Heritability Estimation of Dichotomous Phenotypes Using Liability Threshold Model

Wonji Kim1, Sungho Won1,2\*

1 Interdisciplinary program in Bioinformatics, Seoul National University, Korea

2Department of Public Health Science, Seoul National University, Korea

\*To whom correspondence should be addressed.

**Abstract**

Numerous methods for estimating the heritability have been proposed. However unlike quantitative phenotypes, parameter estimation for dichotomous phenotypes suffers from computational and statistical complexity, and heritability has been limitedly utilized. In this article, we developed statistical method to estimate heritability of dichotomous phenotypes with the liability threshold model for ascertained samples. Liability threshold model assumes dichotomous phenotypes are determined with unobserved latent variable which are normally distributed, and it can be applied to the general pedigree data. The proposed methods were applied to simulated data, and accuracy of estimates by the proposed methods was compared with other methods.

**Keywords**

Heritability, Liability threshold model, ascertainment bias

**Introduction**

Phenotypes are affected by both environmental factors and genes, and individuals are expected to be similar, because of their genetic similarity. Heritability was defined to quantify such phenotypic similarity attributable to heritable components, and has been widely used to understand the genetic architecture of phenotypes. For instance, heritability can be used to compare the importance of genetic components among different phenotypes. Furthermore if large-scale genetic data are available, genetic correlation matrix can be estimated, and its incorporation to linear mixed model enabled the SNP heritability estimation. SNP heritability indicates the relative proportion of variance attributable to the genotyped SNPs, and they were used to identify the amount of the missing heritability.

Estimation of broad-sense heritability requires bilinear relatives such as sibling or monozygotic twins, and in practice, narrow-sense heritability has been often utilized. Narrow-sense heritability is defined as the proportion of the total phenotypic variation explained by additive genetic effects [1]. Various methods have been developed for estimating the heritability of the continuous traits. For instance, restricted maximum likelihood (REML) method based on the linear mixed model (LMM) [2-4] or polygenic score method [5] can be used for estimating the heritability of the continuous traits. For the dichotomous traits, generalized linear mixed models (GLMM) or liability threshold model (LTM) have been often utilized [6, 7]. The liability threshold model assumes there are unobserved continuous liability scores, and subjects are affected if they exceed certain threshold [8-11].

There are multiple factors which can bias variance estimation for dichotomous traits. In particular, family-based samples are usually ascertained with probands. Probands means family member through whom a family's members comes to study. Multiple literature demonstrated that such ascertainment can produce substantial bias of variance estimates [3]. For instance, if phenotypes are rare and families are randomly selected, the number of affected individuals is often very small. Therefore families are ascertained by using affected probands. In such scenario, most of their relative may be unaffected, and unless family size is large, negative correlation can be observed because probands are affected and their relatives are unaffected. Several approaches have been proposed to adjust such ascertainment bias. GCTA adjusts estimated heritabilities by assuming that the level of ascertainment bias is same among individuals [3]. However, families are ascertained with probands and the effect of ascertainment bias is heterogeneous according to their familiar relationship [12]. For instance, ascertainment bias for grandparents of the proband is expected to be around a half compared to parents.

In this article, we developed a new method to estimate heritability based on the liability threshold model for binary traits which can be applied to the extended pedigree structure. Using Expectation-Maximization (EM) algorithm, the proposed method jointly estimates maximum likelihood estimators for heritability and coefficients of covariates [13]. Furthermore the proposed method maximizes the conditional likelihood conditioned on disease statuses of probands via conditional EM (CEM) algorithm [14], and ascertainment bias can be adjusted. We also developed the conditional expected score test (CEST) to test whether heritability is equal to zero. Extensive simulation studies showed that heritability estimates from the proposed methods are generally unbiased even for the ascertained samples. Estimates from GCTA are unbiased for no ascertained families, but otherwise the bias becomes substantial. Also we found that CEST for the heritability was statistically conservative, but achieved reasonable statistical power estimates. Last we applied the proposed method to estimate the heritability of type-2 diabetes with ascertained family-based samples of Korea and those estimates were shown to illustrate the practical value of the proposed methods.

**Methods and Materials**

**Notations and Disease Model**

We assume that there are *n* independent families and family *i* has family members (). We consider the liability threshold model, and dichotomous phenotypes are determined by the unobserved continuous liability score. The liability score of subject *j* in family *i* is denoted by and they are determined by summing the environmental/genetic effects, polygenic effect and random error. The covariates including environmental/genetic effects for subject *j* in family *i* is denoted by andwe assumed that covariates are standardized. In this article, we assumed there are *p* covariates. The random effects including polygenic effect and random error for subject *j* in family *i* are denoted by . The vector form of each component for family *i* are denoted by

and .

Liability scores of family members are often correlated, and we assumed that liability scores are normally distributed as follows:

where . We let be the kinship coefficient matrix multiplied by two, and be the dimensional identity matrix. Under the polygenic model with additivity of genetic effects across loci and linkage equilibrium among loci, we can get

where , and are the variances of additive, dominance and environmental effects in the population, and and are the dominance genetic variance and the covariance of additive and dominance effects in the homozygous population respectively [15-17]. , and are the functions of the condensed coefficients of identity [17]. For simplicity, we assume that all variance components other than and are zero, and the sum of and is equal to one. If we denote the heritability as , then the variance-covariance matrix of is expressed by

The dichotomous phenotypes for subject *j* in family *i* is denoted by and they are coded as 1 for cases and 0 for controls. In a liability threshold model, is determined by ; if is above a certain threshold value *c*, becomes 1, and otherwise it becomes 0. *c* can be determined from the prevalence of the diseases, and the phenotype vector for family *i* is denoted by

.

Cumulative distribution function of standard normal was used to obtain *c* with prevalence, *q*, known a *priori*. Therefore, for each observed , we can infer the range of the liability, , which are the lower and upper bound of respectively. For instance, if is equal to zero, then is bounded in , and otherwise, is bounded in . The lower and upper bounds of the liability for the family *i* are denoted by

and **.**

Based on above notations, all subjects can be expressed in the following vector forms:

and.

Under those notations, we assumed that **L** follows multivariate normal distribution with mean and variance-covariance matrix which are block diagonal matrix consisting of .

## Estimation of the Heritability using the EM Algorithm

The EM (Expectation-Maximization) algorithm [13] was used to estimate based on the complete data consisting of observed phenotypes, **Y**, and unobserved liabilities, **L**. The joint probability density function (pdf) of the complete data can be decomposed into the marginal pdf of **L** and the conditional pdf of **Y** given **L** which has the support of (**a, b**). It can be formulated as

.

If we denote the parameters of interest as , and the log-likelihood of the complete data will be the sum of log-likelihoods for each family, as follows,

In the E-step of the EM algorithm, the conditional expectation of **L** given **Y** was taken to the , where the estimates for the parameters of the previous iteration were used. If we assumed that the *k*th iteration has been performed and defined the estimates for the parameters at the *k*th iteration as , then the conditional expectation will be

and

where and .

In M-step of EM algorithm, we maximize with respect to. Since is the concave function, we can find the maximizer by solving the equation that the first derivative of is equal to **0**. The partial derivative with respect to is

and, which satisfies will be

To emphasize that the root is the function of , it was denoted by . Unfortunately, there is no closed form of the root which . In this process, we need a numerical algorithm to find root and Newton-Raphson algorithm [18] was utilized here. The detailed algorithm is provided in the Appendix (A).

After we obtained the maximizer of in maximization step, we updated to and repeat EM steps until convergence. Note that is the unbiased estimator of and it can be easily proven by using the low of total expectation:

assuming we got after *m* iterations [19].

**Lagrangian Multiplier and Karush-Kuhn-Tucker Condition**

Unlike , the parameter space of is restricted to . Thus, should be estimated in with the following objective function at the *k*th iteration

subject to .

We can optimize the above objective function via the method of Lagrange multiplier [20] under the Karush-Kuhn-Trucker (KKT) conditions [21]. Changing the constraint to and , the Lagrangian, , is denoted by

where . We can find the solution that maximizes subject to by finding and satisfying the following three conditions known as KKT conditions. The first KKT condition known as *Stationarity* is

Since does not depend on the constraint, is identical to , providing that . Replacing to , we get

and it is equivalence to

Note that the left hand side is the function of , denoted by . The second condition which is called as *Complementary slackness* is and , and is equivalent to

.

Since and , cannot be satisfied simultaneously, will be , or . Of them, and are related to the last condition, *Dual feasibility*, which is for . If we assume and , then and it will be non-positive if the assumptions are met by the *Dual feasibility* condition. Similarly, when and are assumed, and it will be non-negative if the assumptions are satisfied. If none of them are met, and are automatically zero, and the problem becomes the optimization with no restrictions on since the constraints have no meaning. The concept is illustrated in Figure 1.

**Robust Estimator to the Misspecified Prevalence**

We showed that is the unbiased estimator regardless the choices of the heritability and the prevalence. On the other hand, the estimator for the heritability is biased and the prevalence is involved in   in such a complicated manner. In our optimization procedure, the only part where the prevalence is involved in is the E-step of EM algorithm, which takes the conditional expectation to the liabilities. Performing optimization by setting the prevalence as a parameter is too complex in terms of calculating the conditional expectation of multivariate truncated normal distribution. Alternatively, if the true prevalence is lower than the assumed prevalence, it can be corrected by shifting liabilities to the right as a whole and it is achievable quite simply by including a positive intercept. This concept is equally applicable in the opposite situation. Therefore, inclusion of the intercept in the parameters can have the effect of correcting the misspecified prevalence. We denoted the parameter for the intercept as .

If we let and where is the dimensional vector which elements are all one, we will get the estimates for and at the *k*th iteration of the EM algorithm as follows,

where , , and =.

**Estimates of Parameters Adjusted for Ascertainment Bias**

In the notation, we added superscripts *P* and *NP* for proband and non-proband respectively in order to distinguish them. If we assume there are non-probands in family *i*, then liabilities, covariates, phenotypes and bounds for a liability for non-probands in family *i* are denoted by

and

respectively. Assuming that there is a proband in each family, a liability, covariates, a phenotype and bounds for a liability for proband in family *i* are defined as , ,, and , respectively. Finally, the liabilities for proband, non-proband and all subjects across entire families are defined as

and ,

and other variables are also defined in the same way.

The likelihood is conditioned on the phenotypes of probands as follows,

and the log-likelihood will be . The objective function of the EM algorithm is a global lower bound for the log-likelihood [22] and it can be obtained by using the lower bound for and the upper bound for given by their relationship,

.

At , we can achieve by considering the EM algorithm for the joint log-likelihood . That is,

where is the entropy. The upper bound for can be defined as [14]. Therefore, the global lower bound of the log-likelihood is achieved by

at . Since probands are independent each other, the log-likelihood for proband is simply given by the sum of log-likelihoods for Bernoulli distribution with the mean which is the function of for each family:

where .

Based on the liability threshold model, is formulated as the function of the cumulative distribution function of the standard normal, , as follows,

The maximum likelihood estimates for parameters are obtained by iteratively maximizing the objective function with respect to until convergence. The detailed algorithm for maximization is provided in Appendix (B).

**Conditional Expected Score Tests**

Since and are the parameters on the liability scale, statistical significance of the estimators cannot be directly tested with the score test based on the observed data, **Y**. Alternatively, the score based on **Y** can be derived using the fact that the Fisher score based on the incomplete data is equal to the conditional expectation of the score based on the complete data given the observed data [13, 23] which can be formulated in our cases as follows,

The score test with the above statistics is also called Conditional Expected Score Test (CEST) in that it takes the conditional expectation to the score based on the complete data [24]. For simplicity, we assumed that the prevalence is correctly specified and samples are randomly selected.

The conditional expected score (CES) based on the complete data for family *i* is

where , and . Note that and are also the function of . The CES statistics can be calculated by adding up for entire families, that is, . If we let and denote and respectively, then the score statistics are

, , , and .

According to the standard likelihood theory [25], the variance-covariance matrix of is approximately calculated based on the empirical Fisher information matrix which is given by

and it could be expressed as

.

Then the variances for and, and ,will beapproximately equal to and respectively.

We let denote the maximum likelihood estimates of under the null hypothesis. For the hypothesis test of , the test statistics is asymptotically distributed as where *p* is the number of parameters to be tested. Whereas, for the test of , the likelihood will be maximized at in case of   with about 50% probability and on the parameter space for with about 50% probability under . Hence, the asymptotic distribution of the test statistics which is calculated by is given by

where . Therefore, the asymptotic distribution of is the mixture of and the distribution with a probability of 1 at zero, with 50% probability for each [26] and whenever , P-value will be half of itself.

**Simulation studies**

We generated 50,000 families with three, four, five and six members consisting of a parents and offsprings at the rates of 0.2, 0.3, 0.3 and 0.2 respectively. Liability was assumed to be affected by a main genetic factor with minor allele frequency 0.2 under the Hardy Weinberg Equilibrium, and no environmental effect. Founders’ genotypes in each family were generated from binomial distribution with the number of trials of 2 with the probability for success of 0.2, and the non-founders’ genotypes were obtained by randomly generated by the rule of Mendelian transmissions. The random effects including the polygenic effects and random errors were generated from the multivariate normal distribution with heritability 0.2 and 0.4. Finally, liability is calculated by summing the product of and the number of minor alleles and random effects. Here is obtained by the assumed the relative proportion of variance explained by the SNP (), and minor allele frequency (), with the following equation:

Here, was assumed to be 0.005 and corresponded to 0.1253. Once liabilities were generated, they were transformed into affected if larger than the threshold *c*, and otherwise were considered unaffected. The value of *c* was chosen to preserve the assumed prevalences of *q* = 0.1 or *q* = 0.2.

For each setting of and *q*, simulation was repeated 300 times with 500 families randomly selected among 10,000 families. Simulation studies were conducted according to the following scheme. First, with correctly specified prevalence, we estimated and using the basic estimation method. Second, assuming that the prevalence was 0.1 larger than true value, and were estimated with and without the tuning parameter .

**CRC Dataset**

**Results**

**Results of Simulation Studies**

For the first scheme of the simulation study which assumes prevalence correctly, some descriptive statistics of and are shown in Table 1. The estimates appear to be slightly different from true values but the bias can be reduced by increasing the family size. The estimates are concentrated around the true value even though they seem to be skewed slightly (Figure 2).

When the prevalence is misspecified, the basic estimation method provided seriously biased estimates for (Figure 3A). Interestingly, when the true prevalences were 0.1 and 0.2, was overestimated to be about 0.4 and 0.2 larger than the true respectively. According to the results, the basic estimation method seems sensitive to the specification of the prevalence and bias depends on not only the difference between true prevalence and assumed prevalence, but the true prevalence itself. Overestimation of is not surprising because observing a lower proportion of affected subjects than the assumed prevalence implies a high correlation between subjects. This phenomenon would appear to be stronger in rare disease. Whereas, was estimated to be relatively stable compared to but somewhat underestimated (Figure 3B).

In Figure 4, we displayed the distributions of estimates using the robust estimation method when the prevalence is misspecified. The bias in based on the basic estimation method seems to be drastically reduced (Figure 4A) and s are well distributed around the true value (Figure 4B). The tuning parameter, , was estimated to have negative values, which reduces the prevalence by shifting the liabilities to the right (Figure 4C).

**References**

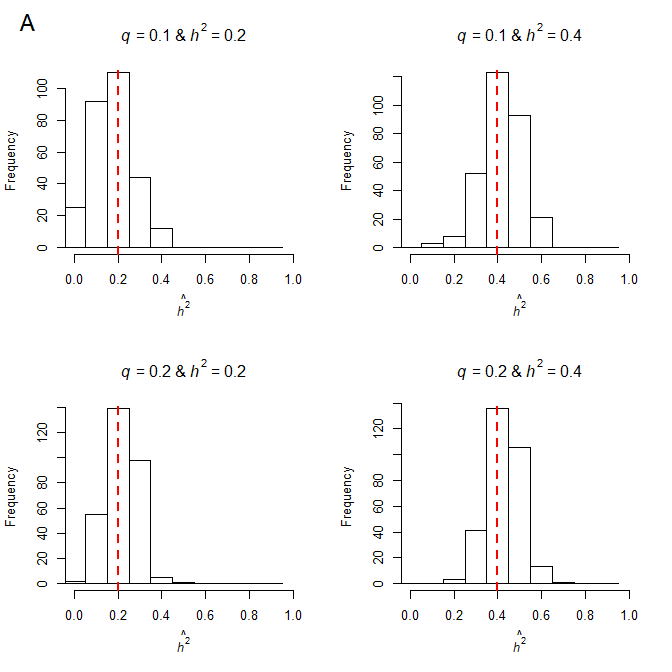
**Table 1. The Descriptive statistics for and**  **based on 300 replicates when the prevalence was correctly specified**. The estimates from 300 replicates were summarized using some descriptive statistics; Mean : observed mean of the 300 estimates, Median : observed median of the 300 estimates, SD : observed standard deviation of the 300 estimates. The true value for is 0.1253.

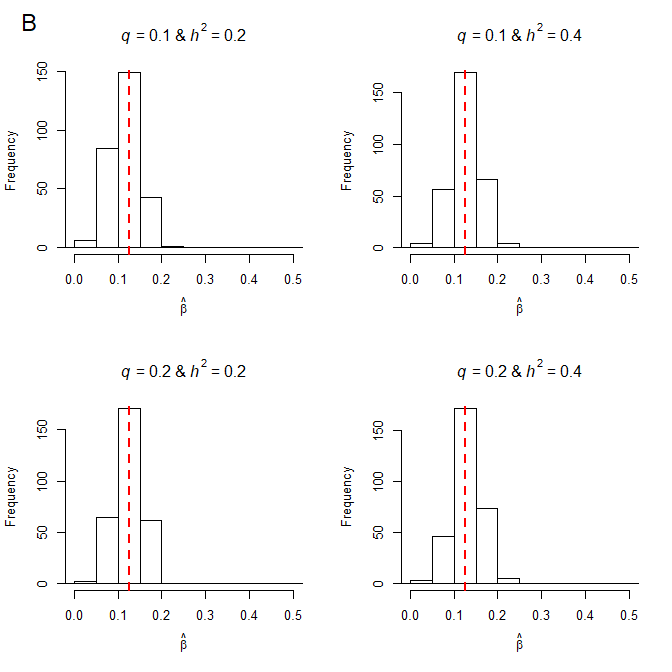
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Prevalence | Heritability | Parameter | Mean | Median | SD |
| 0.1 | 0.2 |  | 0.1146 | 0.1127 | 0.0333 |
|  | 0.1757 | 0.1684 | 0.0917 |
| 0.4 |  | 0.1244 | 0.1241 | 0.0322 |
|  | 0.4221 | 0.4232 | 0.0960 |
| 0.2 | 0.2 |  | 0.1230 | 0.1208 | 0.0300 |
|  | 0.2170 | 0.2183 | 0.0725 |
| 0.4 |  | 0.1311 | 0.1323 | 0.0317 |
|  | 0.4285 | 0.4335 | 0.0774 |

**Figure 1. Illustration of KKT condition using a toy example.** Virtual concave functions, , were created to show how to find the optimal value which maximizes in the parameter space. The parameter, , is restricted to be between zero and one, and the parameter space is grayed out. (A) If the value which maximizes is negative, the tangent slopes at both zero and one will be negative. It is in violation of the KKT condition that tangent slope is negative at one. However, the negative tangent slop at zero satisfies the KKT condition, so the maximizer within the parameter space is zero. (B) When the value which maximizes is greater than 1, the optimal value is one since positive tangent slope at one meets the KKT condition. (C) When the maximizer is located in the parameter space, tangent slopes at both boundaries of the parameter space do not satisfy the KKT condition. Therefore, restrictions do not affect the result of optimization.

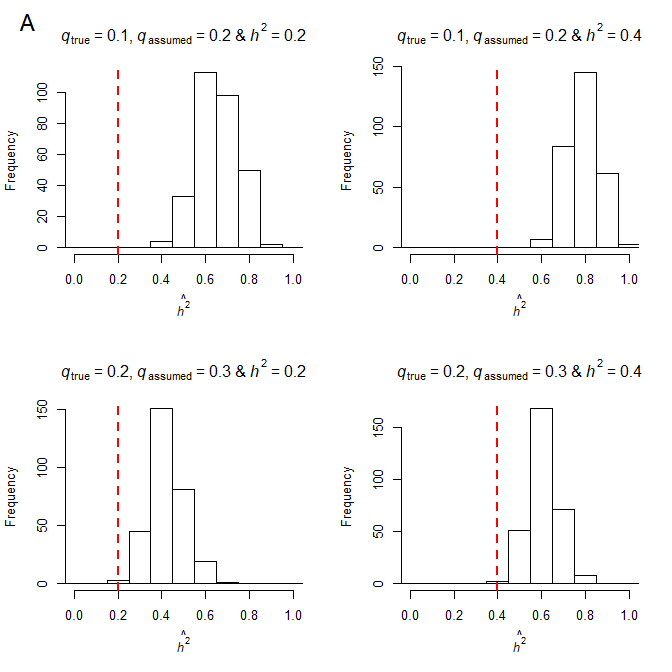
****

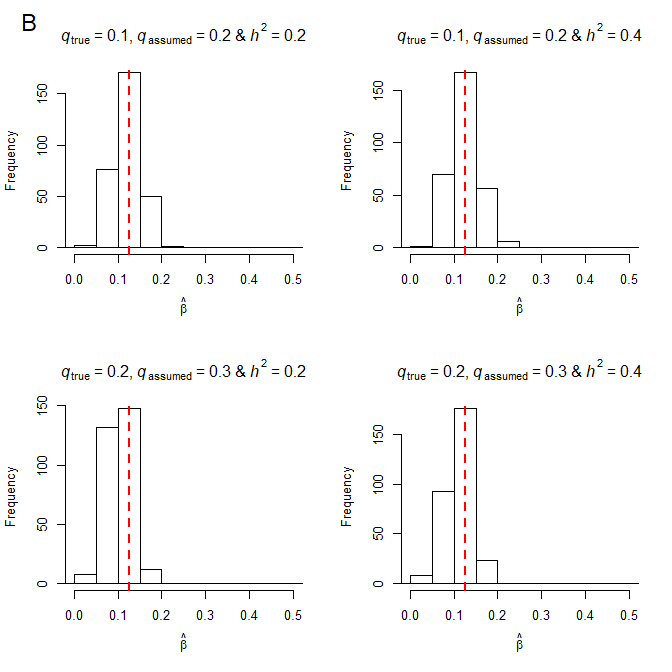
**Figure 2. Distribution of estimates based on 300 replicates when the prevalence was correctly specified.** True prevalence was 0.1 in the upper two figures and 0.2 in the lower two figures. True heritability was 0.2 in the left two figures and 0.4 in the right two figures. (A) The distribution of . Red dashed line indicates true value of . (B) The distribution of . Red dashed line indicates true value of .

****

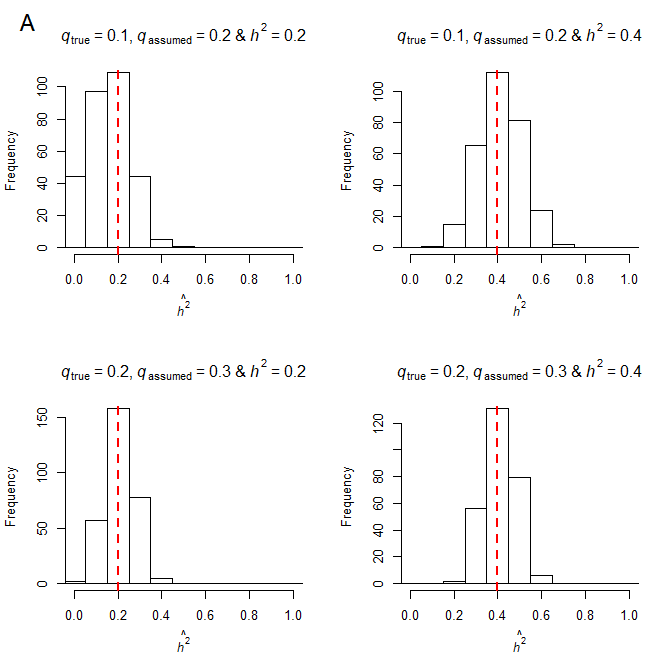
****

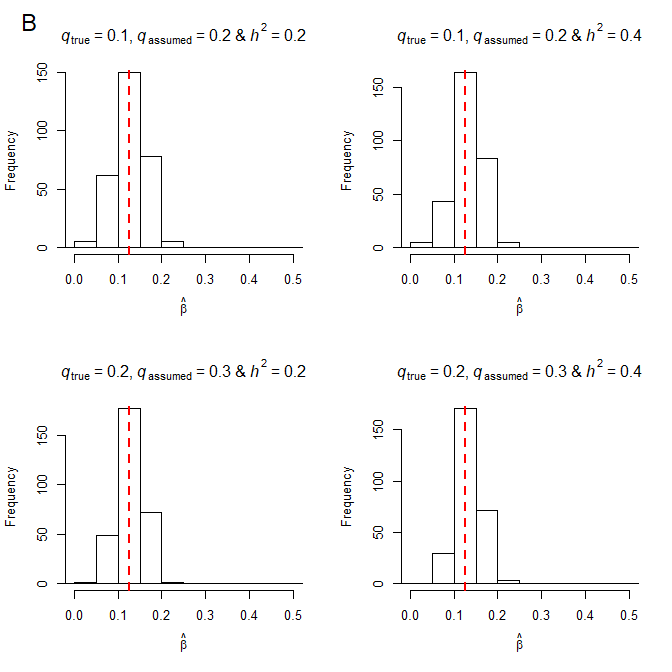
**Figure 3. Distribution of estimates without the intercept based on 300 replicates when the prevalence was misspecified.** The prevalence was assumed to be 0.1 larger than the true prevalence. True prevalence was 0.1 in the upper two figures and 0.2 in the lower two figures. True heritability was 0.2 in the left two figures and 0.4 in the right two figures. (A) The distribution of . Red dashed line indicates true value of . (B) The distribution of . Red dashed line indicates true value of .

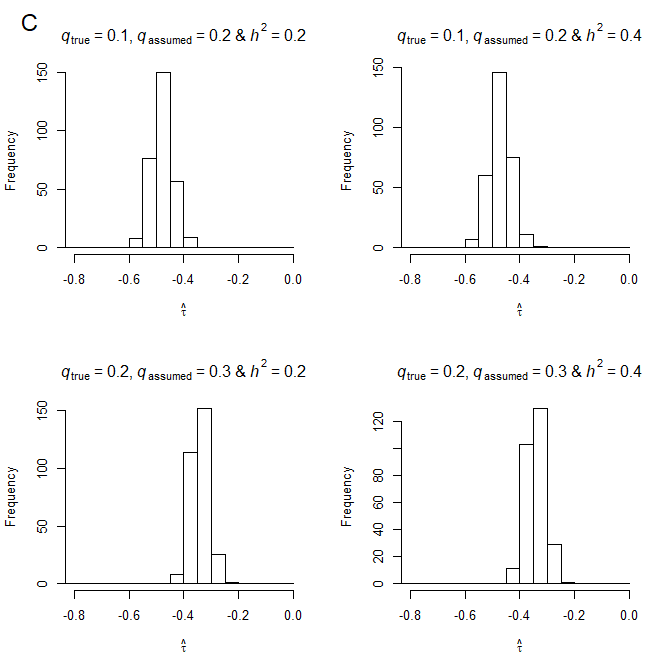
****

****

**Figure 4. Distribution of estimates with the intercept based on 300 replicates when the prevalence was misspecified.** The prevalence was assumed to be 0.1 larger than the true prevalence. True prevalence was 0.1 in the upper two figures and 0.2 in the lower two figures. True heritability was 0.2 in the left two figures and 0.4 in the right two figures. (A) The distribution of . Red dashed line indicates true value of . (B) The distribution of . Red dashed line indicates true value of . (C) The distribution of .





****

**Appendix (A)**

**Numerical analysis for optimization of the heritability in M-step of EM algorithm**

The first derivative of with respect to is given by

where . Then, the objective function will be

Similarly, we can get the first derivative of with respect to as follows,

where and

Finally, is updated according to the following iterative steps keeping and fixed,

**Appendix (B)**

**Numerical analysis for maximizing the global lower bound**

If we denote the global lower bound for the conditional log-likelihood as , then the first derivative of with respect to ,,is given by

and we previously calculated . Using the chain rule, we can easily obtain the first derivative of with respect to as follows,

where is the probability density function for the standard normal. To apply Newton-Raphson algorithm for the objective function , we derive the first derivative of with respect to , , as follows,

and each term is given by

and

With these terms, we iteratively update using following equation until convergence,

**References**

1. Visscher, P.M., W.G. Hill, and N.R. Wray, *Heritability in the genomics era—concepts and misconceptions.* Nature reviews genetics, 2008. **9**(4): p. 255.

2. Yang, J., et al., *Common SNPs explain a large proportion of the heritability for human height.* Nature genetics, 2010. **42**(7): p. 565.

3. Yang, J., et al., *GCTA: a tool for genome-wide complex trait analysis.* The American Journal of Human Genetics, 2011. **88**(1): p. 76-82.

4. Vattikuti, S., J. Guo, and C.C. Chow, *Heritability and genetic correlations explained by common SNPs for metabolic syndrome traits.* PLoS genetics, 2012. **8**(3): p. e1002637.

5. Dudbridge, F., *Power and predictive accuracy of polygenic risk scores.* PLoS genetics, 2013. **9**(3): p. e1003348.

6. Papachristou, C., C. Ober, and M. Abney, *Genetic variance components estimation for binary traits using multiple related individuals.* Genetic epidemiology, 2011. **35**(5): p. 291-302.

7. Burton, P.R., et al., *Genetic variance components analysis for binary phenotypes using generalized linear mixed models (GLMMs) and Gibbs sampling.* Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society, 1999. **17**(2): p. 118-140.

8. Dempster, E.R. and I.M. Lerner, *Heritability of threshold characters.* Genetics, 1950. **35**(2): p. 212.

9. Van Vleck, L., *Estimation of heritability of threshold characters.* Journal of Dairy Science, 1972. **55**(2): p. 218-225.

10. Hoeschele, I. and B. Tier, *Estimation of variance components of threshold characters by marginal posterior modes and means via Gibbs sampling.* Genetics Selection Evolution, 1995. **27**(6): p. 519.

11. Lee, S.H., et al., *Estimating missing heritability for disease from genome-wide association studies.* The American Journal of Human Genetics, 2011. **88**(3): p. 294-305.

12. Park, S., et al., *Adjusting heterogeneous ascertainment bias for genetic association analysis with extended families.* BMC medical genetics, 2015. **16**(1): p. 62.

13. Dempster, A.P., N.M. Laird, and D.B. Rubin, *Maximum likelihood from incomplete data via the EM algorithm.* Journal of the royal statistical society. Series B (methodological), 1977: p. 1-38.

14. Jebara, T. and A. Pentland. *Maximum conditional likelihood via bound maximization and the CEM algorithm*. in *Advances in neural information processing systems*. 1999.

15. Fisher, R.A., *XV.—The correlation between relatives on the supposition of Mendelian inheritance.* Earth and Environmental Science Transactions of the Royal Society of Edinburgh, 1919. **52**(2): p. 399-433.

16. Abney, M., M.S. McPeek, and C. Ober, *Estimation of variance components of quantitative traits in inbred populations.* The American Journal of Human Genetics, 2000. **66**(2): p. 629-650.

17. Jacquard, A., *The genetic structure of populations*. Vol. 5. 2012: Springer Science & Business Media.

18. Atkinson, K.E., *An introduction to numerical analysis*. 2008: John Wiley & Sons.

19. Weiss, N.A., *A course in probability*. 2006: Addison-Wesley.

20. Bertsekas, D.P., *Constrained optimization and Lagrange multiplier methods*. 2014: Academic press.

21. Kuhn, H.W. and A.W. Tucker, *Nonlinear programming*, in *Traces and emergence of nonlinear programming*. 2014, Springer. p. 247-258.

22. Neal, R.M. and G.E. Hinton, *A view of the EM algorithm that justifies incremental, sparse, and other variants*, in *Learning in graphical models*. 1998, Springer. p. 355-368.

23. Fisher, R.A. *Theory of statistical estimation*. in *Mathematical Proceedings of the Cambridge Philosophical Society*. 1925. Cambridge University Press.

24. Finkelstein, D.M., et al., *A score test for association of a longitudinal marker and an event with missing data.* Biometrics, 2010. **66**(3): p. 726-732.

25. Rao, C.R. *Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation*. in *Mathematical Proceedings of the Cambridge Philosophical Society*. 1948. Cambridge University Press.

26. Stram, D.O. and J.W. Lee, *Variance components testing in the longitudinal mixed effects model.* Biometrics, 1994: p. 1171-1177.