PEN Package Example

Ting-Huei Chen

June 13, 2014

PEN: a penalized estimation method for binary and continuous response variables.

1 Overview

```
> library(PEN)
```

This penalized estimation method is motivated by multiple loci mapping for gene expression and case-control study. This vignette describes how to use PEN based on the two example dataset for those studies.

2 PEN: Examples Using Simulated Data

2.1 Example for case-control study

We first illustrate the usage of PEN by a simulated data. Specifically, we simulated 1,100 SNPs on chromosomes 1 to 22 with 50 SNPs on each chromosome by GWAsimulator (**?**). Among them, 3 disease loci (3 SNPs) are located on chromosomes 2, 11, and 22. The three randomly selected SNPs have minor allele frequencies, 0.24, 0.29 and 0.30 respectively. For the case-control disease model, the parameters setting for the simulated data are summarized in the table 1. Let RR1 be the genotypic relative risk of the genotype with one copy of the risk allele versus that with zero copy of the risk allele and RR2 be the genotypic relative risk of the genotype with two copies of the risk allele versus that with zero copy of the risk allele. In addition, the disease prevalence

Table 1: Parameter setting for GWAsimulator

SNP's chromosome index	RR1	RR2
2	3.0	3.0
11	1.5	2.0
22	2.0	3.0

is set as 0.05, and the sample size is 300 with 150 cases and controls respectively.

To conduct the analysis, first, load the data into R workspace, and check the dimensions and a few rows of the data.

```
> data(XData)
> data(XInfo)
> dim(XData)

[1]  300 1101

> mode(XData)

[1] "numeric"

> data(XIndex)
```

For `XData`, the mode should be numerical matrix format with the column names as covariates indexes, and rows names as samples. The indexes of covariates start from non-genotypic variables and followed by genotypic data such as SNPs, which are ordered in genomic positions. `XInfo` is a numerical vector matched to the covariates indexes of `XData` with 0 to index non-genotypic covariates and 1 to 22 autosomal chromosomes and 23 for X chromosomes. The index vector `XIndex` with length of number of covariates consists 0 and 1 to index covariates to be unpenalized and penalized respectively. In this example, the first covariate is age, which is considered to be an unpenalized variable. Therefore, the first element of `XInfo` and `XIndex` is assigned to be 0.

For case-control study example, assume additive effect of the copy of the risk allele and the logistic model on the disease status y, $y = 1$ for cases and $y = 0$ for controls. Let X denote the `XData` matrix, α be the intercept, and β be the coefficient matrix. The logistic regression model is

$$\log\{\Pr(y = 1|X)/\Pr(y = 0|X)\} \quad = \quad \alpha + X\beta.$$

Next, load the response vector `Y`.

```
> data(Y)
```

To obtain the estimates of the coefficient matrix, run the R function `PEN`. The following illustration is based on SICA penalty and default chromosomal updating order, from chromosomes 1 to 22 for the iterative coefficient estimation procedure.

```
> out <- PEN( X=XData,  y=Y, family=c("binomial"), penalty
+ =c("SICA"), ChooseXindex=XIndex)
> out$beta[which(out$beta!=0)]

NULL
```

Next, a regular logistic regression is applied to the response and the selected SNPs to obtain the p-values corresponding to the regression coefficients for those SNPs. A comparison to the p-values obtained by univariate logistic regression on each of 1100 SNPs and the true simulated disease loci position is summarized in figure 1.

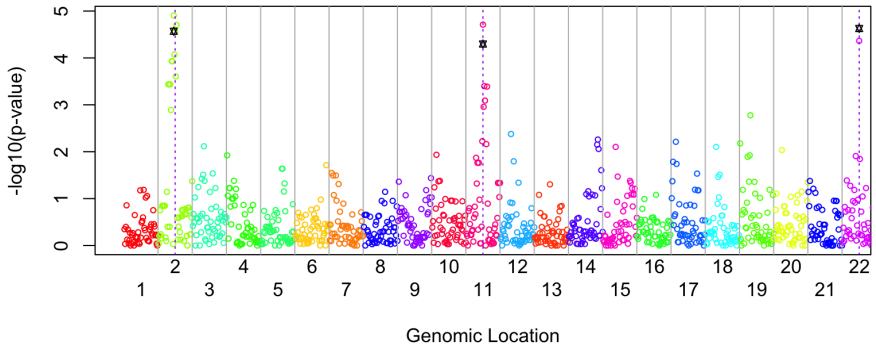

Figure 1: The manhattan plot of the $-\log_{10}$p-values obtained by univariate logistic regression on each of 1100 SNPs. The three purple lines denote the three simulated true disease loci, and the three black star symbols denote the $-\log_{10}$p-values obtained by the regular logistic regression on the joint three selected SNPs by the penalized estimation approach.

2.2 Example for gene expression study

For the example of gene expression study, the genotype matrix is the same as that of the case-control example, and three SNPs are randomly picked to be the causal ones. The the gene expression response vector, `Y_g` is simulated by linear model with coefficients vector $(0.5, -0.5, 0.5)$ and residuals e following standard normal distribution.

$$y_g \quad = \quad \alpha + X\beta + e.$$

Similarly, user can run the R function `PEN` to obtain the estimates of the coefficient matrix.

```
> data(Y_g)
> out <- PEN( X=XData,  y=Y_g, family=c("gaussian"), penalty
+ =c("SICA"), ChooseXindex=XIndex)
> out$beta[which(out$beta!=0)]
```

```
NULL
```

A comparison to the p-values obtained by univariate linear regression on each of 1100 SNPs and the true simulated causal loci position is summarized in figure 2.

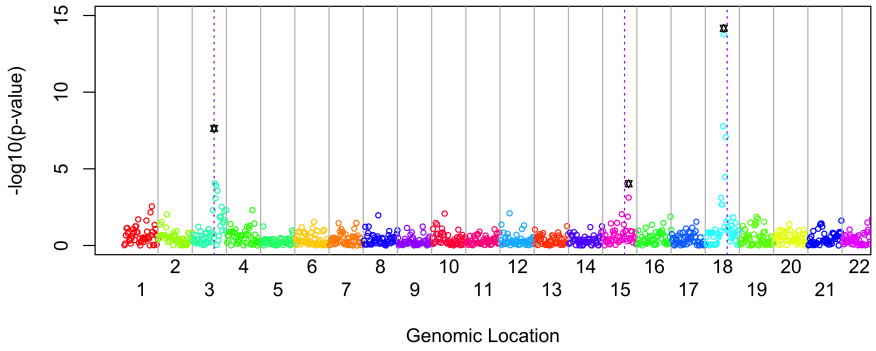

Figure 2: The manhattan plot of the $-\log_{10}$p-values obtained by univariate linear regression on each of 1100 SNPs. The three purple lines denote the three simulated true disease loci, and the three black star symbols denote the $-\log_{10}$p-values obtained by the regular linear regression on the joint three selected SNPs by the penalized estimation approach.

References

Li, C. and Li, M. (2008). Gwasimulator: a rapid whole-genome simulation program. *Bioinformatics*, **24**(1), 140–142.