sojo Tutorial

Zheng Ning

5/19/2017

What is SOJO?

SOJO stands for **S**election **O**perator for **JO**intly analyzing multiple variants. It is a tool for implementing Least Absolute Shrinkage and Selection Operator (LASSO) using GWAS summary statistics. LASSO was introduced and applied to variable selection problems in various disciplines because of its better interpretability and prediction accuracy. More specifically, Instead of only considering the square loss function $\frac{1}{2}\|\mathbf{y}-\mathbf{X}\hat{\beta}\|_2^2$, LASSO takes the ℓ_1-norm regularization $\|\hat{\beta}\|_1$ into account and solves

$$\min_{\hat{\beta}\in R^p} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\hat{\beta}\|_2^2+\lambda\|\hat{\beta}\|_1,$$

where the tuning parameter $\lambda \geq 0$. The regularization term makes LASSO allow large coefficients only when they lead to substantially better fit. The LASSO result at any tuning parameter can be approximated by using (i) the covariance structure between variants and the trait and (ii) the LD structure between variants. The method is described in:

Ning et al. (2017) A selection operator for summary association statistics reveals locus-specific allelic heterogeneity of complex traits. *Submitted*

What data are required?

SOJO needs genome-wide association study (GWAS) summary statistics for a trait. The data frame of GWAS summary statistics should contain columns for variant names (column name `SNP`), effect alleles (column name `A1`), reference alleles (column name `A2`), allele frequencies of A1 (column name `Freq1`), effect sizes (column name `b`), standard errors (column name `se`), and sample sizes (column name `N`). An example data frame is given later.

Installation

Run the following command in R to install the **sojo** package:

For Windows users, please download and install the R package for

For Mac and Linux users, please download the source .tar.gz and install via:

```
R CMD INSTALL sojo_1.0.tar.gz
```

in your terminal.

In R, load the package via:

```
library(sojo)
```

or

```
require(sojo)
```

LASSO using Summary Statistics

Example Summary Statistics

You need to load a data frame of GWAS summary statistics for a trait into your working directory. Let us demonstrate this via an example included in the package. Here, we have the summary statistics for height across 963 variants around rs11090631 on chromosome 22. The top of the summary statistics file looks like:

```
data(sum.stat.raw)
head(sum.stat.raw)
```

```
##          SNP A1 A2 Freq1       b     se      N
## 1  rs1022622  C  G 0.942  0.0072 0.0050 241337
## 2  rs2024708  A  G 0.833  0.0052 0.0038 253216
## 3  rs2073239  A  G 0.833  0.0051 0.0038 253248
## 4  rs5765102  A  G 0.167 -0.0061 0.0037 252230
## 5  rs5766231  A  G 0.833  0.0052 0.0038 253216
## 6 rs11703912  A  G 0.833  0.0052 0.0038 252265
```

A Simple sojo analysis

Once the data are successfully loaded and all necessary columns are present, the LASSO solution can be computed by:

```
res <- sojo(sum.stat.raw, chr = 22, nvar = 20)
```

The result is a list with two sub-objects `$lambda.v` and `$beta.mat`. By setting `nvar = 20`, the computation stops when the model include 20 variants with non-zero coefficients. The vector of lambdas when a variant is added into or removed from the model can been seen by:

```
res$lambda.v
```

```
##  [1] 0.012017 0.011131 0.009834 0.009622 0.009194 0.007176 0.006401
##  [8] 0.005839 0.005157 0.004852 0.004730 0.004518 0.004410 0.004284
## [15] 0.004102 0.004016 0.003913 0.003896 0.003830 0.003814 0.003618
## [22] 0.003569 0.003531 0.003522 0.003510
```

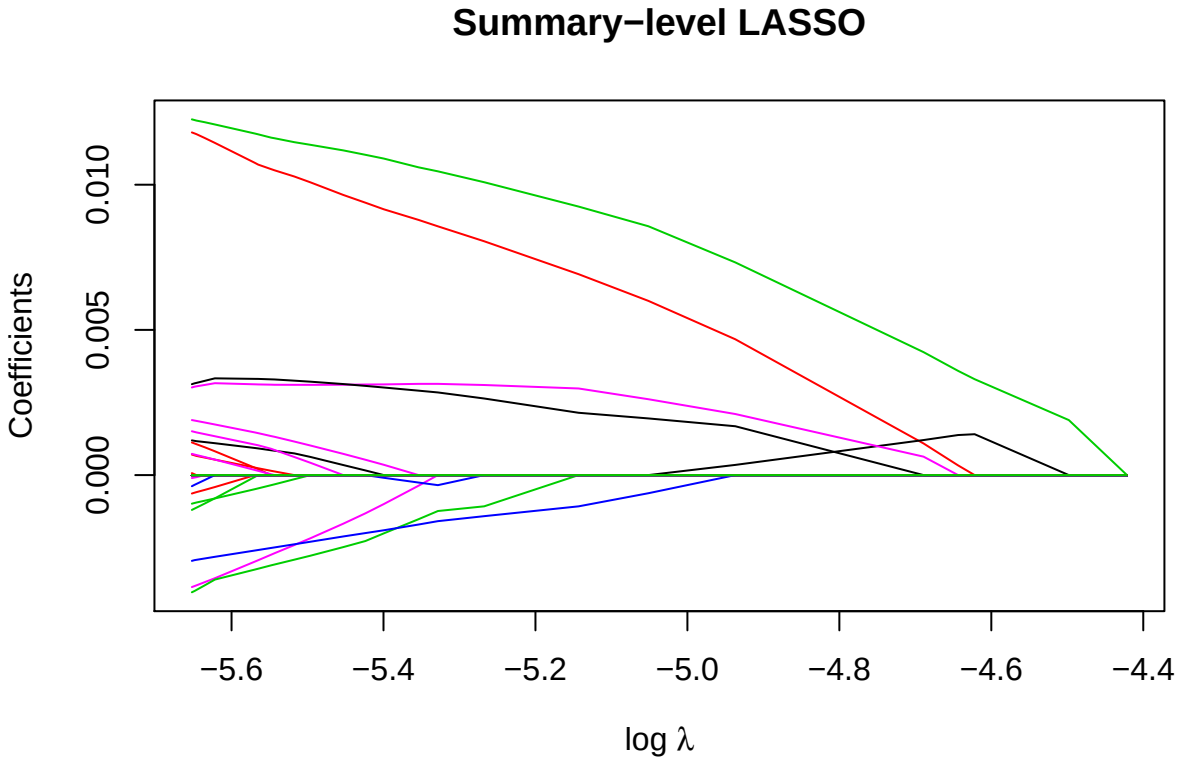
We can check the LASSO solutions for some variants at lambdas among `lambda.v`:

```
snp_selected <- which(res$beta.mat[,5] != 0)
res$beta.mat[snp_selected,1:4]
```

```
## 5 x 4 sparse Matrix of class "dgCMatrix"
##           beta      beta       beta      beta
## rs2142661    . .         -6.986e-18 3.330e-04
## rs9614670    . 2.596e-17  1.408e-03 1.379e-03
## rs6006753    . 1.894e-03  3.302e-03 3.589e-03
## rs714022     . .          .         1.920e-18
## rs8141212    . .          .         .
```

The LASSO path plot can be obtained by:

```
matplot(log(res$lambda.v), t(as.matrix(res$beta.mat)), lty = 1, type = "l",
    xlab = expression(paste(log, " ", lambda)), ylab = "Coefficients", main = "Summary-level LASSO")
```



LASSO solution at some specific tuning parameters can also be computed via:

```
res2 <- sojo(sum.stat.raw = sum.stat.raw, chr = 22, lambda.vec = c(0.004,0.002))
```

For Help

For direct R documentation of `sojo` function, you can simply use question mark in R:

```
?sojo
```

If you have specific questions, you may email the maintainer of sojo via zheng.ning@ki.se.