphi + (1 - \hat{h}^2)I_n = \begin{pmatrix} \hat{\Gamma}^{(11)} & \hat{\Gamma}^{(10)} \\ \hat{\Gamma}^{(01)} & \hat{\Gamma}^{(00)} \end{pmatrix}.$$
Test statistics and $p$-value computation

A completed vector of residuals has the form

$$r = \begin{pmatrix} \hat{q}^{(1)} \\ \tilde{q}^{(0)} \end{pmatrix},$$

where the vector of imputed values $\tilde{q}^{(0)}$ is a random draw from the truncated multivariate normal distribution with mean

$$\hat{\Gamma}^{(01)}\hat{\Gamma}^{(11)^{-1}}\hat{q}^{(1)},$$

covariance matrix

$$\hat{\Gamma}^{(00)} - \hat{\Gamma}^{(01)}\hat{\Gamma}^{(11)^{-1}}\hat{\Gamma}^{(10)},$$

and support

$$[\hat{q}_{n1}+1, \infty] \times \cdots \times [\hat{q}_n, \infty].$$
Test statistics and $p$-value computation

- Typically, imputation procedures require the generation of multiple completed data sets.
Test statistics and $p$-value computation

- Typically, imputation procedures require the generation of multiple completed data sets.
- Separate analysis are performed on each completed data set and the results are aggregated afterwards using the methodology of Rubin (Multiple imputation for nonresponse in surveys, New York: Wiley 1987).
Test statistics and $p$-value computation

- Typically, imputation procedures require the generation of multiple completed data sets.
- Separate analysis are performed on each completed data set and the results are aggregated afterwards using the methodology of Rubin (*Multiple imputation for nonresponse in surveys*, New York: Wiley 1987).
- However, this strategy is mainly useful when the quantity of interest (e.g. parameter estimator or test statistics) follows a normal distribution.
Test statistics and $p$-value computation

- Typically, imputation procedures require the generation of multiple completed data sets.
- Separate analysis are performed on each completed data set and the results are aggregated afterwards using the methodology of Rubin (Multiple imputation for nonresponse in surveys, New York: Wiley 1987).
- However, this strategy is mainly useful when the quantity of interest (e.g. parameter estimator or test statistics) follows a normal distribution.
- This is not the case here as the considered test statistics is a quadratic form following a mixture of chi-squared variables.
Test statistics and $p$-value computation

- Typically, imputation procedures require the generation of multiple completed data sets.

- Separate analysis are performed on each completed data set and the results are aggregated afterwards using the methodology of Rubin (Multiple imputation for nonresponse in surveys, New York: Wiley 1987).

- However, this strategy is mainly useful when the quantity of interest (e.g. parameter estimator or test statistics) follows a normal distribution.

- This is not the case here as the considered test statistics is a quadratic form following a mixture of chi-squared variables.

- In this work, we consider an alternative approach.
Test statistics and $p$-value computation

- We generate $M$ completed vectors of residuals and compute their mean, which we denote $r^{(M)}$. 
Test statistics and $p$-value computation

- We generate $M$ completed vectors of residuals and compute their mean, which we denote $r^{(M)}$.

- We show that $r^{(M)}$ follows approximately a multivariate normal distribution with mean zero and covariance matrix $\rho^{(M)}\hat{\Gamma}$. 
We generate $M$ completed vectors of residuals and compute their mean, which we denote $r^{(M)}$.

We show that $r^{(M)}$ follows approximately a multivariate normal distribution with mean zero and covariance matrix $\rho^{(M)}\hat{\Gamma}$.

$\rho^{(M)}$ is a scale parameter that reflects the fact that we are using multiple imputations rather than real observations.
We generate $M$ completed vectors of residuals and compute their mean, which we denote $r^{(M)}$.

We show that $r^{(M)}$ follows approximately a multivariate normal distribution with mean zero and covariance matrix $\rho^{(M)}\hat{\Gamma}$.

$\rho^{(M)}$ is a scale parameter that reflects the fact that we are using multiple imputations rather than real observations.

This parameter is estimated by its maximum likelihood estimator $\hat{\rho}^{(M)} = r^{(M)}\top \hat{\Gamma} r^{(M)} / n$. 
The test statistics is then $Q = r^{(M)\top} K r^{(M)}$. 
The test statistics is then $Q = r^{(M)^\top}Kr^{(M)}$.

It is distributed as

$$r^{(M)^\top}Kr^{(M)} \sim \sum_{l=1}^{L} \mu_l \chi^2_{l,1},$$

where $0 < \mu_1 < \cdots < \mu_L$ are the $L$ positive eigenvalues of $\hat{\rho}^{(M)}\hat{\Gamma}^{1/2}K\hat{\Gamma}^{1/2}$. 
The test statistics is then $Q = r^{(M)\top} K r^{(M)}$.

It is distributed as

$$r^{(M)\top} K r^{(M)} \sim \sum_{l=1}^{L} \mu_l \chi^2_{l,1},$$

where $0 < \mu_1 < \cdots < \mu_L$ are the $L$ positive eigenvalues of $\hat{\rho}^{(M)} \hat{\Gamma}^{1/2} K \hat{\Gamma}^{1/2}$.

The Davies (Davies, JRSS Series C 1980) method is then employed to obtain a $p$-value for the test.
Summary

Estimate $H$ and $\xi$ by the algorithm of Chen et al. (Biometrika, 2002).
Summary

1. Estimate $H$ and $\xi$ by the algorithm of Chen et al. (Biometrika, 2002).

2. Estimate $h^2$ by the algorithm of Othus and Li (Stat. Biosc., 2010) and compute $\hat{\Gamma} = \hat{h}^2 \varphi + (1 - \hat{h}^2)I_n$. 
Summary

1. Estimate $H$ and $\xi$ by the algorithm of Chen et al. (Biometrika, 2002).

2. Estimate $h^2$ by the algorithm of Othus and Li (Stat. Biosc., 2010) and compute $\hat{\Gamma} = \hat{h}^2 \varphi + (1 - \hat{h}^2)I_n$.

3. Rearrange the indices $\{1, \cdots, n\}$ and deduce the partitioned expressions of $\hat{q}$ and $\hat{\Gamma}$. 
Summary

1. Estimate $H$ and $\xi$ by the algorithm of Chen et al. (Biometrika, 2002).

2. Estimate $h^2$ by the algorithm of Othus and Li (Stat. Biosc., 2010) and compute $\hat{\Gamma} = \hat{h}^2 \varphi + (1 - \hat{h}^2)I_n$.

3. Rearrange the indices $\{1, \cdots, n\}$ and deduce the partitioned expressions of $\hat{q}$ and $\hat{\Gamma}$.

4. Generate $M$ completed vector of residuals and compute their mean $r^{(M)}$ and the test statistics $Q = r^{(M)^\top} Kr^{(M)}$. 
Summary

1. Estimate $H$ and $\xi$ by the algorithm of Chen et al. (Biometrika, 2002).

2. Estimate $h^2$ by the algorithm of Othus and Li (Stat. Biosc., 2010) and compute $\hat{\Gamma} = \hat{h}^2 \varphi + (1 - \hat{h}^2)I_n$.

3. Rearrange the indices $\{1, \cdots, n\}$ and deduce the partitioned expressions of $\hat{q}$ and $\hat{\Gamma}$.

4. Generate $M$ completed vector of residuals and compute their mean $r^{(M)}$ and the test statistics $Q = r^{(M)\top}Kr^{(M)}$.

5. Compute $\hat{\rho}^{(M)} = r^{(M)\top}\hat{\Gamma}r^{(M)}/n$ and deduce the positive eigenvalues of $\hat{\rho}^{(M)}\hat{\Gamma}^{1/2}K\hat{\Gamma}^{1/2}$. 
Summary

1. Estimate $H$ and $\xi$ by the algorithm of Chen et al. (Biometrika, 2002).

2. Estimate $h^2$ by the algorithm of Othus and Li (Stat. Biosc., 2010) and compute $\hat{\Gamma} = \hat{h}^2 \varphi + (1 - \hat{h}^2)I_n$.

3. Rearrange the indices $\{1, \cdots, n\}$ and deduce the partitioned expressions of $\hat{q}$ and $\hat{\Gamma}$.

4. Generate $M$ completed vector of residuals and compute their mean $r^{(M)}$ and the test statistics $Q = r^{(M)^\top} Kr^{(M)}$.

5. Compute $\hat{\rho}^{(M)} = r^{(M)^\top} \hat{\Gamma} r^{(M)}/n$ and deduce the positive eigenvalues of $\hat{\rho}^{(M)} \hat{\Gamma}^{1/2} K \hat{\Gamma}^{1/2}$.

6. Obtain the $p$-value by the Davies approximation.
Simulations: setting

- Simulations were carried out to evaluate the empirical properties of the proposed genomic region-based association test and to compare it to existing approaches.
Simulations: setting

- Simulations were carried out to evaluate the empirical properties of the proposed genomic region-based association test and to compare it to existing approaches.
- Samples of $n = 600$ individuals from 120 families were generated:
Simulations: setting

- Simulations were carried out to evaluate the empirical properties of the proposed genomic region-based association test and to compare it to existing approaches.
- Samples of $n = 600$ individuals from 120 families were generated:
  - 40 families of two parents and one child.
Simulations: setting

Simulations were carried out to evaluate the empirical properties of the proposed genomic region-based association test and to compare it to existing approaches.

Samples of \( n = 600 \) individuals from 120 families were generated:

- 40 families of two parents and one child.
- 40 families of two parents and two children.
Simulations: setting

Simulations were carried out to evaluate the empirical properties of the proposed genomic region-based association test and to compare it to existing approaches.

Samples of $n = 600$ individuals from 120 families were generated:

- 40 families of two parents and one child.
- 40 families of two parents and two children.
- 40 families of 3 generations: two grand parents, four parents and 2 grand children.
Simulations: setting

- Simulations were carried out to evaluate the empirical properties of the proposed genomic region-based association test and to compare it to existing approaches.
- Samples of $n = 600$ individuals from 120 families were generated:
  - 40 families of two parents and one child.
  - 40 families of two parents and two children.
  - 40 families of 3 generations: two grand parents, four parents and 2 grand children.
- We set $s = 10$ and $q = 2$.
Simulations: setting

• Simulations were carried out to evaluate the empirical properties of the proposed genomic region-based association test and to compare it to existing approaches.
• Samples of \( n = 600 \) individuals from 120 families were generated:
  • 40 families of two parents and one child.
  • 40 families of two parents and two children.
  • 40 families of 3 generations: two grand parents, four parents and 2 grand children.
• We set \( s = 10 \) and \( q = 2 \)
• Simulations were ran under \( H_0 \) and \( H_1 \), in both presence and absence of familial dependence.
Simulations: setting

- Simulations were carried out to evaluate the empirical properties of the proposed genomic region-based association test and to compare it to existing approaches.

- Samples of $n = 600$ individuals from 120 families were generated:
  - 40 families of two parents and one child.
  - 40 families of two parents and two children.
  - 40 families of 3 generations: two grand parents, four parents and 2 grand children.

- We set $s = 10$ and $q = 2$

- Simulations were ran under $H_0$ and $H_1$, in both presence and absence of familial dependence.

- We ran simulations with the Gaussian copula either correctly specified or misspecified.
Simulations: results

- Under $H_0$, the distribution of $p$-values of the proposed method matches very well the expected values in all scenarios.
Simulations: results

- Under $H_0$, the distribution of $p$-values of the proposed method matches very well the expected values in all scenarios.

- Under $H_0$, both existing methods have significant inflated type I error in the presence of familial dependence.
Simulations: results

- Under $H_0$, the distribution of $p$-values of the proposed method matches very well the expected values in all scenarios.

- Under $H_0$, both existing methods have significant inflated type I error in the presence of familial dependence.

- Under $H_1$, the power of the three methods are comparable.
Simulations: results

- Under $H_0$, the distribution of $p$-values of the proposed method matches very well the expected values in all scenarios.

- Under $H_0$, both existing methods have significant inflated type I error in the presence of familial dependence.

- Under $H_1$, the power of the three methods are comparable.

- Under $H_1$, the power of the proposed test increases with $M$, until reaching a plateau.
Simulations: results

- Under $H_0$, the distribution of $p$-values of the proposed method matches very well the expected values in all scenarios.

- Under $H_0$, both existing methods have significant inflated type I error in the presence of familial dependence.

- Under $H_1$, the power of the three methods are comparable.

- Under $H_1$, the power of the proposed test increases with $M$, until reaching a plateau.

- The proposed method still performs well in both control of type I error and power when the Gaussian copula is misspecified.
Breast cancer data

The Consortium of Investigators of Modifiers of BRCA1 and BRCA2 (CIMBA) aims to identify genetic factors associated with breast cancer risk in BRCA1 and BRCA2 mutation carriers.
Breast cancer data

The Consortium of Investigators of Modifiers of BRCA1 and BRCA2 (CIMBA) aims to identify genetic factors associated with breast cancer risk in BRCA1 and BRCA2 mutation carriers.

A sample of 13,465 women of European ancestry who carry a mutation in the BRCA1 gene.
The Consortium of Investigators of Modifiers of BRCA1 and BRCA2 (CIMBA) aims to identify genetic factors associated with breast cancer risk in BRCA1 and BRCA2 mutation carriers.

A sample of 13,465 women of European ancestry who carry a mutation in the BRCA1 gene.

25 countries.
Breast cancer data

- The Consortium of Investigators of Modifiers of BRCA1 and BRCA2 (CIMBA) aims to identify genetic factors associated with breast cancer risk in BRCA1 and BRCA2 mutation carriers.

- A sample of 13,465 women of European ancestry who carry a mutation in the BRCA1 gene.

- 25 countries.

- 9,544 clusters.
Breast cancer data

- The Consortium of Investigators of Modifiers of BRCA1 and BRCA2 (CIMBA) aims to identify genetic factors associated with breast cancer risk in BRCA1 and BRCA2 mutation carriers.
- A sample of 13,465 women of European ancestry who carry a mutation in the BRCA1 gene.
- 25 countries.
- 9,544 clusters.
- 24% of the clusters have a size greater than 1.
Breast cancer data

- The Consortium of Investigators of Modifiers of BRCA1 and BRCA2 (CIMBA) aims to identify genetic factors associated with breast cancer risk in BRCA1 and BRCA2 mutation carriers.

- A sample of 13,465 women of European ancestry who carry a mutation in the BRCA1 gene.

- 25 countries.

- 9,544 clusters.

- 24% of the clusters have a size greater than 1.

- Among these, the family size varies between 2 and 52, for a total of 6,250 subjects (46% of the sample).
Breast cancer data

- The SNP set analyzed comprises 111 variants across the TERT locus at 5p15.33 (positions 1,280,693-1,414,669).
Breast cancer data

- The SNP set analyzed comprises 111 variants across the \textit{TERT} locus at 5p15.33 (positions 1,280,693-1,414,669).
- The MAF of the SNPs ranges from 0.0134 to 0.5.
Breast cancer data

- The SNP set analyzed comprises 111 variants across the *TERT* locus at 5p15.33 (positions 1,280,693-1,414,669).

- The MAF of the SNPs ranges from 0.0134 to 0.5.

- The kinship coefficient for each pair of individuals was estimated using the available genotype data.
Breast cancer data

- The SNP set analyzed comprises 111 variants across the *TERT* locus at 5p15.33 (positions 1,280,693-1,414,669).
- The MAF of the SNPs ranges from 0.0134 to 0.5.
- The kinship coefficient for each pair of individuals was estimated using the available genotype data.
- The phenotype is the age at breast cancer diagnostic or age at last follow-up.
Breast cancer data

- The SNP set analyzed comprises 111 variants across the TERT locus at 5p15.33 (positions 1,280,693-1,414,669).
- The MAF of the SNPs ranges from 0.0134 to 0.5.
- The kinship coefficient for each pair of individuals was estimated using the available genotype data.
- The phenotype is the age at breast cancer diagnostic or age at last follow-up.
- The observation is censored if cancer wasn’t observed by last follow-up.
Breast cancer data

- The SNP set analyzed comprises 111 variants across the \textit{TERT} locus at 5p15.33 (positions 1,280,693-1,414,669).

- The MAF of the SNPs ranges from 0.0134 to 0.5.

- The kinship coefficient for each pair of individuals was estimated using the available genotype data.

- The phenotype is the age at breast cancer diagnostic or age at last follow-up.

- The observation is censored if cancer wasn’t observed by last follow-up.

- The censoring rate is equal to 0.49%.
Breast cancer data

- We obtained $\hat{h}^2 = 0.566$, which confirms the breast cancer high polygenic heritability.
Breast cancer data

- We obtained $\hat{h}^2 = 0.566$, which confirms the breast cancer high polygenic heritability.

- We applied the proposed kinship-adjusted association test with $M = 50$. We obtained a $p$-value equal to $2.61 \times 10^{-4}$, which indicates an evidence of association between $TERT$ and breast cancer risk.
Breast cancer data

- We obtained $\hat{h}^2 = 0.566$, which confirms the breast cancer high polygenic heritability.

- We applied the proposed kinship-adjusted association test with $M = 50$. We obtained a $p$-value equal to $2.61 \times 10^{-4}$, which indicates an evidence of association between $TERT$ and breast cancer risk.

- Such evidence is weaker for the approaches of Lin et al. (AJHG, 2011) and Chen et al. (AJHG, 2014), for which we obtained $p$-values equal to $3.18 \times 10^{-2}$ and $2.96 \times 10^{-2}$, respectively.
Breast cancer data

- We obtained $\hat{h}^2 = 0.566$, which confirms the breast cancer high polygenic heritability.

- We applied the proposed kinship-adjusted association test with $M = 50$. We obtained a $p$-value equal to $2.61 \times 10^{-4}$, which indicates an evidence of association between TERT and breast cancer risk.

- Such evidence is weaker for the approaches of Lin et al. (AJHG, 2011) and Chen et al. (AJHG, 2014), for which we obtained $p$-values equal to $3.18 \times 10^{-2}$ and $2.96 \times 10^{-2}$, respectively.

- We also considered a sliding window of 25 SNPs, with 15 SNPs each overlapping with the previous and subsequent windows, except for the last window, which contained 21 SNPs.
Breast cancer data

<table>
<thead>
<tr>
<th>SNPs</th>
<th>Lin et al. (2011)</th>
<th>Chen et al. (2014)</th>
<th>Imputed residuals</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–25</td>
<td>8.70 × 10⁻³</td>
<td>8.00 × 10⁻³</td>
<td>8.87 × 10⁻⁴*</td>
</tr>
<tr>
<td>11–35</td>
<td>7.13 × 10⁻²</td>
<td>6.77 × 10⁻²</td>
<td>4.46 × 10⁻⁴*</td>
</tr>
<tr>
<td>21–45</td>
<td>1.35 × 10⁻¹</td>
<td>1.41 × 10⁻¹</td>
<td>5.59 × 10⁻⁴*</td>
</tr>
<tr>
<td>31–55</td>
<td>4.90 × 10⁻²</td>
<td>4.38 × 10⁻²</td>
<td>9.21 × 10⁻³</td>
</tr>
<tr>
<td>41–65</td>
<td>1.11 × 10⁻¹</td>
<td>1.10 × 10⁻¹</td>
<td>1.82 × 10⁻¹</td>
</tr>
<tr>
<td>51–75</td>
<td>2.85 × 10⁻¹</td>
<td>2.91 × 10⁻¹</td>
<td>7.32 × 10⁻³</td>
</tr>
<tr>
<td>61–85</td>
<td>3.93 × 10⁻¹</td>
<td>3.96 × 10⁻¹</td>
<td>9.51 × 10⁻³</td>
</tr>
<tr>
<td>71–95</td>
<td>2.62 × 10⁻¹</td>
<td>2.64 × 10⁻¹</td>
<td>7.00 × 10⁻³</td>
</tr>
<tr>
<td>81–105</td>
<td>1.85 × 10⁻¹</td>
<td>1.84 × 10⁻¹</td>
<td>1.11 × 10⁻²</td>
</tr>
<tr>
<td>91–111</td>
<td>1.83 × 10⁻¹</td>
<td>1.80 × 10⁻¹</td>
<td>4.22 × 10⁻²</td>
</tr>
</tbody>
</table>

0.05/10 = 5 × 10⁻³
Conclusion

- We have developed an association test for censored traits in family-based designs.
Conclusion

- We have developed an association test for censored traits in family-based designs.

- The proposed test is based on a multiple imputation approach.
We have developed an association test for censored traits in family-based designs.

The proposed test is based on a multiple imputation approach.

The proposed test can be extended to any kernel matrix $K$. 
We have developed an association test for censored traits in family-based designs.

The proposed test is based on a multiple imputation approach.

The proposed test can be extended to any kernel matrix $K$.

The proposed test is suitable for genome wise association studies since the vector $r^{(M)}$ has to be generated only once.
Acknowledgements

Collaborators: Martin Leclerc, Jacques Simard, Antonis Antoniou.
Acknowledgements

- Collaborators: Martin Leclerc, Jacques Simard, Antonis Antoniou.
- Special thanks to the CIMBA consortium for providing the breast cancer data.
Acknowledgements

- Collaborators: Martin Leclerc, Jacques Simard, Antonis Antoniou.
- Special thanks to the CIMBA consortium for providing the breast cancer data.

Thank you for your attention
Acknowledgements

- Collaborators: Martin Leclerc, Jacques Simard, Antonis Antoniou.
- Special thanks to the CIMBA consortium for providing the breast cancer data.

Thank you for your attention

Any Questions?