Getting started with subgxe

Introduction

`subgxe` is an R package for combining summary data from multiple association studies or multiple phenotypes in a single study by incorporating potential gene-environment (G-E) interactions into the testing procedure. It is an implementation of the *p value-assisted subset testing for associations (pASTA)* framework proposed by Yu et al(2019). The goal is to identify a subset of studies or traits that yields the strongest evidence of associations and give a meta-analytic p-value. This vignette offers a brief introduction to the basic use of `subgxe`. For more details on the algorithms used by `subgxe` please refer to the paper.

subgxe Example

We use simulated data of $K = 5$ independent case-control studies that come along with the package to illustrate the basic use of `subgxe`. In each data set, `G`, `E`, and `D` denote the genetic variant, environmental factor, and disease status (binary outcome), respectively. In this case, `G` is coded as binary (under a dominant or recessive susceptibility model). It can also be coded as allele count (under the additive model). Two of the 5 studies have non-null genetic associations with the true marginal genetic odds ratio being 1.09. Each study has 6,000 cases and 6,000 controls, with the total sample size n_k being 12,000. For the specific underlying parameters of the data generating model, please refer to the original article (Table A4, Scenario 2).

```r
# library(devtools)
# install_github("umich-cphds/subgxe", build_opts = c())
library(subgxe)
```

We first obtain a $K \times 1$ vector of input p-values by conducting association test for each study. For study k, $k = 1, \cdots, K$, the *joint* model with G-E interaction is

$$\text{logit}[E(D_{ki}|G_{ki}, E_{ki})] = \beta_0^{(k)} + \beta_G^{(k)} G_{ki} + \beta_E^{(k)} E_{ki} + \beta_{GE}^{(k)} G_{ki} E_{ki}$$

where $i = 1, \cdots, n_k$. The model can be further adjusted for potential confounders, which we drop from the presentation for the simplicity of notation.

To detect the genetic association while accounting for the G-E interaction, one can test the null hypothesis

$$\beta_G^{(k)} = \beta_{GE}^{(k)} = 0$$

based on the joint model for each study k. The coefficients can be estimated by maximum likelihood using the `glm` function. For alternative null hypotheses and methods for estimation of coefficients, see the reference mentioned above.

A common choice for testing the null hypothesis $\beta_G^{(k)} = \beta_{GE}^{(k)} = 0$ is the likelihood ratio test (LRT) with 2 degrees of freedom, which can be carried out with the `lrtest` function in the package `lmtest`. We use the results of LRT as an example to demonstrate the use of `subgxe`. For comparative purposes, we also look at the p-values of the *marginal* genetic associations obtained by Wald test, i.e. the p-values of $\hat{\alpha}_G^{(k)}$ in the model

$$\text{logit}[E(D_{ki}|G_{ki})] = \alpha_0^{(k)} + \alpha_G^{(k)} G_{ki}.$$

```r
library(lmtest)

K <- 5 # number of studies
study.pvals.marg <- NULL
study.pvals.joint <- NULL
```

```
for(i in 1:K){
  joint.model <- glm(D ~ G + E + I(G*E), data=studies[[i]], family="binomial")
  null.model <- glm(D ~ E, data=studies[[i]], family="binomial")
  marg.model <- glm(D ~ G, data=studies[[i]], family="binomial")
  study.pvals.marg[i] <- summary(marg.model)$coef[2,4]
  study.pvals.joint[i] <- lmtest::lrtest(null.model, joint.model)[2,5]
}
```

Then we use the `pasta()` function in the `subgxe` package to conduct subset analysis and obtain a meta-analytic p-value for the genetic association.

- The `cor` parameter is a correlation matrix of the study-specific p-values. In this example, since the studies are independent, the p-values are independent as well, and therefore the `cor` should be an identity matrix. In a *multiple-phenotype* analysis where the phenotypes are measured on the same set of subjects, one way to approximate the correlations among p-values is to use the phenotypic correlations.

```
study.sizes <- c(nrow(studies[[1]]), nrow(studies[[2]]), nrow(studies[[3]]),
                 nrow(studies[[4]]), nrow(studies[[5]]))

cor.matrix <- diag(1, K)
pasta.joint <- pasta(p.values=study.pvals.joint, study.sizes=study.sizes, cor=cor.matrix)
pasta.marg <- pasta(p.values=study.pvals.marg, study.sizes=study.sizes, cor=cor.matrix)

pasta.joint$p.pasta # delete 'joint'
#> [1] 0.001859015
pasta.joint$test.statistic$selected.subset
#>   Var1 Var2 Var3 Var4 Var5
#> 4    1    1    0    0    0

pasta.marg$p.pasta # delete 'joint'
#> [1] 0.03312643
pasta.marg$test.statistic$selected.subset
#>   Var1 Var2 Var3 Var4 Var5
#> 8    1    1    1    0    0
```

From the output we observe that when the G-E interaction is taken into account, `pasta` yields a meta-analytic p-value of 0.002 and identifies the first two studies as non-null. On the other hand, if we only consider the marginal associations, the meta-analytic p-value becomes much larger (p=0.033) and the first three studies are identified as having significant associations.

Reference

- Yu Y, Xia L, Lee S, Zhou X, Stringham H, M, Boehnke M, Mukherjee B: Subset-Based Analysis Using Gene-Environment Interactions for Discovery of Genetic Associations across Multiple Studies or Phenotypes. *Hum Hered* 2019. doi: 10.1159/000496867