Summary report about proposed method

Xuelong Wang

2019-01-18

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.

3.1 the proposed method

The approach of the proposed method is to uncorrelated the observed covariates by using a linear transformation, i.e. $Z = A^{-1}X$, So that $Var(Z) = I_p$. In this way, we could get the uncorrelated predictors and use them as the input of the regular GCTA method.

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

4.1.1 Independent covariates

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

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

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

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. But because the covariates follows normal distribution, which means that after decorrelation, they become independent variables. The simulation result could be found on my report with date 08/15/2018

Let's assume $X \sim N(\mu, \Sigma)$, after standardization we have

$$E(X) = 0 \quad Var(X) = \tilde{\Sigma},$$

Where $diag(\tilde{\Sigma}) = \vec{1}$, and $\tilde{\Sigma}_{ij} = \rho_{ij}$.

If we uncorrelated the X by $Z = A^{-1}X$, where $A = \tilde{\Sigma}^{1/2}$, then we will have

$$Z \sim N(0, I_p),$$

We are back to the independent case again.

A thougth for take advantage of the properties of normal distirbution: The normal distribution is relatively easy to deal with because of the property. We could just add a 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. Similar with the independent normal distribution, the covariance of main and interaction effect will be zero. One condition is that we need the covariates to be standardized.

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 make them indepedent by just some linear operation, 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 effects correctly.

However, we can take one step back, which means instead of considering the main and interaction effect separately, we look at the total variance related to covariates:

$$\begin{aligned} Var(Y) &= Var(X^T\beta + X^T\Gamma X) + Var(\epsilon) \\ &= Var(X^T\beta + X_{inter}^T\gamma) + Var(\epsilon) \\ &= \begin{bmatrix} X \\ X_{inter} \end{bmatrix}^T \begin{bmatrix} \beta \\ \gamma \end{bmatrix} + Var(\epsilon) \\ &= Var(X_{total}^T\beta_{total}) + Var(\epsilon) \end{aligned}$$

In this way, we combine the main and interaction effect together as the new covariates. The good news for estimation of the total variance is we don't need to worry about the covariance part, because it is always included in the total variance term. The downside is that we will never make the total effect to be independent themselves anymore, because the interaction terms contains the main effects.

Based on the simulation, we find the proposed method works fine, and it works much better than the origianl GCTA method. After decorrelation, the GCTA is able to estimate the total variance accuretly based on the simualtion result.

4.3 Decorrelation method

For Decoorelation step, it could be done by using the Spectral decomposition if the observed covariates X is full column rank. After adding this decorrelation step, the GCTA can work well in term of estimation the main and interaction variance. However, if $p > n$, which means the number of parameter is larger than the number of observation, so the sample covariance is not full rank, then the proposed method suddenly work much worse. One working solution for that issuse is to reduce the dimension of covariate and then apply the decorrealtion.

4.3.1 Spectral decomposition

$$Var(X) = \Sigma_X = U\Lambda U^T,$$

- X is the random vector with dim as $p \times 1$,
- Σ_X is $p \times p$ symmetry and p.d. matrix,
- Λ is a diagonal matrix with each diagonal element as the eigenvalue.

4.3.2 Assume the Σ_X is full rank

To decorreate the X, we could just take the reciprocal of each square root of eigenvalue as following.

$$\Sigma_X^{-\frac{1}{2}} = U\Lambda^{-\frac{1}{2}}U^T,$$

where $\Lambda^{-\frac{1}{2}} = \begin{bmatrix} e_1^{-\frac{1}{2}} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & e_p^{-\frac{1}{2}} \end{bmatrix}$

So that after transformation the $\Sigma_X^{-\frac{1}{2}}X$ has identity covariance matrix as following,

$$Var(\Sigma_X^{-\frac{1}{2}}X) = \Sigma_X^{-\frac{1}{2}}\Sigma_X\Sigma_X^{-\frac{1}{2}} = U\Lambda^{-\frac{1}{2}}U^TU\Lambda^{-1}U^TU\Lambda^{-\frac{1}{2}}U^T = I_p.$$

4.3.2.1 Assume the Σ_X is not full rank

$$Var(X) = \Sigma_X = U\Lambda U^T = \begin{bmatrix} U_1 & U_2 \end{bmatrix} \begin{bmatrix} \Lambda_1 & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U_1^T \\ U_2^T \end{bmatrix} = U_1\Lambda_1U_1^T,$$

- U_1 is a $p \times r$ matrix with r < p and in most of case r = n the sample size.

Then after applying the same procedure we get following,

$$\Sigma_X^{-\frac{1}{2}} = U_1\Lambda_1^{-\frac{1}{2}}U_1^T,$$

Note that in this case, I'm using Moore Penrose inverse.

After transformation the X we have,

$$Var(\Sigma_X^{-\frac{1}{2}}X) = \Sigma_X^{-\frac{1}{2}}\Sigma_X\Sigma_X^{-\frac{1}{2}} = U_1\Lambda_1^{-\frac{1}{2}}U_1^TU_1\Lambda_1^{-1}U_1^TU_1\Lambda_1^{-\frac{1}{2}}U_1^T = U_1U_1^T,$$

Note that by the property of the U we have

$$U_1U_1^T + U_2U_2^T = I_p, \qquad (U_1U_1^T)^TU_1U_1^T = U_1U_1^T,$$

Besides, $U_1U_1^T$ and $U_2U_2^T$ are indempotent and $rank(U_2U_2^T) + rank(U_1U_1^T) = p$.

So if the X is not full rank we cannot decorrelation the covariance matrix to an identity matrix.

4.3.3 SVD with dimension reduction step

One possible solution for the issue of too many predictors is to reduce the column dimension of covariates first and then used the reduced covariate as the input for the prosposed method. I have tried several dimension reduction method, and found that a modified SVD method provides promosing results based on simulation study. the details as following:

$$\begin{aligned} X = UDV^T &= \begin{bmatrix} U_r & U_2 \end{bmatrix} \begin{bmatrix} D_r & 0 \\ 0 & D_2 \end{bmatrix} \begin{bmatrix} V_r & V_2 \\ V_3 & V_4 \end{bmatrix}^T \\ &= \begin{bmatrix} U_rD_r & U_2D_2 \end{bmatrix} \begin{bmatrix} V_r^T & V_3^T \\ V_2^T & V_4^T \end{bmatrix} = \begin{bmatrix} U_rD_rV_r^T + U_2D_2V_2^T & U_rD_rV_3^T + U_2D_2V_4^T \end{bmatrix} \end{aligned}$$

Ignore V_2, V_3 and V_4 , then we have the X_r as following

$$X_r = U_r D_r V_r^T.$$

We use X_r as the new covariates to the proposed methd. Therefore, we reduce the dimension from p to n. After calculating X_r, we can regard X_r as our new predictors and use it as the input to the proposed method Note that we could use this blocking method to reduce X's dimension to $k, k \leq min(p, n)$.

4.3.4 PCA method

4.4 sub sampling method for evaluating methods' performance