Simulation_summary

Xinru Wang

6/16/2020

Objective.

To effectively analyze different variable selection methods for high dimensional data, we have to generate a dataset containing a combination of **strong**, **WAI**, **WBC**, and **Null** predictors. The signals were created using the following criterias:

1. **Strong signals**

$$S_{strong} = j : |\beta_j| > c\sqrt{\frac{log(p)}{n}}, \text{ for some c} > 0, 1 \leq j \leq p$$

2. **Weak-and-independent (WAI)**

$$S_{WBC} = j : |\beta_j| \leq c\sqrt{\frac{log\,(p)}{n}}, \text{ for some c} > 0, corr(X_j, X_j') \neq 0$$

3. **Weak-and-correlated (WBC)**

$$S_{WBC} = j : |\beta_j| \leq c\sqrt{\frac{log\,(p)}{n}}, \text{ for some c} > 0, corr(X_j, X_j') = 0$$

4. **Null signals**

$$S_{null} = j : \beta_j = 0, 1 \leq j \leq p$$

The general idea of generating data

1. Create a 50x50 positive-definite variance-covariance matrix with WBC variables being correlated to the first strong predictor with $corr(X_j, X_j') = 0.3$
2. Generate a multivariate normal distribution with mean 0 and sigma equal to the variance-covariance matrix generated in step one.
3. The matrix of true coefficient values was created with the strong signals set to 5 and the weak predictors (WAI and WBC) set to the threshold value defined by the changing c value.
4. Finally, get the linear response Y values:

$$Y = 1 + X\beta + \epsilon$$

```r
# generate_data function
generate_data = function(n = 1000, c =1, correlation = 0.3,
                         strong_coeff = strong_coeff,
                         strong_num = st_weak_num[k,1],
                         wai_num = st_weak_num[k,2],
                         wbc_num = st_weak_num[k,3],
                         num_p = 50){

  null_num = num_p - wbc_num - wai_num - strong_num

  ## the coeff for wai and wbc
  threshold = c * (log(num_p)/n)^0.5

  ## the condition that cor(Xj,Xj')=correlation for j' belongs to strong signals
  matrix = diag(num_p)
  matrix[1, (strong_num + wai_num + 1):(strong_num + wai_num + wai_num)] = correlation
  matrix[(strong_num + wai_num + 1):(strong_num + wai_num + wai_num), 1] = correlation

  ## generate data from multivariate normal
  X = mvrnorm(n = n, mu = rep(0, num_p), Sigma = matrix, empirical = F, tol = 0.1)

  ## set the coeffecient beta
  b_true = c(
    rep(strong_coeff, strong_num),
    rep(threshold, wai_num),
    rep(threshold, wbc_num),
    rep(0, null_num)
  )

  ## get response Y
  Y = 1 + X %*% b_true + rnorm(n)
  data = as_tibble(data.frame(cbind(X, Y)))
```