Knockoff and its Applications

Spandan Ghoshal and Debarshi Chakraborty

Indian Statistical Institute, Kolkata

April 27, 2023

Contents

Introduction

Problem Statement

Methodology

Theoritical Results

Choice of Tuning Parameter

Simulations and Experiment Results

Further Improvements and Applications

References

Introduction

- A very common problem that we encounter in various statistical problems, especially in statistical learning, is the problem of **variable selection**
- For example, identifying the important/significant predictors in a supervised learning problem of regression or classification
- Also, we are interested to control the number of false discoveries at some certain level (FDR/FWER control etc)
- **Knockoff** is a powerful,versatile, data driven framework for the above situation of controlled variable selection

Problem Statement

- Suppose we have a response variable Y and p potential predictors $X_1, X_2, ..., X_p$. Given n i.i.d samples from $(X_1, X_2, ..., X_p, Y)$, we would like to know which predictors are important to make inferences about the response
- This problem is motivated by the idea that in many practical problems, $F(Y|X = x)$ depends only on a small subset $S \subseteq \{X_1, ..., X_p\}$ of the predictors. In other words
$$Y \perp\!\!\!\perp \{X_j\}_{j \notin S} | \{X_j\}_{j \in S}$$
- We say that the j^{th} variable is a null variable if and only if Y is conditionally independent of X_j given others. Let H_0 denote the set of all null variables
- For any variable selection method that selects a subset S of the predictors, **the false discovery rate** (FDR) is given by
$$\text{FDR} = E[\tfrac{|S \cap H_0|}{max(1, |S|)}]$$
- Goal is to discover as many significant variables as possible while keeping the false discovery rate (FDR) under control

Methodology

- Suppose we have a linear model of the form $y = X\beta + \epsilon$
- X is the $n \times p$ data matrix
- y is the $n \times 1$ vector of responses
- β is the vector of unknown coefficients
- $\epsilon_{n\times 1} \sim N(0, \sigma^2 I)$
- The key idea behind knockoffs is creating a **negative control group** for the predictors that behaves in the same way as the original null variables but, unlike them, is known to be null

Methodology

Intuitive idea of constructing knockoffs

For each observation of a predictor X_j, we construct a copy of it, say $\tilde{X}_j$ in such a manner that

- The correlation between $\tilde{X}_j$ and $\tilde{X}_k$ (for $j \neq k$) is same as the correlation between X_j and X_k
- The correlation between X_j and $\tilde{X}_k$ (for $j \neq k$) is same as the correlation between X_j and X_k
- The copies are created without looking at the response y

Working Principle

- By construction, the knockoff copies are not important for the response, in other words they are null predictors
- Hence, the importance of a predictor X_j can be deduced by comparing its predictive power for y to that of its knockoff copy $\tilde{X}_j$, which in essence works as a negative control for the original explanatory variable

Methodology

Comparing Importance

- Once we get the knockoffs, apply any traditional variable selection method (say LASSO) on the augmented set of explanatory variables $(X_1, ..., X_p, \tilde{X}_1, ...\tilde{X}_p)$
- Then we can compute some measure of variable importance in predicting y and compare between original variable and its copy
- For example, we can compare between $|\hat{\beta}_j|$ and $|\hat{\beta}_{j+p}|$ $\forall$ $j = 1, 2, ..., p$ and then select only those variables which are ”clearly”better than their corresponding knockoff copies
- the above was just an example, the knockoff procedure is not restricted to the LASSO coefficients, any statistic (say Z_j) that captures variable importance can be used
- Once we get $\{Z_j\}_{j=1}^{p}$, different contrast functions can be used to compare the knockoff copy to the original variable, in general we choose anti symmetric functions
- For the simplest of cases, one can use $W_j = Z_j - \tilde{Z}_j$, certainly many other alternatives are there. Clearly, idea is to select those variables which have higher values of W_j, for which we must set an appropriate threshold

Methodology

Now we can build up the mathematical details of the knockoff filter .The entire procedure can be carried out in 3 steps.

Step 1: Constructing the Knockoffs

- For each predictor X_j (i.e. the j^{th} column of the design matrix) construct a copy $\tilde{X}_j$ such that

$$\tilde{X}^T\tilde{X} = \Sigma \text{ and } X^T\tilde{X} = \Sigma - \text{ diag}(s)$$

  where $\Sigma = X^T X$, after normalising such that $||X_j||_2^2 = 1 \forall j = 1, 2, .., p$
- It is easy to check that the following conditions hold
  1. $X_j^T\tilde{X}_k = X_j^T X_k$
  2. $X_j^T\tilde{X}_j = \Sigma_{jj} - s_j = 1 - s_j$
  3. $\tilde{X_j}^T\tilde{X}_j = X_j^T X_j = 1$
- One basic strategy to construct knockoffs is
  1. Choose $s \in \mathbb{R}_p^+$ such that diag$(s) \preccurlyeq 2\Sigma$
  2. Set $\tilde{X} = X(I - \Sigma^{-1}\text{diag}(s)) + UC$

  where $U \in \mathbb{R}_{n\times p}$ is a orthonormal matrix that is orthogonal to the span of the features of X and C is a Cholesky decomposition of the Schur complement of the Gram matrix of $[X, \tilde{X}]$

Methodology

Step 2: Calculating statistics to compare original and knockoff variables

- As mentioned earlier, we need an existing variable selection method. For familiarity, we will use LASSO.
- LASSO will return the coefficient vector

  $$\hat{\beta}=\text{argmin}_{\beta}(\tfrac{1}{2}||y - X\beta||_2^2 + \lambda||\beta||_1)$$

  as a function of the penalty term λ
- Z_j =sup$\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}$
- Run LASSO on $[X, \tilde{X}]$ and obtain $(Z_1, ..., Z_p, \tilde{Z}_1, ...\tilde{Z}_p)$
- Define

  $$\mathrm{W}_j = \begin{cases} Z_j & \text{if } Z_j > \tilde{Z}_j \\ -\tilde{Z}_j & \text{if } Z_j < \tilde{Z}_j \end{cases}$$
- Note that, large positive values of W_j give indication of a true signal
- Only task left is how to determine W_j is "large"or not, we need an appropriate threshold

Methodology

Since this is a completely data driven procedure, we will use a data driven threshold.

Step 3: Calculating data driven threshold for W_j

- A data-dependent threshold T to allow for performing the final variable selection (with the FDR control guaranteed) is given by
$$T_\alpha = min[t \in W : \frac{\#\{j : W_j \leqslant -t\}}{\#\{j : W_j > t\} \vee 1} \leqslant \alpha]$$
or $T = \infty$ is the set is empty (for a particular level α)
- Here $W = \{W_j : |W_j| \, \forall j = 1, 2, ..., p\} \backslash \{0\}$, the set of unique non zero $|W_j|$'s

In a nutshell, the knockoff procedure can be written as follows :

1. Construct the knockoff design matrix $\tilde{X}$, the test statistics W_j and the threshold T by the steps mentioned above
2. Select the set of variables $S = \{X_j : W_j \geqslant T\}$

Theoritical Results

Theorem

For any $\alpha \in (0,1)$, *the above knockoff method satisfies*

$$E[\frac{\#\{j:\beta_j=0,\ j\in S\}}{\#\{j:\beta_j=0\}+\alpha^{-1}}] \leqslant \alpha$$

where the expectation is taken over gaussian noise of the model, treating X *and* $\tilde{X}$ *as fixed.*

The "modified FDR" bounded by this theorem is very close to the FDR in settings where a large number of features are selected (as adding α^{-1} in the denominator then has little effect), but often we prefer to control the FDR exactly. To achieve that, a slightly more conservative procedure is proposed.

Theorem

If we use the same knockoff method but with a higher threshold

$$T_\alpha = min[t \in W : \frac{1+\#\{j:W_j\leqslant -t\}}{\#\{j:W_j>t\}\vee 1} \leqslant \alpha]$$

then FDR $\leqslant \alpha$*.This is also called the* ***Knockoff*** $+$ *method.*

Some Comments on the choice of test statistic and tuning parameter

- The strategy we had pointed out in the first step of constructing knockoffs gives us the desired correlation structure as follows
$$[X, \tilde{X}]^T [X, \tilde{X}] = \begin{bmatrix} \Sigma & \Sigma - diag\{s\} \\ \Sigma - diag\{s\} & \Sigma \end{bmatrix}$$
- A necessary and sufficient condition for $\tilde{X}$ to exist is the above block matrix should be positive semi definite $\iff$ $diag\{s\} \geqslant 0$ and $2\Sigma \geqslant diag\{s\}$
- Now, we are to choose the vector s with the above constraints
- Now let us remember our intuition, if X_j truly belongs to the model, we want it to enter before its knockoff copy $\tilde{X}_j$, for which we need to make the correlation between the knockoff and the true signal to be small, so that $\tilde{X}_j$ does not enter the LASSO model very early
- In other words, we want to make any variable and its knockoff to be as orthogonal as possible, which essentially implies maximising $s_j \forall j$ under the given constraints

Some Comments on the choice of test statistic and tuning parameter

- One strategy is to choose $s_j = \text{min}\{2\lambda_{min}(\Sigma), 1\} \ \forall j = 1, 2, ..., p$ so that all the correlations take the same value, the corresponding knockoffs are known as **Equicorrelated Knockoffs**
- Another possibility is to select knockoffs so that the average correlation between an original variable and its knockoff is minimum. This can be achieved by solving the convex optimization problem

$$\text{minimize } \sum_j |1 - s_j| \text{ subject to } 0 \leqslant s_j \leqslant 1 \text{ and } diag\{s\} \leqslant 2\Sigma$$

- Note that, if we are able to construct the knockoff matrix $\tilde{X}$ in such a way that the s_j are high, then we are more likely achieve higher power

Some Comments on the choice of statistic

- Note that, the proposed method is useful (power is high) only if the true variables X_j enter before the knockoffs $\tilde{X}_j$ do.
- Otherwise, because we cut off and only select the first consecutive k variables, we will be selecting few of the true variables, hence low power.
- Hence it is a desirable trait for X_j to enter before the corresponding knockoff $\tilde{X}_j$.
- Thus we want a near-zero correlation (near-orthogonality) between the original variable X_j and its knockoff $\tilde{X}_j$.

$$X_j \underset{\sim}{\perp} \tilde{X}_j, \text{ for } j = 1, \ldots, p$$

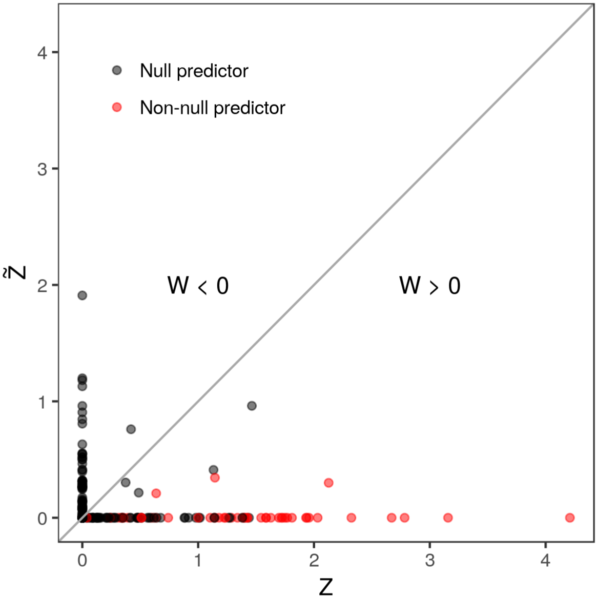
- The intuition is that, if there is a true relationship between the original variable X_j and $\boldsymbol{y}$, then by construction $\tilde{X}_j$ should not have relation with $\boldsymbol{y}$.
- On the other hand if X_j is a null variable then there is no reason why either X_j & $\tilde{X}_j$ should be any more related to $\boldsymbol{y}$ than the other.
- These are good properties for our selection procedure! Hence, we want:

$$\boldsymbol{X}_j^T \tilde{\boldsymbol{X}}_j = 1 - s_j \approx 0$$

Some Comments on the choice of test statistic and tuning parameter

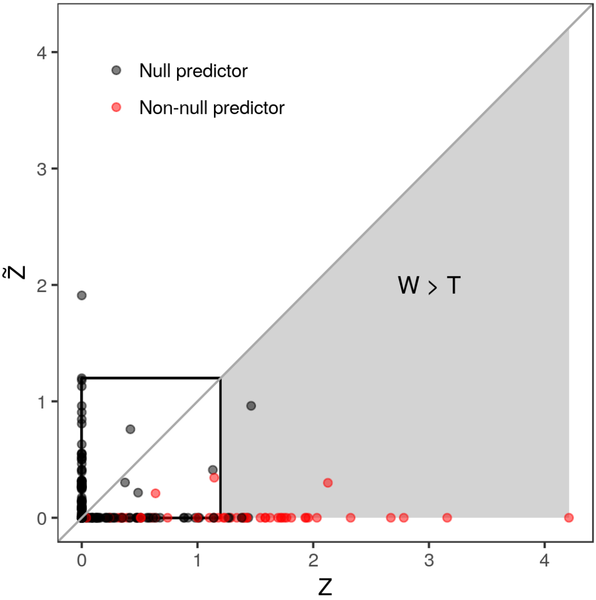
- Despite the simple example that we presented, the knockoffs procedure is by no means restricted to statistics based on the lasso, as many other options are available for assessing the importance of X_j and $\tilde{X}_j$.
- In general, it is required that the method used to compute the Z_j and $\tilde{Z}_j$'s satisfy a fairness requirement, so that swapping X_j and $\tilde{X}_j$ would only have effect of swapping Z_j and $\tilde{Z}_j$.
- Once the Z_j and $\tilde{Z}_j$'s have been computed, different contrast functions can be used to compare them. In general, we must choose an anti-symmetric function h and we compute the symmetrized knockoff statistics $W_j = h\left(Z_j, \tilde{Z}_j\right) = -h\left(Z_j, \tilde{Z}_j\right)$, such that $W_j > 0$ indicates X_j that appears to be more important than its own knockoff copy. Another simple example may be $W_j = Z_j - \tilde{Z}_j$, but many other alternatives are possible.

Simulations and Experiment Results



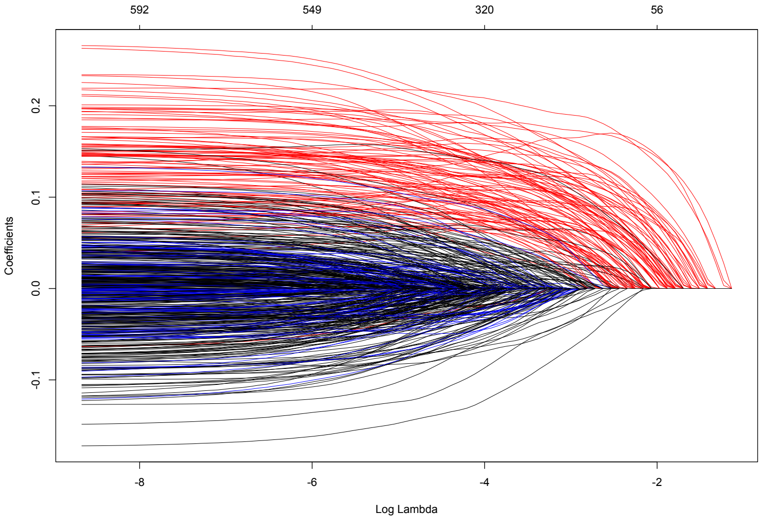
Where do the important variables lie?

Simulations and Experiment Results



The Critical Region

Simulations and Experiment Results

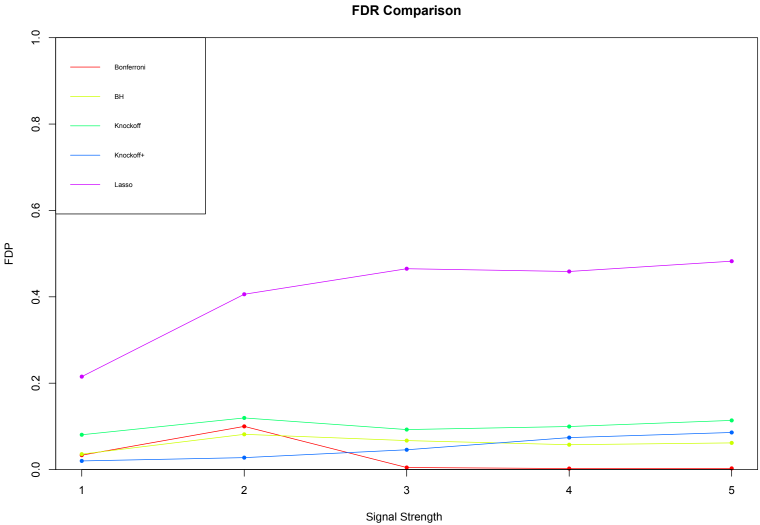


The Knockoff Path

Simulation and Experiment Results

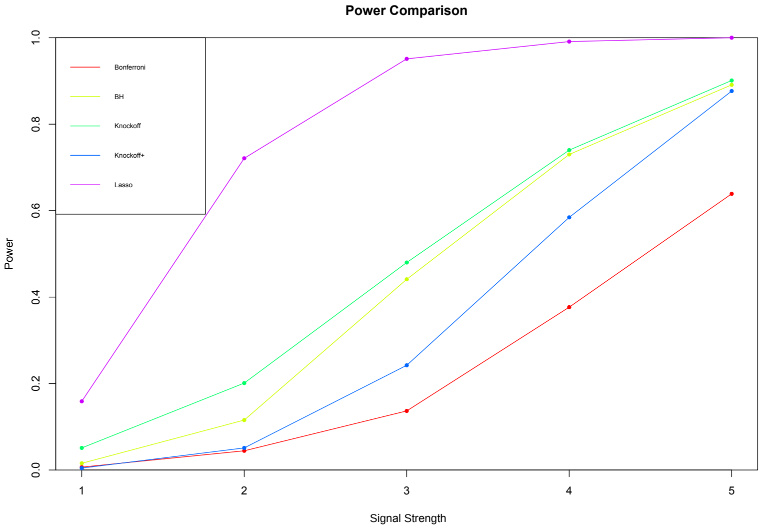
- We generate to data from model $\boldsymbol{y} \sim N\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 I_k\right)$ with $\boldsymbol{y} \in \mathbb{R}^{200}$ and $\boldsymbol{X} \in \mathbb{R}^{200\times 80}$ where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{30}, 0, \ldots, 0)$ where each $\beta_i \in \{-b, b\}$ randomly. Then we use different methods to detect the important variables and calculate the FDP and Power by repeating the simulation around 50 times.
- We plot the obtained values against different choices of $b \in \{1, 2, 3, 4, 5\}$. This values of b basically denotes the strength of the signal.

Simulations and Experiment Results



FDR comparison of different methods

Simulations and Experiment Results



Power comparison for the methods under the same scenario

Analysis of HIV Drug Resistance Data

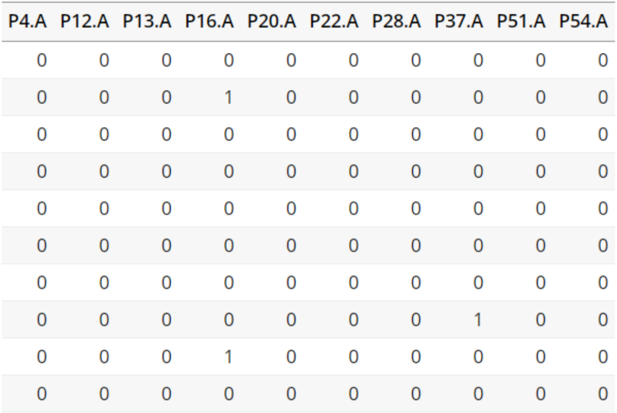
- We illustrate the analysis of a real (non-simulated) data set.
- To be specific, the scientific goal is to determine which mutations of the Human Immunodeficiency Virus Type 1 (HIV-1) are associated with drug resistance.
- The data set consists of measurements for three classes of drugs: protease inhibitors (PIs), nucleoside reverse transcriptase (RT) inhibitors (NRTIs), and nonnucleoside RT inhibitors (NNRTIs).
- Protease and reverse transcriptase are two enzymes in HIV-1 that are crucial to the function of the virus. This data set seeks associations between mutations in the HIV-1 protease and drug resistance to different PI type drugs, and between mutations in the HIV-1 reverse transcriptase and drug resistance to different NRTI and NNRTI type drugs.
- We modify the dataset to give it the desired structure for our analysis

Understanding the Data Structure

- The features (columns of $\boldsymbol{X}$) are given by mutation/position pairs. Define

$$X_{i,j} = \begin{cases} 1 & \text{if the } i\text{th patient has the } j\text{th mutation/position pair and} \\ 0 & \text{otherwise} \end{cases}$$

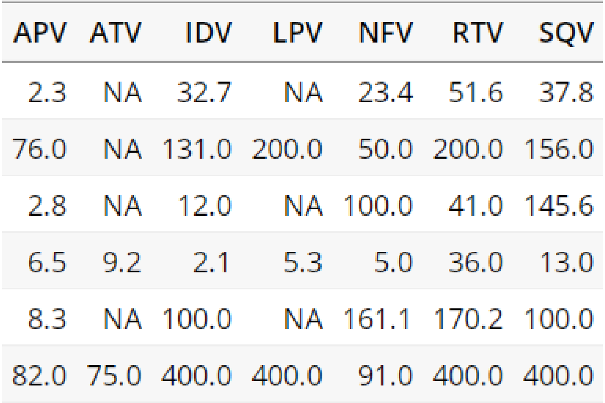
and $Y_{i,j}$ = resistance of patient i to drug j. These is an excerpt of the design matrix :-



P4.A	P12.A	P13.A	P16.A	P20.A	P22.A	P28.A	P37.A	P51.A	P54.A
0	0	0	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0
0	0	0	1	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0

Understanding the Data Structure

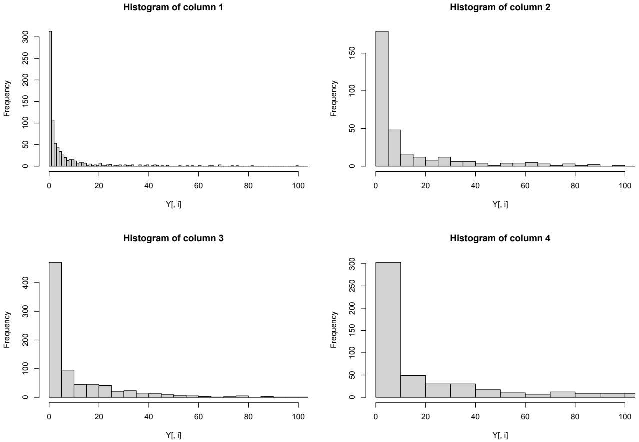
- The response matrix looks like:



APV	ATV	IDV	LPV	NFV	RTV	SQV
2.3	NA	32.7	NA	23.4	51.6	37.8
76.0	NA	131.0	200.0	50.0	200.0	156.0
2.8	NA	12.0	NA	100.0	41.0	145.6
6.5	9.2	2.1	5.3	5.0	36.0	13.0
8.3	NA	100.0	NA	161.1	170.2	100.0
82.0	75.0	400.0	400.0	91.0	400.0	400.0

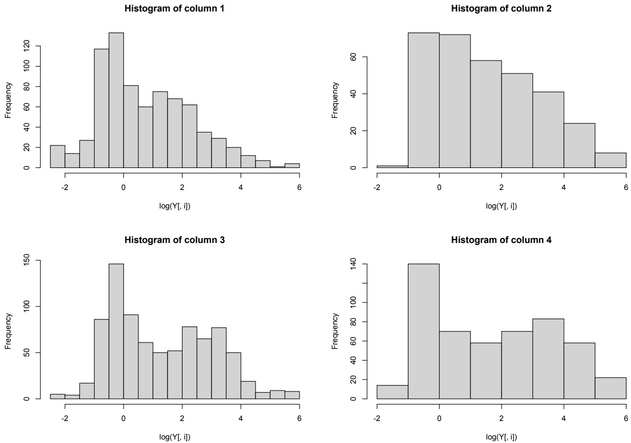
Transforming the response

- The knockoff filter is designed to control the FDR under Gaussian noise. A quick inspection of the response vector shows that it is highly non-Gaussian.



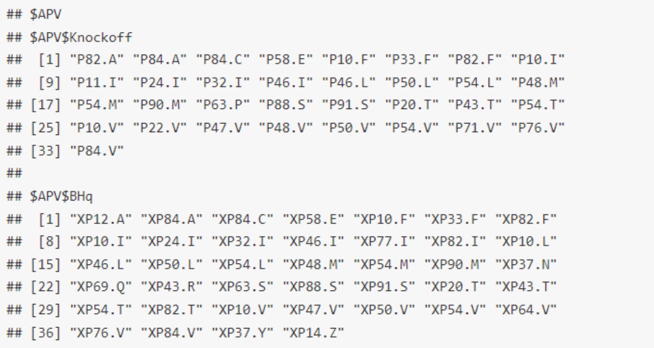
Transforming the response

- Hence we take the log transform as it seems to help considerably, so we will use the log-transformed drug resistancement measurements below.



Variable Selection using Knockoff & BH for comparison

- We now run the knockoff filter on each drug separately. We also run the Benjamini-Hochberg (BHq) procedure for later comparison.
- For example here are the selected variables associated with the first drug:



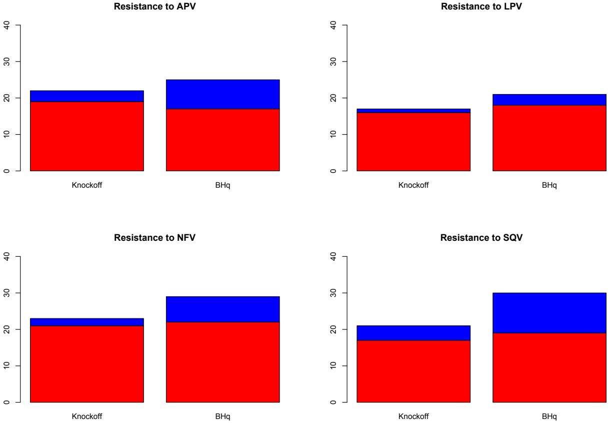
```
## $APV
## $APV$Knockoff
##  [1] "P82.A" "P84.A" "P84.C" "P58.E" "P10.F" "P33.F" "P82.F" "P10.I"
##  [9] "P11.I" "P24.I" "P32.I" "P46.I" "P46.L" "P50.L" "P54.L" "P48.M"
## [17] "P54.M" "P90.M" "P63.P" "P88.S" "P91.S" "P20.T" "P43.T" "P54.T"
## [25] "P10.V" "P22.V" "P47.V" "P48.V" "P50.V" "P54.V" "P71.V" "P76.V"
## [33] "P84.V"
##
## $APV$BHq
##  [1] "XP12.A" "XP84.A" "XP84.C" "XP58.E" "XP10.F" "XP33.F" "XP82.F"
##  [8] "XP10.I" "XP24.I" "XP32.I" "XP46.I" "XP77.I" "XP82.I" "XP10.L"
## [15] "XP46.L" "XP50.L" "XP54.L" "XP48.M" "XP54.M" "XP90.M" "XP37.N"
## [22] "XP69.Q" "XP43.R" "XP63.S" "XP88.S" "XP91.S" "XP20.T" "XP43.T"
## [29] "XP54.T" "XP82.T" "XP10.V" "XP47.V" "XP50.V" "XP54.V" "XP64.V"
## [36] "XP76.V" "XP84.V" "XP37.Y" "XP14.Z"
```

Evaluating the results

- In this case, we actually have a “ground truth” obtained by another experiment. Using this, we can compare the results from the knockoff and BHq procedures.
- Here is the table of fdp values obtained for the two methods for different drug responses :-

Drug	Knockoff_FDP	BHq_FDP
APV	0.136	0.32
ATV	0.258	0.259
IDV	0.387	0.333
LPV	0.059	0.143
NFV	0.087	0.241
RTV	0.296	0.308
SQV	0.19	0.367

Evaluating the results



Proportion of correctly detected predictors by two different methods

Further Improvements and Applications

- The above paradigm we discussed was for fixed model matrix, it can also be implemented when we consider the predictors as stochastic variables
- One limitation is the assumption $n > 2p$, but the same method can be improved for $2p > n > p$ case
- Can be used in structure estimation of Gaussian Graphical Models
- Controlled variable selection is particularly relevant in the context of Statistical Genetics

References

- The main paper : *"Controlling the False Discovery Rate via Knockoffs", Rina Foygel Barber and Emmanuel Candès. Ann. Statist. 43 (2015)*
- R Package Repository : *https://github.com/msesia/knockoff-filter*
- To view the codes and the output for the simulations, click here.