STAT41520 **(Always take notes)**

Final Project

Name: Weihao LI, Ming Gao, Zihao Zheng

1 Knockoff background

A plethora of variants of Knockoff has been developed since the original version is proposed. As an useful tool to control false discovery rate (FDR) and ensure true and replicable scientific research, it takes into account the relationship among potential explanatory variables and serves as a flexible framework, under which a wide class of statistical tests can be applied.

The goal of selective inference is to adjust for a valid p-value when selecting and estimating models, and then control for the portion of true positive. An important quantity to pay attention to is FDR, defined as the expectation of FDP

$$\text{FDR} = \mathbb{E}\left[\frac{\text{FP}}{\text{P} \vee 1}\right]$$

where P is the number of positive, FP is the number of false positive. Knockoff focuses on variable selection problem in regression setting where we have a response variable Y and d variables X and test which X_j has significant influence on response Y. The traditional permutation test does not account for the correlation among the variables, which might lead to false positives and severely underestimate the FDP. Therefore the basic idea of Knockoff is to create a "fake" version of design matrix $\tilde{X}$ to mimic the null distribution of X, then compare the difference between X_j and $\tilde{X}_j$ on Y and determine whether they are significant or not.

Algorithm 1 Knockoff workflow

Input: X, Y, α

1. Create knockoff $\tilde{X}$ by X and Y.
2. Calculate some importance statistic $I_j, \tilde{I}_j$ between Y and both X_j and $\tilde{X}_j$ for all $j = 1, \cdots, d$.
3. Compute $W_j = I_j - \tilde{I}_j$ for $j = 1, \cdots, d$, positive and large W_j means X_j is more important than $\tilde{X}_j$.
4. Compute a threshold T_α to control FDR below α. Then select the variables with $W_j > T_\alpha$.

The workflow of Knockoff can be found in algorithm 1. We can see each step of Knockoff is separated and flexible for any customized building block as long as they satisfy some requirements. To be more specific, the FDP is estimated by

$$\widehat{\text{FDP}}(t) = \frac{\sum_j \mathbb{1}\{W_j \leq -t\} + 1}{\sum_j \mathbb{1}\{W_j \geq t\}}$$

Then the choice of T_α is computed by $T_\alpha = \min\{t > 0 : \widehat{\text{FDP}}(t) \leq \alpha\}$. There have been plenty of extension of Knockoff on how to create the mimic $\tilde{X}$ like deep Knockoff and conditional Knockoff. Also, the comparison of behavior of Knockoff under different scenarios with existing FDR control approaches are fully investigated and understood. While the discussion on the choice of importance statistic is still interesting to researchers. Therefore, we conduct experiments on different importance statistic under different possible scenarios, and evaluate the performance of Knockoff in terms of FDR control and discovery Power where let TP be true positive and FN be false negative

$$\text{Power} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Knockoff has two basic versions, Fixed-X Knockoff and Model-X Knockoff. For Fixed-X Knockoff, X is treated as fixed covariates and the relationship of between X and Y should be known. For simplicity, Gaussian linear model $Y = X\beta + \epsilon$ is considered here and there is a neat form for knockoff $\tilde{X}$. For Model-X Knockoff, X is treated as random and subject to some joint distribution, which should be known in advance. In this way, we

can create the knockoff $\tilde{X}$ through the joint distribution. No assumption on the relationship between X and Y is needed. Both versions are examined in our experiments.

2 Importance statistic

We consider three sets of importance statistics from literature. Some of them are simple and easy to compute, some of them are based on certain machine learning algorithm therefore require fitting a model to implement. We have different expectations on the performance of them under different scenarios. The detailed simulation settings would be introduced in next section.

2.1 Correlation based

Correlation The basic importance statistic is the inner product of response and variable, defined as $W_j = |X_j^\top Y| - |\tilde{X}_j^\top Y|$, we named it as "Correlation" for simplicity. This measures the marginal dependency rather than partial dependency.

Spearman correlation Apart from linear correlation, the Spearman correlation is rank based and a suitable choice for measuring dependency in nonparametric setting, especially when the linear relationship between X and Y is violated but still maintains monotonicity. Let $W_j = |\rho(X_j, Y)| - |\rho(\tilde{X}_j, Y)|$ where

$$\rho(x, y) = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)} \qquad d_i = \text{rank}(x_i) - \text{rank}(y_i)$$

Kendall correlation Another widely considered correlation in non-parametric statistics is the Kendall correlation, which measures the ordinal association between response and variables. Let $W_j = |\tau(X_j, Y)| - |\tau(\tilde{X}_j, Y)|$ where

$$\tau(x, y) = \frac{2}{n(n-1)} \sum_{i<j} \text{sgn}(x_i - x_j)\text{sgn}(y_i - y_j)$$

2.2 Lasso based

Lasso is a classic variable selection method, which provides a rich class of importance statistics along with its estimate. Here we consider three of them. We first run Lasso on Y against $(X, \tilde{X})$. Denote the lasso estimate for X_j and $\tilde{X}_j$ to be $\hat{\beta}_j$ and $\tilde{\beta}_j$ respectively, and $\lambda_j(\tilde{\lambda}_j) = \sup\{\lambda : \hat{\beta}_j(\tilde{\beta}_j) > 0\}$ is when variable $X_j(\tilde{X}_j)$ first enter the Lasso path.

Lasso coefdiff $W_j = |\hat{\beta}_j| - |\tilde{\beta}_j|$ compares the scales of importance of the original version and knockoff, which behaves like regression coefficients taking partial correlation into account.

Lasso lambdadiff $W_j = \lambda_j - \tilde{\lambda}_j$ is another measure of how large regression coefficient is. The larger β_j is, the earlier it enters the Lasso path in term of pernalty λ.

Lasso lambdasmax $W_j = (-1)^{\mathbb{1}\{\lambda_j < \tilde{\lambda}_j\}} \max(\lambda_j, \tilde{\lambda}_j)$ is similar to the one above but more harsh on the output. We expect the Lasso based importance statistics have a very good performance in Gaussian linear model case but fail for highly correlated variables or other nonlinear cases.

2.3 Others

Some other importance statistics from literature can be applied for more general settings, like nonlinear relationship between X and Y. Here we consider two of them.

Random forest Random forest is a widely used nonparametric learning algorithm for classification or regression. Briefly it builds a lot of trees on given features X to approximate response Y by leveraging the fitted trees. The row and column samplings in the algorithm endows it with a byproduct which measures the importance of each variable, which is related to how many times the variable is used in building the trees and how much residual is reduced by introducing the variable. Denote the variable importance given by random forest as RF_j and $\widetilde{\text{RF}}_j$, then $W_j = \text{RF}_j - \widetilde{\text{RF}}_j$. We expect good performance from it when the relationship between X and Y are nonlinear.

Mutual information Mutual information between two variables X and Y is defined as

$$I(X;Y) = \mathbb{E}\log\frac{p(X,Y)}{p(X)p(Y)}$$

which measures the dependency between X and Y with information theoretic property. This is useful in fully nonparametric setting. One thing to bear in mind is

$$I(X;Y) = 0 \Longleftrightarrow X \perp\!\!\!\perp Y$$

Mutual information is well-defined in discrete case, although there is density version for continuous variables, we discretize the them to estimate the mutual information in our following experiments. Therefore $W_j = \hat{I}(X_j;Y) - \hat{I}(\tilde{X}_j;Y)$.

3 Simulation

In this section we present some simulation studies under different scenarios for knockoff and assumptions about relationship between Y and X. We use the following importance statistics for simulation: `lasso_coefdiff`, `lasso_lambdadiff`, `lasso_lambdasmax`, `correlation`, `kendall`, `spearman` and `random forest`. We didn't put results of using mutual information here because its power is always 0. The possible reason is that we need to discretize variables to implement `mutual_information` and a lot of information will be thrown away.

To analyze the performances of different importance statistics, we choose 5 levels of the number of true signal k ($k = 20, 40, 60, 80, 100$) and 2 levels of the magnitude of true signal β_M ($\beta_M = 1, 3.5$). We let $p = 200$, $n = 500 > 2p$. For each combination of k and β_M, we generate true coefficient $\beta \in R^p$ where β satisfies

1. $\|\beta\|_0 = k$.
2. $|\beta_j| = 0$ or β_M. for $j = 1, \ldots, p$.

We also generate $X \sim N(\mathbf{0}_n, \Sigma)$ and $\varepsilon \sim N(\mathbf{0}_n, \mathbf{I}_n)$, where Σ is an n×n Toeplitz matrix. Since we are interested in the performance of all importance statistics under different situations, such as linear model, non-linear model and generalized linear model, therefore we design several assumptions between X and Y in simulation studies and calculate Y using the assumptions we set. Then we follow Algorithm 1 to compute threshold T_α where $\alpha = 0.10$. Algorithm 1 returns $\hat{S} = \{j : W_j \geq T\} \subset \{1, \ldots, p\}$. In order to compare the performance of different importance statistics, we look into the FDP $= \frac{\#\{j:\beta_j=0 \text{ and } j\in\hat{S}\}}{\#\{j:j\in\hat{S}\}\vee 1}$ and power $= \frac{\#\{j:\beta_j\neq 0 \text{ and } j\in\hat{S}\}}{k}$ and take average over 100 repetitions of Algorithm 1.

3.1 Fixed-X setting and linear model

We assume the model to be $Y = X\beta + \varepsilon$ and use function `create.fixed` in the R package `knockoff` to create fixed-X knockoff variables. The average FDP and power are shown at Figure 1.

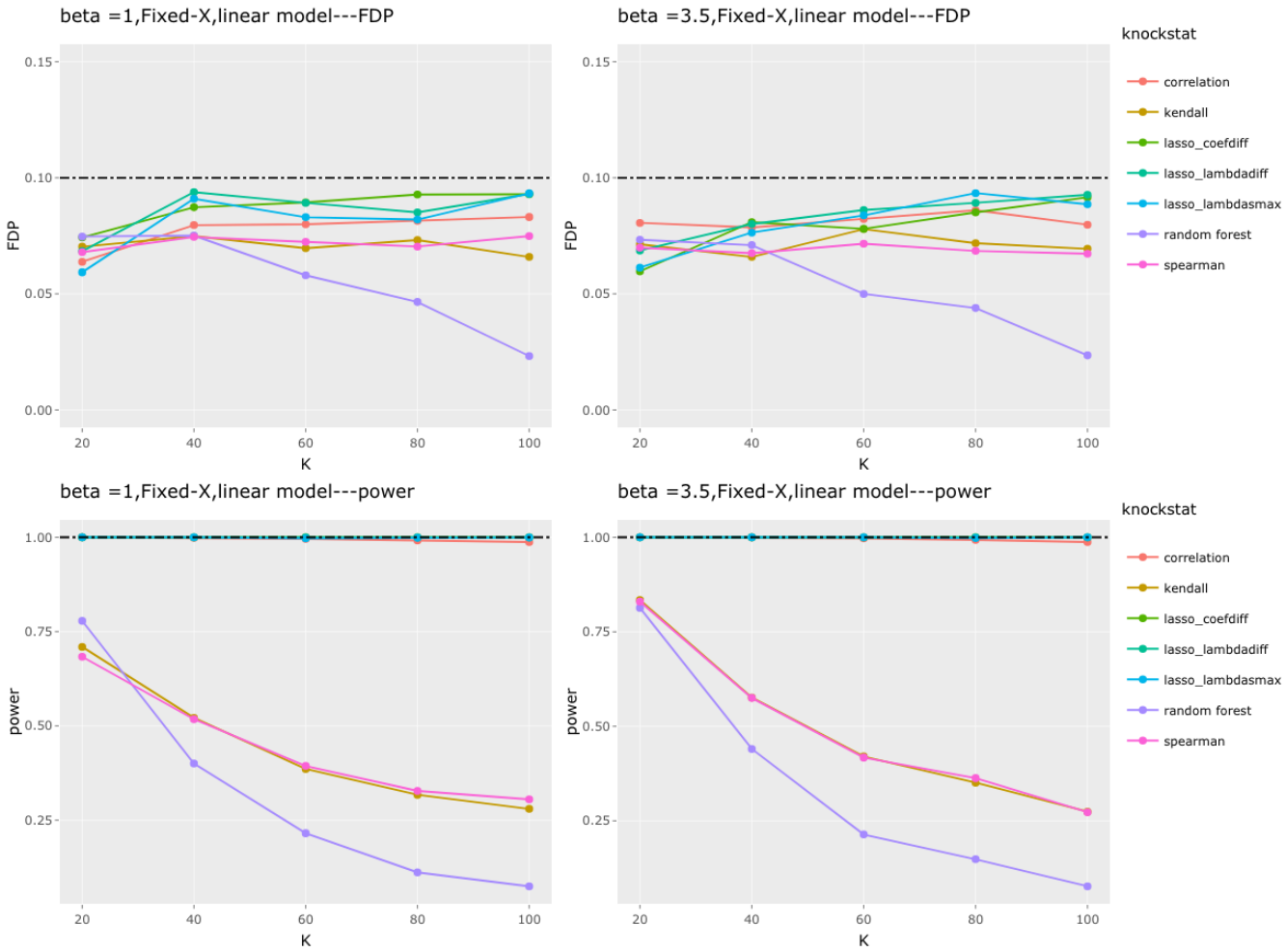

Figure 1: Fixed-X and linear assumption

From the Figure 1, FDRs have been controlled for all importance statistics. However, the power result shows that only correlation and statistics based on Lasso have average power close to 1. The possible reason is that these statistics are theoretical meaningful under the linear model assumption, they utilize the linear model assumption to capture the true importance, while the remaining statistics are based on non-parametric methods and they simply throw away the information behind the correct assumption of linear model. That's why these non-parametric statistics will lose power.

Besides, from the Figure 1, the power decreases as the number of true signal k increases. The main reason is that the threshold t decreases slowly when k increases, therefore the increment of true positives is much less than the increment of k. However, the power of random forest method drops more quickly than others, which is because the threshold t for random forest method will increases as k increases, so less true signals will be rejected.

Finally, the power is larger if the magnitude of true signal is larger. It's obvious that if the true signals are more significant, they will be found out and rejected more easily.

3.2 Model-X setting and linear model

We still assume the model to be $Y = X\beta + \varepsilon$. Since we know the true distribution of X, we use function `create.gaussian` to create model-X knockoff variables here. The average FDP and power are shown at Figure 2.

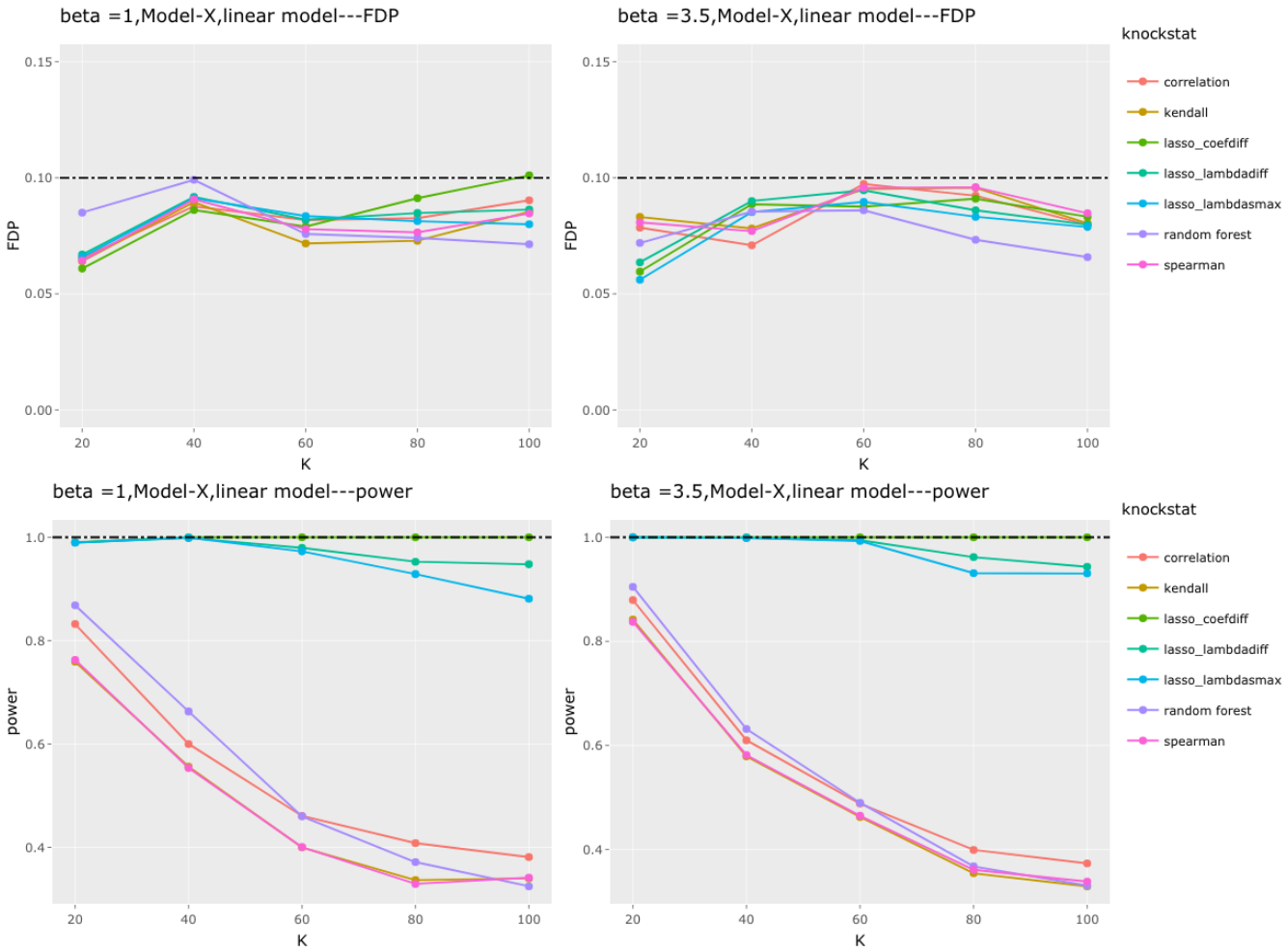

Figure 2: Model-X and linear assumption

From the Figure 2, FDRs also have been controlled for all importance statistics, and the properties of power are similar to the above section except that:

1. The power of inner product (correlation) is no longer as good as statistics based on Lasso, it now has power similar to other statistics based on non-parametric methods. The possible reason is that Fixed-X setting standardizes original X_i by multiplying $\frac{1}{\|X_i\|_2}$ to each X_i for $i \in \{1, \ldots, p\}$, which means Fixed-X setting generates standardized data X_i' and knockoffs $\tilde{X}_i$ which satisfy $\|X_i'\|_2 = \|\tilde{X}_i\|_2 = 1$, therefore the correlation statistics can reveal the true importance of X accounting for the issues of scale. However, Model-X doesn't address the issue. It generates knockoff copies from conditional distribution randomly, so the actual norms of X_i and $\tilde{X}_i$ are different. That's a possible reason for why inner product has smaller power in Model-X setting. More discussion can be found be Section 5.

2. The power of statistics based on random forest now performs similar to or even better than other statistics based on non-parametric methods.

3.3 Model-X and non-linear model

In this section, we investigate the performances of above importance statistics under non-linear model assumption.We assume the model to be $y = \sin(X) \cdot \beta + \varepsilon$. Similarly, we use function `create.gaussian` to create model-X knockoff variables here. The average FDP and power are shown at Figure 3.

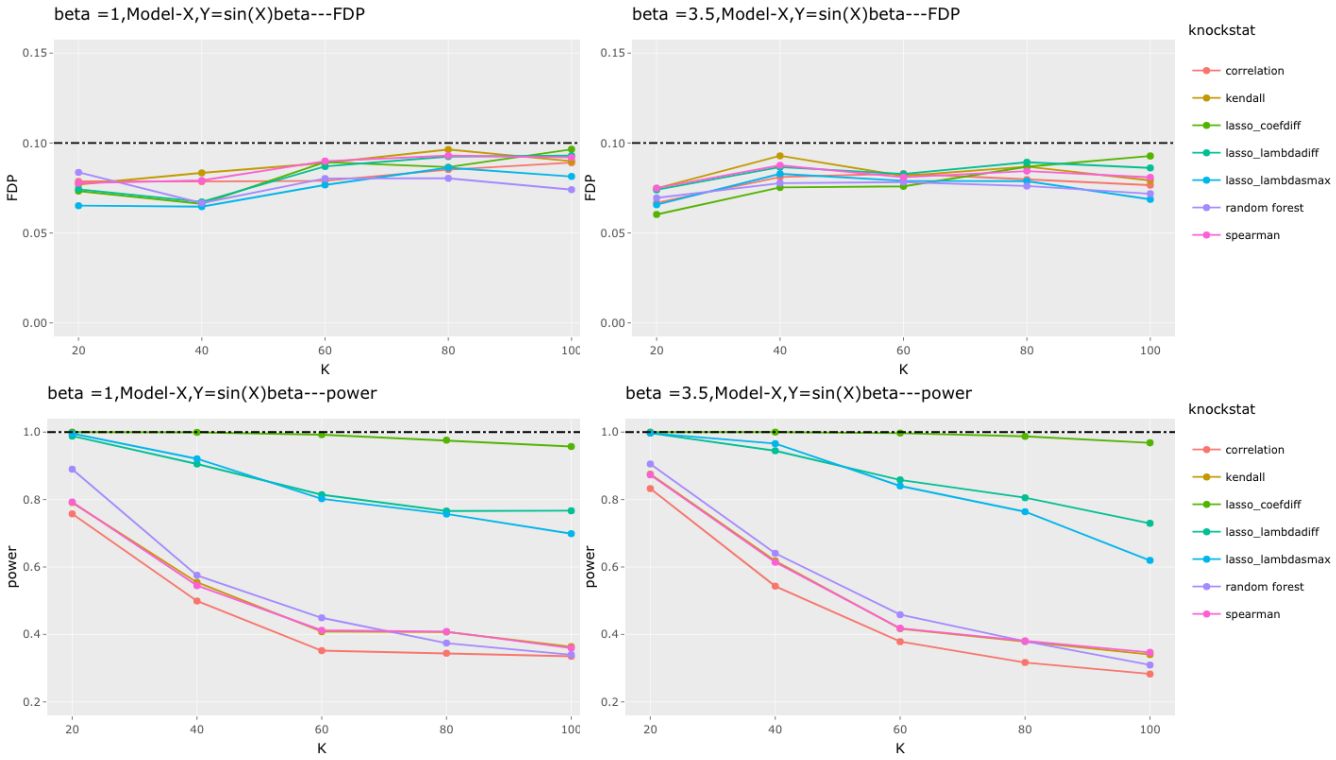

Figure 3: Model-X and non-linear assumption

The result of Figure 3 is surprising. Because we expect that under non-linear model assumption, the importance statistics based on non-parametric methods would have larger power comparing to importance statistics. But the Figure 3 shows that the Lasso based statistics have powers larger than non-parametric statistics, and lasso-coefdiff still has power close to 1. One possible reason is that we sample X from multivariate normal distribution with $\mu_p = 0$, therefore the entries of X are actually close to 0 and the non-linear model we used is similar to the linear model. However, powers of lasso based statistics do decrease, so we conclude that non-linear model assumption will affect the performance of lasso based statistics.

3.4 Model-X and GLM model

In this section, we assume a generalized linear model: $\text{logit}\left(\mathbb{P}(y=1)\right) = X\beta$, since importance statistics based on correlation are uninterpretable, we abandon these statistics here and use importance statistics based on regularized logistic regression. We also implement original Lasso based statistics and random forest for comparison.

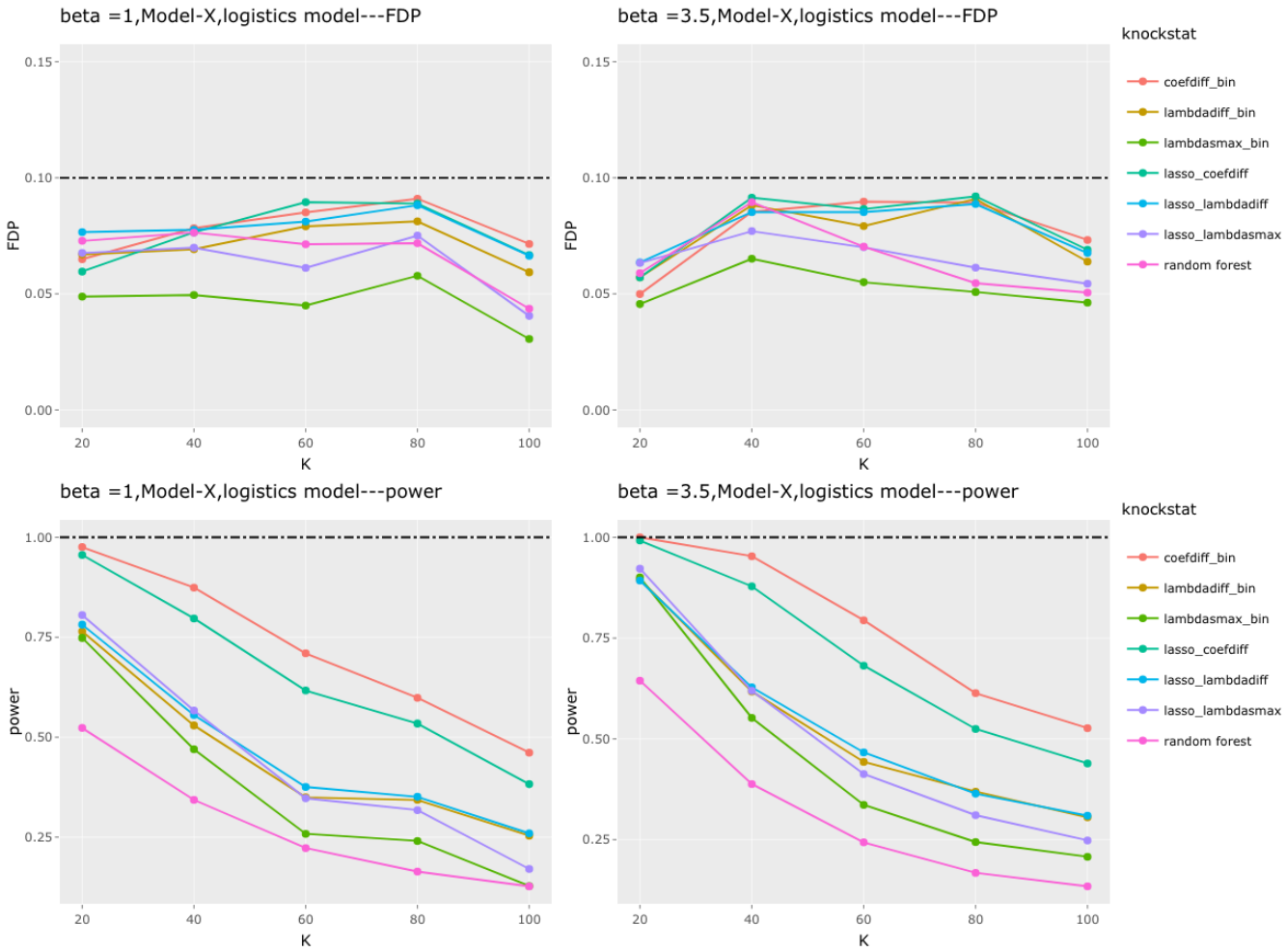

Figure 4: Model-X and GLM assumption

From the Figure 4, we get different results for three different pairs of Lasso based importance statistics. For coefdiff, the power will be larger if the importance statistics are based on regularized logistic regression; for lambdadiff, both regularized logistic regression and regularized linear regression lead to similar power; for lambdasmax, the power will be larger if the importance statistics are based on regularized linear regression. Besides, coefdiff leads to largest power and lambdasmax leads to smallest power. However, all Lasso based statistics lead to power larger than random forest method.

The possible reasons for opposite results of coefdiff and lambdasmax are because logistic regression reveal the true importances of variables, so in principle statistics based on regularized logistic regression should have better performance, which is larger power. But the lambdasmax based on regularized logistic regression may lead to an over conservative result, which could be observed from the average FDP. That's why it has lower power comparing to lambdasmax based on regularized linear regression. More discussion can be found be Section 5.

4 Conclusion

In conclusion, all importance statistics proposed above are able to control the target FDR, but different statistics have different performances on power. In general, Lasso based statistics have larger power comparing to correlation based statistics, especially under linear model assumption. Among all Lasso based statistics, coefdiff has the largest power in all settings we designed. Since random forest is a general method fits for all conditions, we don't expect it to have large power.

5 Further discussion

ISSUE 1 From our experiments, we find using inner product as importance statistic has much smaller power in model-X setting than fixed-X setting. In both cases, the relationship between X and Y is linear.

Examination We check the code of generating knockoffs in both fixed-X and model-X setting. We find in fixed-X setting, the generated knockoff will be scaled to unit column norm together with original X. While in model-X setting, both knockoff and original X stay the raw scale. Can this make a difference?

So we restore the knockoff and X to the raw scale in fixed-X setting, then conduct the same experiment, the result is pretty much the same. We further plot the histogram of the inner product statistic, the shape is quite similar with the one in model-X setting, only difference is the scale. Our implementation is the same as the function `knockoff.filter` in package `knockoff`, which also uses scaled X and $\tilde{X}$ to generate knockoff copies. Therefore, this issue might be left for future research.

ISSUE 2 In the experiments in model-X with logistic model, we observe competitive performances of linear Lasso based statistic and logistic Lasso based one, which is weird since the logistic based one without model misspecification should work better generally.

Examination This issue might be caused by too simple symmetrical distribution used in simulation, like the multivariate Gaussian we used here. Due to the symmetry, different loss functions will lead to the same variable selection results.

Therefore we try a complex distribution, which is a Gaussian mixture, then apply both linear Lasso and logistic Lasso on it. The logistic Lasso does show a larger power. However, the result is highly unstable because we use `create.second_order` to create knockoffs copy, which implies we assume the true distribution of X is multivariate normal. Therefore, this result is invalid since we don't know how to derive the conditional distribution for $\tilde{X}|X$. This might also be left for future research.