Genotype Conditional Association Test Vignette

Wei Hao, Minsun Song, John D. Storey

October 29, 2024

1 Introduction

The `gcatest` package provides an implementation of the Genotype Conditional Association Test (GCAT) [2]. GCAT is a test for genetic association that is powered by Logistic Factor Analysis (LFA) [1]. LFA is a method of modeling population structure in a genome wide association study. GCAT performs a test for association between each SNP and a trait (either quantitative or binary). We have shown that GCAT is robust to confounding from population structure.

2 Sample usage

We include a sample dataset with the package. `sim_geno` is a simulated genotype matrix. `sim_trait` is a simulated trait. There are 10,000 SNPs and 1,000 individuals. The first five SNPs are associated with the trait. This simulations were done under the Pritchard-Stephens-Donnelly model with $K = 3$, with Dirichlet parameter $\alpha = 0.1$ and variance allotment in the trait corresponding to 5% genetic, 5% environmental, and 90% noise. This dataset is simulated under identical parameters as the PSD simulation in Figure 2 of the paper [2], except that we have adjusted the size of the simulation to be appropriate for a small demo.

```
library(lfa)
library(gcatest)
dim(sim_geno)

## [1] 10000  1000

length(sim_trait)

## [1] 1000
```

2.1 `gcat`

The first step of `gcat` is to estimate the logistic factors:

```
LF <- lfa(sim_geno, 3)
dim(LF)

## [1] 1000    3
```

Then, we call the `gcat` function, which returns a vector of p-values:

```
gcat_p <- gcat(sim_geno, LF, sim_trait)
```

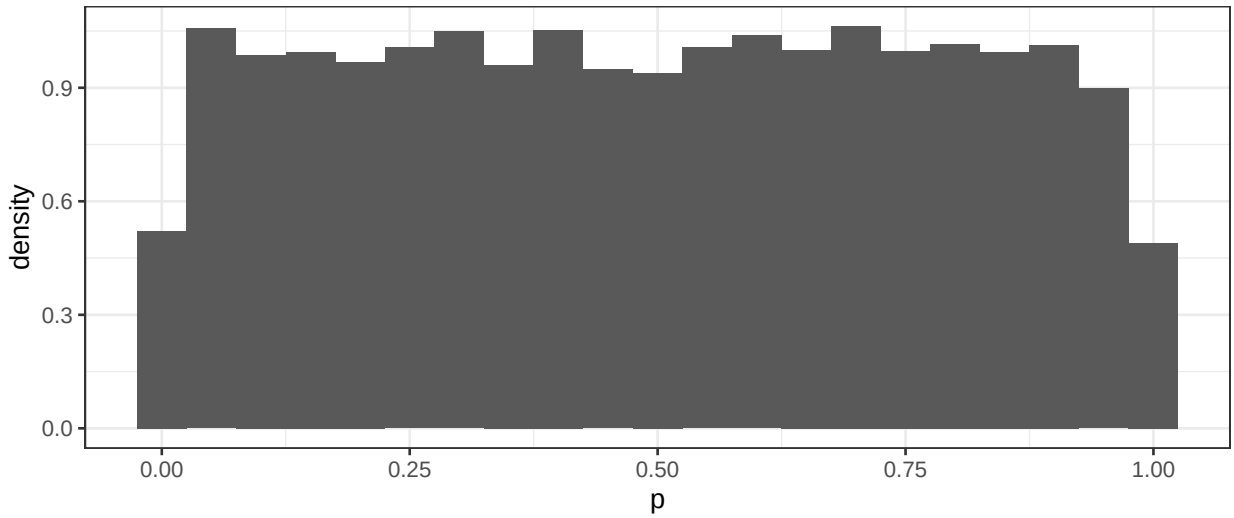
We can look at the p-values for the associated SNPs:

```
gcat_p[1:5]

## [1] 5.873675e-01 1.570672e-01 1.240494e-03 4.799294e-01 8.607748e-11
```

And also plot the histogram of the unassociated SNPs:

```
library(ggplot2)
dat <- data.frame(p = gcat_p[6:10000])
ggplot(dat, aes(p, after_stat(density))) + geom_histogram(binwidth=1/20) + theme_bw()
```



3 Data Input

The `genio` package provides the function `read_plink` for parsing PLINK binary genotypes (extension: `.bed`) into an R object of the format needed for the `gcat` function. A `BEDMatrix` object (from the eponymous function and package) is also supported, and can result in reduced memory usage (at a small runtime penalty).

References

[1] Wei Hao, Minsun Song, and John D. Storey. Probabilistic models of genetic variation in structured populations applied to global human studies. *Bioinformatics*, 32(5):713–721, March 2016.

[2] Minsun Song, Wei Hao, and John D. Storey. Testing for genetic associations in arbitrarily structured populations. *Nat. Genet.*, 47(5):550–554, May 2015.