Bioinformatics Homework2 Question2

Rhodes 2013011551

2015 年 12 月 2 日

在原 paper 的 Method 中给出了具体运算的步骤，为了说明得更加清楚，使用中文表明具体的运算过程。同时，一些步骤使用 R 编写了简单地脚本来进行运算

整个运算步骤分为三步

1°得到新的特征坐标轴

Inference of axes of variation. Let g_{ij} be a matrix of genotypes for SNP i and individual j, where $i = 1$ to M and $j = 1$ to N. We subtract the row mean $\mu_i = (\Sigma_j g_{ij})/N$ from each entry in row i to obtain a matrix with row sums equal to 0; missing entries are excluded from the computation of μ_i and are subsequently set to 0. We then normalize row i by dividing each entry by $\sqrt{p_i(1-p_i)}$, where p_i is a posterior estimate of the unobserved underlying allele frequency of SNP i defined by $p_i = (1 + \Sigma_j g_{ij})/(2 + 2N)$, with missing entries excluded from the computation. We denote the resulting matrix X. We compute an $N \times N$ covariance matrix Ψ of individuals, where $\Psi_{jj'}$ is defined to be the covariance of column j and column j' of X. We define the kth axis of variation to be the kth eigenvector of Ψ (that is, the eigenvector with kth largest eigenvalue). Thus, the ancestry a_{jk} of individual j along the kth axis of variation equals coordinate j of the kth eigenvector. We note that eigenvectors are orthonormal by definition; thus, $\Sigma_j a_{jk} = 0$, $\Sigma_j a_{jk}^2 = 1$ and $\Sigma_j a_{jk} a_{jk'} = 0$ for distinct axes k and k'. In particular, the ancestry values a_{jk} can be either positive or negative and should not be interpreted as percentages. Each axis is invariant to multiplying by a factor of −1, which does not change its interpretation.

The above procedure is motivated by the decomposition $X = USV^T$, where U is an $M \times N$ matrix whose kth column contains coordinates of each SNP along the kth principal component, S is a diagonal matrix of singular values and V is an $N \times N$ matrix whose kth column contains ancestries a_{jk} of each individual j along the kth principal component. It follows that $X^TX = VS^2V^T$; thus, the columns of V are the eigenvectors of the matrix X^TX. The matrix X^TX is equivalent up to a constant to the covariance matrix Ψ, and the matrix S^2 of squared singular values is equivalent up to a constant to the diagonal matrix of eigenvalues of Ψ.

根据 Figure 1 中的例子，现有 SNP 数据如下：

```
##    [,1] [,2] [,3] [,4] [,5]
## [1,]    1    1    1    0    0
## [2,]    0    1    2    1    2
## [3,]    2    1    1    0    1
## [4,]    0    0    1    2    2
## [5,]    2    1    1    0    0
## [6,]    0    0    1    1    1
## [7,]    2    2    1    1    0
```

对于现有的 SNP 数据，行表示特征，列表示样本，即现在有 5 个样本，7 种特征。以下公式中的下标 i 表示行，j 表示列 对于每行特征，我们求其平均数$\mu_i = (\Sigma g_{ij}/N)$ 然后再求p_i值如下$p_i = (1 + \Sigma g_{ij})/(2 + 2N)$。最后对于 SNP 数据中的每一格都数据$x_{ij}$都这么处理 $x'_{ij} = (x_{ij} - \mu_i)/(\sqrt{p_i(1 - p_i)})$（$\mu_i$和$p_i$定义前面给出）。此步骤即相当于对原数据的“归一化处理”。

此步骤的 R 脚本如下

```
sum<-as.data.frame(0)
mean<-as.data.frame(0)
p<-as.data.frame(0)
divider<-as.data.frame(0)
for(i in 1:7)
{
  sum[i,1]=sum(Genotypes[i,])
  mean[i,1]=sum[i,1]/5
  p[i,1]<-(1+sum[i,1])/(2+2*5)
  divider[i,1]<-(p[i,1]*(1-p[i,1]))^(0.5)
}

data <- Genotypes
```

```
for(i in 1:7)
  data[i,]=(data[i,]-mean[i,1])/divider[i,1]
```

处理后得到数据如下

```
##       [,1]     [,2]     [,3]     [,4]     [,5]
## [1,]  0.8485281  0.8485281  0.8485281 -1.2727922 -1.2727922
## [2,] -2.4340443 -0.4056740  1.6226962 -0.4056740  1.6226962
## [3,]  2.0000000  0.0000000  0.0000000 -2.0000000  0.0000000
## [4,] -2.0000000 -2.0000000  0.0000000  2.0000000  2.0000000
## [5,]  2.4340443  0.4056740  0.4056740 -1.6226962 -1.6226962
## [6,] -1.2727922 -1.2727922  0.8485281  0.8485281  0.8485281
## [7,]  1.6226962  1.6226962 -0.4056740 -0.4056740 -2.4340443
```

紧接着计算对于个体的协方差矩阵，大小为 5*5。同时计算协方差矩阵的本征向量，并根据本征值的大小排列。

```
cor_table=cor(data)
eigen=eigen(cor_table,symmetric=TRUE)
pca=princomp(data,cor=T)
vector<-pca$loadings
vector<-as.data.frame(vector[,])
summary(pca,loadings=TRUE)
```

```
## Importance of components:
##                           Comp.1    Comp.2     Comp.3     Comp.4
Comp.5
## Standard deviation     1.8604339 1.0346573 0.60930838 0.31146960
0
## Proportion of Variance 0.6922428 0.2141032 0.07425134 0.01940266
0
## Cumulative Proportion  0.6922428 0.9063460 0.98059734 1.00000000
1
##
```

```
## Loadings:
##      Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
## [1,] -0.511 -0.120  0.406 -0.433  0.610
## [2,] -0.488  0.190 -0.537  0.546  0.373
## [3,]  0.241  0.850 -0.167 -0.385  0.207
## [4,]  0.428 -0.473 -0.566 -0.301  0.428
## [5,]  0.509         0.447  0.524  0.513
```

可见使用前两个新特征已达到 90%,可以认为基本达到要求。而输出的 Loading 部分就是映射关系，根据 paper 的 Method 部分，他对每个向量都乘了-1，所以第一列的结果应该是 0.5,0.4,-0.3,-0.4,-0.5,这与论文里的结果不太一样哎（反复确认不知道哪里算错了，）。

12 月 9 日更新：找到了为什么不一样，R 语言的 princomp 很贴心的自带了 scale，影响了论文中自己定义的 scale。代码改为

```
pca=princomp(data,scale=F)
vector<-pca$loadings
vector<-as.data.frame(vector[,])
summary(pca,loadings=TRUE)
```

得到新的结果

```
Importance of components:
                          Comp.1     Comp.2     Comp.3     Comp.4
Comp.5
Standard deviation     2.7797462 0.96136767 0.83265859 0.33901330
0
Proportion of Variance 0.8168525 0.09770401 0.07329381 0.01214974
0
```

```
Cumulative Proportion  0.8168525 0.91455646 0.98785026 1.00000000
1



Loadings:

     Comp.1 Comp.2 Comp.3 Comp.4 Comp.5

[1,]  0.650 -0.106  0.585 -0.155 -0.447

[2,]  0.366        -0.593  0.559 -0.447

[3,] -0.096  0.470 -0.330 -0.679 -0.447

[4,] -0.386 -0.787        -0.150 -0.447

[5,] -0.534  0.383  0.434  0.425 -0.447
```

所以第一列的结果是 0.7,0.4,-0.1,-0.4,-0.5, 这样就和原来 paper 一样了。感谢宋绍铭同学在检查这个 BUG 时提供的帮助和思路。

检验所得到的本征向量是否符合正交基的特点：

```
test<-as.data.frame(0)


for(i in 1:5)
  for(j in 1:5)
    test[i,j]=sum(vector[,i]*vector[,j])
```

得到数据如下

```
##        0      V2      V3      V4      V5
## 1  1.000000e+00 -1.040834e-17 -1.318390e-16  2.359224e-16 -5.551115e-17
## 2 -1.040834e-17  1.000000e+00  1.977585e-16  0.000000e+00 -5.551115e-17
## 3 -1.318390e-16  1.977585e-16  1.000000e+00 -1.387779e-17  3.816392e-16
## 4  2.359224e-16  0.000000e+00 -1.387779e-17  1.000000e+00 -1.665335e-16
## 5 -5.551115e-17 -5.551115e-17  3.816392e-16 -1.665335e-16  1.000000e+00
```

符合定义。

2° 生成匹配的 control 和 case

Adjustment of genotypes and phenotypes using axes of variation. Let g_{ij} be the genotype of individual j ($g_{ij} = 0$, 1 or 2) at SNP i, and let a_j be the ancestry of individual j along a given axis of variation. We define $g_{ij,\ adjusted} = g_{ij} - \gamma_i a_j$, where $\gamma_i = \Sigma_j a_j g_{ij} / \Sigma_j a_j^2$ is a regression coefficient for ancestry predicting genotype across individuals j with valid genotypes at SNP i. (If there are no missing genotypes at SNP i, then $\Sigma_j a_j^2 = 1$ by definition, and thus $\gamma_i = \Sigma_j a_j g_{ij}$.) A similar adjustment is performed for each axis of variation. The adjustment of phenotype p_j is analogous. We note that the procedure we have described is equivalent to using the axes of variation as covariates in a multilinear regression, but is simpler because the axes of variation are orthogonal, and thus the adjustments can be performed independently for each axis of variation.

第二步是不断调整基因型和表现型沿着新特征空间的可以被回归得到量，从而得到残差。即产生了成对匹配的 case 和 control。计算过程如下。“坐标轴”以论文中的[0.7,0.4, −0.1, −0.4, −0.5]为例，Candidate SNP 为[2,2,1,1,0]，先算得$\gamma_i = (\Sigma a_j g_{ij})/(\Sigma a_j^2)$ 在本例中$\gamma_i = 1.589$ 然后对于$g'_{ij} = g_{ij} - \gamma_i a_j$计算得到新的一列向量[0.9,1.4,1.1,1.6,0.8]除了第一个和原论文有所出入外，其余均一致。然后对于 Pheotype：[1,1,0,0,0]减去特征向量[0.7,0.4, −0.1, −0.4, −0.5]即得到向量[0.3,0.6,0.1,0.4,0.5].以上即为论文中那两列向量得到的计算过程。

3° 计算 Armitage trend χ^2值

Computation of Armitage trend χ^2 statistic. As discussed in ref. 10, the Armitage trend χ^2 statistic[29] is more appropriate than a χ^2 statistic obtained from a 2×2 allelic or 2×3 genotypic χ^2 table. The Armitage trend χ^2 statistic is equal to N times the squared correlation between genotype (0, 1 or 2) and phenotype (0 or 1), where N is the number of samples. Though we believe that $(N - 1)$ times the squared correlation is a more appropriate statistic, we used the original definition of Armitage in all of our calculations.

第三步即计算χ^2值.论文中计算的是 Armitage trend χ^2,在 R 中可用 prop.trend.test 函数完成。论文中说使用这种统计量比常见的从 2×2 allelic 或 2×3 genotype 的χ^2表中得到的统计量更加科学。由于对于这块的统计学知识不太了解,无法给出具体的计算步骤了。