Summary report about proposed method

Xuelong Wang

2019-01-11

Contents

1 Topic

The overall goal of this project is to understand the relationships among chemical exposures and health outcomes. Since the relation could be very complicated and the effect of each chemical factor could very weak, one may want to model the relation of variance between chemical factors and health outcome.

To achieve that goal we need to break things into steps, so the current goal of this project is to estimate the main and interactive effects given simulated responses.

More specifically, we are trying to adopt and modify an approach called GCTA method, which is used for estimating of heritablity in genoew-wide study.

2 Model

The model we are using is mixed model with main effect and interaction effect. The effects could either be fixed or random effect. But for now, we assume that both of them are fixed.

$$y = \alpha + \sum_{j=1}^{p} x_j \beta_j + \sum_{j \neq k} \gamma_{jk} x_j x_k + \epsilon.$$

Matrix form

$$y = X^T \beta + X^T \Gamma X + \epsilon,$$

Where

- $X = (x_1, \dots x_p)^T$, in our case assumpe $X \sim N(0, \Sigma_p)$
- $\epsilon \perp\!\!\!\perp x_{ji}$
- $\beta = (\beta_1, \dots, \beta_p)^T$ is fixed
- Γ is a $p \times p$ matrix with diagonal elements equal to 0.

3 GCTA and proposed method (a modified GCTA method)

The details of the GCTA and proposed method could be found in previous report (simulation of fixed and random effect). The main idea of the proposed method is to add a decorrelation step, so that the GCTA method could deal with correlated data.

There is a suggestion (Aim 1(b) Proposal) of GCTA method. In order to let the method work correctly, the causal covariates to be independent themselves and independent of non-causal covariates. But based on the simulation study and some theoretical results, we found that as long as the main effect and the interaction effect are uncorrelated to each other, $\mathbf{Cov}(\mathbf{X_m^T}\beta_\mathbf{m}, \mathbf{X_i^T}\beta_\mathbf{i})$, then we are able to estimate both of the effects' variance without much bias. This suggests that the **Independence** of covariates may not be that crucial.

However, if the correlation between main and interaction is not zero, then it will cause some trouble in variances estimation. The correlation term, which is not considered by the GCTA method, will affect the estimation result for both effect. One solution for walking round that problem is a two-step method. Firstly, we estimate the total variance, which is the summation of main and interaction and their correlation method. And then, we use some statistical test to determine if there exits an interaction effect. Followings are some details of the methods.

4 Variance Estimation under different conditions

Before we go into details, Let me just rewrite the issue part in math formula so that we could get a better understand.

$$\begin{aligned} Var(Y) &= Var(X^T\beta + X^T\Gamma X) + Var(\epsilon) \\ &= Var(X^T\beta) + Var(X^T\Gamma X) + 2Cov(X^T\beta, X^T\Gamma X) + Var(\epsilon) \end{aligned}$$

1. There is an additional terms $Cov(X^T\beta, X^T\Gamma X)$
2. The main effect x_i and the interaction effect x_ix_j are dependent and cannot be independent anymore. Besides, even if X is an independent random vector, the interaction effect are not independent themselves, i.e. x_ix_j and $XX_{j'}$ are dependent.
3. In order to keep the variance structure, we can only apply the linear transformation on the main effects, not the interactive effects.

4.1 Normal distribution

Let's just start with the most straightforward one, which is when covariates follows a Normal distribution. The properties of normal distribution simplify the situation, so that the proposed method can work well. Namely, no matter covariates are independent or not, we can always have

$$\begin{aligned} Cov(X_i^T\beta, X_i^T\Gamma X_i) &= E[(X_i^T\beta - E(X_i^T\beta))(X_i^T\Gamma X_i - E(X_i^T\Gamma X_i))] \\ &= E[X_i^T\beta(X_i^T\Gamma X_i - E(X_i^T\Gamma X_i)) \\ &= E[X_i^T\beta(X_i^T\Gamma X_i - trace(\Gamma\Sigma_p))] && \text{Note that } \gamma_{jj} = 0 \\ &= E[X_i^T\beta \cdot X_i^T\Gamma X_i] \\ &= E[(\sum_m (x_{im}\beta_m))(\sum_j \sum_k \gamma_{jk}x_{ij}x_{ik})] \\ &= 0 && \text{Note that } E(x_i^2 x_j) = 0 \text{ because of centered data} \end{aligned}$$

Then we have,

$$\begin{aligned} Var(Y) &= Var(X^T\beta + X^T\Gamma X) + Var(\epsilon) \\ &= Var(X^T\beta) + Var(X^T\Gamma X) + Var(\epsilon) \end{aligned}$$

. Therefore, we don't have to worry about the covariance term and should expect good variance estimation result

Following are some results and conclusion based on simulation and theoretical study.

4.1.1 Independent covariates

In this situation, both GCTA and proposed method can work well.

4.1.2 Dependent covariates

In this situation, the proposed method's performance is much better then the GCTA itself in term of unbiaseness. This indicates that the correlation structure may be more necessary for the GCTA method than independency. The simulation result could be found on my report with date 08/15/2018

The normal distribution is relatively easy to deal with because of the property. We could just add an decorrelation step before the GCTA method, than we could get a good result. This also indicates an option to deal with a more complicated problem, for example, the non-normal distribution. We could just transfer the data into a normal or normal-like distribution and do the analysis as usual. But we could not or hard to control the magnitude of main and interaction variance and their relative relation after transfermation.

4.2 Non-normal distribution

4.2.1 Independent covariates

For Independent case, GCTA method appears to work fine.

4.2.2 Dependent covariates

Now, we move to a more general and also more complicated situation, non-normal distribution. For the non-normal or long tail distribution, we cannot guarantee that the third monment equal to zero, therefore the covariance of main and interaction effect is no longer zero. Simulation study (08/15/2018) has shown that in such case, even the proposed method cannot estimate both of the effect correctly.

4.3 sub sampling method for evaluating methods' performance