Gene set enrichment analysis using Wilcoxon tests

Chris Wallace
Cambridge Institute for Medical Research
University of Cambridge
email: chris.wallace at cimr.cam.ac.uk

December 5, 2016

Contents

1 Overview

Gene set enrichment analysis (GSEA) is typically based on tests derived from the Kolmogorov-Smirnov, which is underpowered and a need for simpler methods has been identified.[2]

The `wgsea` package contains functions for conducting GSEA using a Wilcoxon test to test for differences in the distribution of p values between SNPs within the gene set under test and a control set of SNPs. The mean of the Wilcoxon test statistic is unperturbed by correlation between the SNPs resulting from linkage disequilibrium, but its variance is. We use permutation of the case control status to generate Wilcoxon statistics under the null. This permutation preserves the correlation structure, and because we are only trying to estimate the variance rather than a permutation p value we need only a modest number of permutations.

This vignette steps through analysis of an example dataset. For further detail, and example of an application to real data, please see [1].

2 Example data

We load the example data from snpStats, and split the SNPs into two groups, according to whether or not they show deviation from Hardy Weinberg equilibrium. This is generally an indicator of poor genotyping and such SNPs are excluded, particularly as they may be examples of differential genotyping error between cases and controls which may induce spurious associations. We will examine here whether that is the case in this small subset of data.

```
> library(snpStats)
> library(wgsea)
> ## load example data from snpStats
> data(for.exercise,package="snpStats")
```

```
> ## generate an artificial indicator of test and control SNPs.
> snpsum <- col.summary(snps.10)
> snp.indicator <- abs(snpsum$z.HWE) > 1.96
> ## subset to case and control objects
> case <- snps.10[subject.support$cc==1,!is.na(snp.indicator)]
> control <- snps.10[subject.support$cc==0,!is.na(snp.indicator)]
> snp.indicator <- snp.indicator[!is.na(snp.indicator)]
```

We use `pairtest`, a wrapper for the `snpStats` function `single.snp.tests`, to generate p values for the association of each SNP with T1D. We also store the minor allele frequencies (MAF) to allow investigation of whether MAF confounds our test.

```
> maf <- col.summary(control)[,"MAF"]
> p <- pairtest(case,control)
```

A naive Wilcoxon analysis of the p values might be

```
> wilcox.test(x=p[snp.indicator==1],
+             y=p[snp.indicator==0])

        Wilcoxon rank sum test with continuity correction

data:  p[snp.indicator == 1] and p[snp.indicator == 0]
W = 68559000, p-value < 2.2e-16
alternative hypothesis: true location shift is not equal to 0
```

but the considerable linkage disequilibrium between SNPs would cause the variance to be under estimated and so cause the test to be anti conservative.

Instead, we generate a set of permuted phenotypes and use these to generate a null distribution of the Wilcoxon test statistic, from which we calculate its variance. This step can take a while, depending on the number of permutations and the size of your dataset. Here, we use only 10 to keep computation down. `pairtest` prints dots to monitor progress of the permutations.

```
> n.perm=10
> p.perm <- pairtest(case,control,n.perm=n.perm)

..........
```

We can now calculate a Wilcoxon statistic for the observed data, and for the permuted data.

```
> ## calculate wilcoxon score
> W <- wilcoxon(p,snps.in=which(snp.indicator==1))
> Wstar <- wilcoxon(p.perm,snps.in=which(snp.indicator==1))
```

The empirical estimate of variance is considerably increased compared to the theoretical, and we can examine how many permutations are required to estimate it reasonably well. Because we have set `n.perm=10` in this vignette, we cannot see convergence; instead we substitute a figure generated previously using 10000 permutations. Our experience suggests that the results here, showing that convergence takes several thousand permutations, are unusual, and relate to the inclusion of SNPs with extreme case-control differences in a relatively small dataset. For larger datasets in which SNPs with genotyping errors have been excluded, convergence is

typically achieved in hundreds, and not thousands of permutations. However, each permutation is computationally expensive, and for this reason, we recommend storing the `p.perm` object once generated, as refinements are often made to the list of SNPs to be tested, and `Wstar` can be regenerated relatively quickly given `p.perm`. Note also, that if the whole genome is being examined, it can make sense to parallelise the computations, by running different chromosomes seperately, for example. The p.perm objects thus created can be combined using `rbind()`.

```
> ## running empirical estimate of variance
> #W.var <- sapply(2:n.perm,function(i) var(Wstar[sample(1:n.perm,size=i)]))
> W.var <- sapply(2:n.perm,function(i) var(Wstar[1:i]))
> ## theoretical variance for comparison
> n1 <- sum(snp.indicator)
> n2 <- sum(!snp.indicator)
> var.theoretical <-  exp(log(n1) + log(n2) + log(n1+n2+1) - log(12))
> ## plot
> plot(2:n.perm,W.var,ylim=range(c(W.var,var.theoretical)),pch=".",
+      xlab="Permutation number",ylab="Var(W*)",main="Estimate of Var(W*) vs number of permu
+      sub="(dotted line shows theoretical value)")
> abline(h=var.theoretical,lty=3)
```

Finally, we calculate an overall significance level for comparing whether our test SNPs are more associated with T1D than our control SNPs. Although we do not expect the expected null value of `W` to be perturbed due to linkage disequilibrium, we report two Z scores

$$Z_{\text{theoretical}} = (W - \mu)/\operatorname{sd}(W^*)$$
$$Z_{\text{empirical}} = (W - \overline{W^*})/\operatorname{sd}(W^*)$$

where μ is the theoretical mean of the Wilcoxon distribution.

```
> Z.value(W=W, Wstar=Wstar, n.in=sum(snp.indicator==1), n.out=sum(snp.indicator==0))

$Z.theoretical

        Wilcoxon theoretical mean

data:  W
Z = -5.576, p-value = 2.462e-08


$Z.empirical

        Wilcoxon empirical mean

data:  W
Z = -5.7486, p-value = 9.001e-09
```

3 Addressing potential confounding using inverse probability weighting

There is a tendency for less common SNPs to show smaller p values than more common, presumably because genotype calls tend to be less accurate for rarer SNPs, and we are contrasting

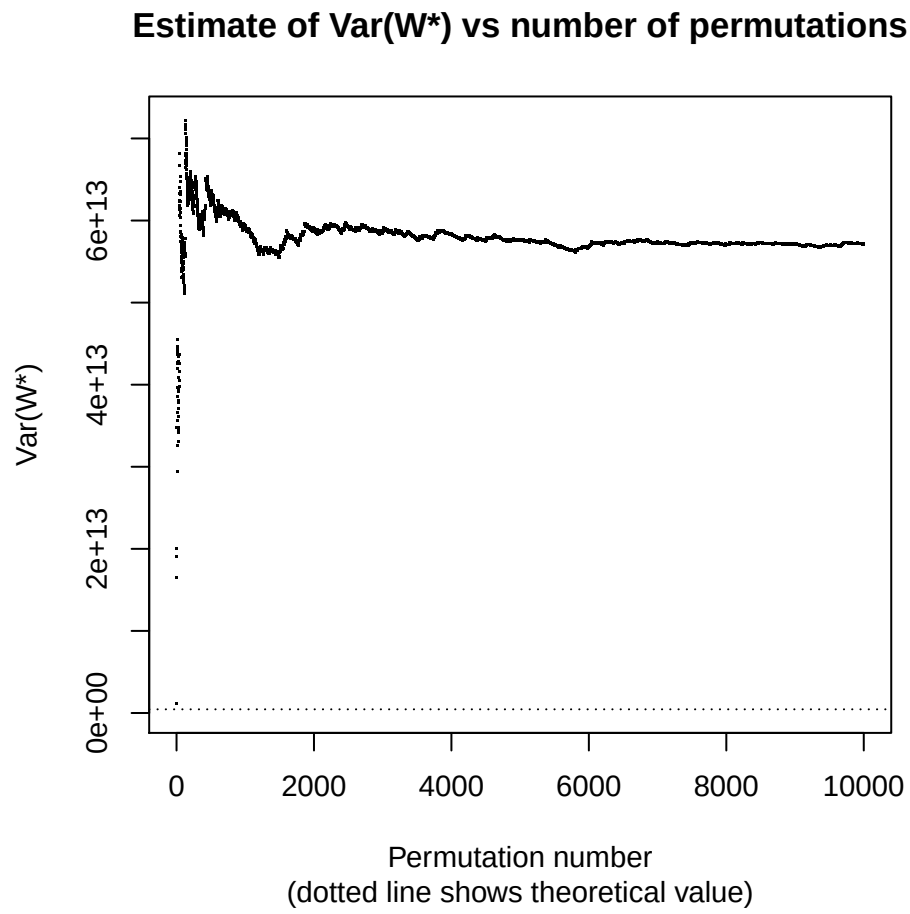

Figure 1: Histogram of p-values.

poorly with well genotyped SNPs, but we have observed this phenomenom to a lesser extent in other datasets which have passed quality control measures.

```
> cor.test(maf,p)

        Pearson's product-moment correlation

data:  maf and p
t = -15.405, df = 28495, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.10238373 -0.07935445
sample estimates:
        cor
-0.09088124
```

We address this, and could address any other similar confounders, using inverse probability weights[4, 3]. `maf` is binned (by default in bins of width 0.05) and we calculate

$$w_b = \frac{n_b/n}{m_b/m}$$

as the weight for SNPs in bin b, with m and n the number of SNPs in the test set, and the total number of SNPs, respectively and m_b, n_b the same within bin b. `maf` may be used to generate these weights by using the `weights` argument of `wilcoxon`

```
> ## calculate wilcoxon score
> W.weighted <- wilcoxon(p,snps.in=which(snp.indicator==1),weights=maf)
> Wstar.weighted <- wilcoxon(p.perm,snps.in=which(snp.indicator==1),weights=maf)
> Z.value(W=W.weighted, Wstar=Wstar.weighted,
+         n.in=sum(snp.indicator==1), n.out=sum(snp.indicator==0))

$Z.theoretical

        Wilcoxon theoretical mean

data:  W.weighted
Z = -5.2028, p-value = 1.963e-07


$Z.empirical

        Wilcoxon empirical mean

data:  W.weighted
Z = -5.3213, p-value = 1.03e-07
```

We see that, in this case, the adjustment makes some difference, but that, unsuprisingly, that SNPs with poor quality control profiles, ie those with high Hardy Weinberg absolute Z scores, do have a different profile of p values compared to SNPs with good quality control profiles. This is true even after allowing for the fact that poor quality calls are more likely for rarer SNPs.

Note that the theoretical mean of the Wilcoxon distribution is calculated ignoring weights, so when using inverse probability weighting, you should anticipate a discrepancy between the p values calculated using empirical and theoretical means, and prefer that calculated using the empirical mean.

4 Combining Wilcoxons from multiple datasets

Many GWAS datasets are now combined from different sources using meta analysis and this data structure can be analysed combining the Wilcoxon statistics over strata defined by SNP chip or GWAS group. We apply van Elteren's combined test[5] which is optimal assuming the effect is constant across strata, which seems a reasonable assumption in this application. We do not have example data from multiple datasets available. But, if required, it is simple to generate a combined test statistic by following the procedure above within each dataset, then applying the `Z.value` function with lists of W statistics. Under this use, the arguments to `Z.value` are:

- `W`: a list of W values from each study
- `Wstar`: a list of Wstar vectors from each study
- `n.in`: a vector of the number of SNPs in the test regions in each study
- `n.out`: a vector of the number of SNPs in the control regions in each study.

For example purposes, we can combine the Wilcoxon stats from above:

```
> Z.value(W=list(W,W), Wstar=list(Wstar,Wstar),
+         n.in=rep(sum(snp.indicator==1),2),
+         n.out=rep(sum(snp.indicator==0),2))

$Z.theoretical

        Wilcoxon theoretical mean

data:  4811.52961611341
Z = -5.576, p-value = 2.462e-08


$Z.empirical

        Wilcoxon empirical mean

data:  4811.52961611341
Z = -5.7486, p-value = 9.001e-09
```

References

[1] Matthias Heinig, Enrico Petretto, Chris Wallace, Leonardo Bottolo, Maxime Rotival, Han Lu, Yoyo Li, Rizwan Sarwar, Sarah R Langley, Anja Bauerfeind, Oliver Hummel, Young-Ae Lee, Svetlana Paskas, Carola Rintisch, Kathrin Saar, Jason Cooper, Rachel Buchan, Elizabeth E Gray, Jason G Cyster, Cardiogenics Consortium, Jeanette Erdmann, Christian Hengstenberg, Seraya Maouche, Willem H Ouwehand, Catherine M Rice, Nilesh J Samani, Heribert Schunkert, Alison H Goodall, Herbert Schulz, Helge G Roider, Martin Vingron, Stefan Blankenberg, Thomas Münzel, Tanja Zeller, Silke Szymczak, Andreas Ziegler, Laurence Tiret, Deborah J Smyth, Michal Pravenec, Timothy J Aitman, Francois Cambien, David Clayton, John A Todd, Norbert Hubner, and Stuart A Cook. A trans-acting locus regulates an anti-viral expression network and type 1 diabetes risk. *Nature*, 467(7314):460–464, Sep 2010.

[2] Rafael A Irizarry, Chi Wang, Yun Zhou, and Terence P Speed. Gene set enrichment analysis made simple. *Stat Methods Med Res*, 18(6):565–575, Dec 2009.

[3] P. R Rosenbaum. Model-based direct adjustment. *Journal of the American Statistical Association*, 82:387–394, 1987.

[4] P. R Rosenbaum and D. B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70:41–55, 1983.

[5] P. van Elteren. On the combination of independent two sample tests of Wilcoxon. *Bulletin of the International Statistical Institute*, 37(3):351–361, 1960.