STAT 260 Project Report

Mengqi Lin

May 2021

Abstract

This project report is about the paper A Power and Prediction Analysis for Knockoffs with Lasso Statistics (Asaf Weinsteina, Rina Barberb, and Emmanuel Cande'sa).

In section 1, I introduce the background of LASSO statistics in multiple hypothesis testing, and indicate the uniqueness of the knockoff procedure, which is very different from the general multiple hypothesis testing. And this is important to judge how much possibility we can work out of the LASSO problem. In section 2, I summarized the main contents of the paper and gave some understandings, you may skip this part if you're already familiar with the paper and watched my presentation. In section 3, I gave my observations and comments to this paper, which is probably the most important things in my report. And in section 4, I gave some extensions that I think of to work on, based on section 3.

1 Background of the LASSO statistics in multiple hypothesis testing

The LASSO statistics is known to be universally sophisticated to analyze both in practice and theory. In practice, a natural question occurs is: what λ should we choose? One scientific criteria is the FDR-TPP criteria, that is, we wish to choose the λ such that the FDR(False Discovery Rate) is below some level, say 0.05 while maximize the TPP(True Positive Rate) as much as possible. And since losing the FDR is a less desiring thing compared to losing the TPP, we called a procedure (which returns a λ) valid if the procedure controls FDR always in finite samples. Therefore, the seek for such valid procedure has become a hot topic in multiple hypothesis testing research area. But finding such a procedure is not an easy thing, since the LASSO statistics is very different from the usual statistics in multiple hypothesis testing: The general multiple hypothesis testing works on "individual" statistics: Usually, they compute p-values from each individual hypothesis testing, defined as "When the hypothesis is null, the probability of being more extreme", for example,

$$p_i = 1 - \Phi(Z_i)$$

But what does it mean for "When the hypothesis is null, the probability of the LASSO statistics being more extreme"? On one hand, we don't know the null distribution for LASSO statistics, i.e we don't know how $\widehat{\beta}(\lambda)$ is distributed when $\beta = 0$. On the other hand, each hypothesis $\mathcal{H}_i$ in LASSO statistics involve not only statistics i, but the whole data set. And so the general multiple hypothesis testing procedure that works directly on p-values can do little on LASSO statistics.

Here, I would like to mention the importance of the null distribution, this is out of the fact that the numerator of FDR requires the knowledge of that.

Until (Barber and Candes, 2015), the knockoff procedure makes a breakthrough. Their procedure controls FDR with mild assumptions and conditions. And the idea of the knockoff procedure is to create some fake LASSO statistics that comes from the null distribution. By doing so, we are having hope of capturing

the null distribution and further controlling FDR.

Despite the utility of the knockoff procedure, the not-much-discussed power of the knockoff procedure has been one of the shortcomings of it.

2 Main Contents of the paper

2.1 A fully charaterization of the LASSO problem

Almost the same time as the knockoff procedure, (Montanari and Bayati, 2011) proposed a fully charaterization of the LASSO problem under a specific setting. The theorem they proved turned into a powerful tool to make it possible to analyze the power of the knockoff procedure, which is familiar to us already:

consider the linearmodel

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{z} \tag{1}$$

in which $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ has i.i.d. $\mathcal{N}(0, 1/n)$ entries and the errors z_i are i.i.d. $\mathcal{N}(0, \sigma^2)$ with $\sigma \geq 0$ fixed and arbitrary. We assume that the regression coefficients $\beta_j,\ j = 1, ..., p$ are i.i.d. copies of a random variable Π with $\mathbb{E}(\Pi^2) < \infty$ and $\mathbb{P}(\Pi \neq 0) = \epsilon \in (0, 1)$ for a fixed constant ϵ. In this section we will be interested in the case where $n, p \to \infty$ with $n/p \to \delta$ for a positive constant δ. Note that Π is assumed to not depend on p, and so the expected number of nonzero elements β_j is equal to $\epsilon \cdot p$.
Let $\widehat{\boldsymbol{\beta}}(\lambda)$ be the Lasso solution,

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\mathbf{b}\in\mathbb{R}^p}{\arg\min} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\mathbf{b}\|^2 + \lambda\|\mathbf{b}\|_1, \tag{2}$$

and denote $V(\lambda) = |\{j : \widehat{\beta}_j(\lambda) \neq 0, \beta_j = 0\}|$, $T(\lambda) = |\{j : \widehat{\beta}_j(\lambda) \neq 0, \beta_j \neq 0\}|$, $R(\lambda) = |\{j : \widehat{\beta}_j(\lambda) \neq 0\}|$ and $k = |\{j : \beta_j \neq 0\}|$. Hence $V(\lambda)$ is regarded as the number of false 'discoveries' made by the Lasso; $T(\lambda)$ is the number of true discoveries; $R(\lambda)$ is the total number of discoveries, and k is the number of true signals. We would like to remark here that $\beta_j = 0$ implies that $\boldsymbol{X}_j$ (the j-th variable) is independent of $\boldsymbol{y}$ marginally *and* conditionally on any subset of $X_{-j} := \{X_1, ..., X_{j-1}, X_{j+1}, ..., X_p\}$; hence the interpretation of rejecting the hypothesis $\beta_j = 0$ as a false discovery is clear and unambiguous. The false discovery proportion (FDP) is defined as usual as

$$\mathrm{FDP}(\lambda) = \frac{V(\lambda)}{1 \vee R(\lambda)}$$

and the true positive proportion (TPP) is defined as

$$\mathrm{TPP}(\lambda) = \frac{T(\lambda)}{1 \vee k}.$$

Also, let Π^* denote a random variable distributed according to the conditional distribution of Π given $\Pi \neq 0$; that is,

$$\Pi = \begin{cases} \Pi^*, & \text{w.p. } \epsilon, \\ 0, & \text{w.p. } 1 - \epsilon. \end{cases} \tag{3}$$

Finally, denote by α_0 the unique root of the equation $(1 + t^2)\Phi(-t) - t\phi(t) = \delta/2$.

Lemma 1. *The Lasso solution with a fixed $\lambda > 0$ obeys*

$$\frac{V(\lambda)}{p} \xrightarrow{\mathbb{P}} 2(1 - \epsilon)\Phi(-\alpha),$$
$$\frac{T(\lambda)}{p} \xrightarrow{\mathbb{P}} \mathbb{P}(|\Pi + \tau W| > \alpha\tau, \Pi \neq 0) = \epsilon\mathbb{P}(|\Pi^* + \tau W| > \alpha\tau),$$

where $W \sim \mathcal{N}(0,1)$ *independently of* Π*, and* $\tau > 0$*,* $\alpha > \max\{\alpha_0, 0\}$ *is the unique solution to*

$$\begin{aligned} \tau^2 &= \sigma^2 + \frac{1}{\delta}\mathbb{E}(\eta_{\alpha\tau}(\Pi + \tau W) - \Pi)^2 \\ \lambda &= \left(1 - \frac{1}{\delta}\mathbb{P}(|\Pi + \tau W| > \alpha\tau)\right)\alpha\tau. \end{aligned} \tag{4}$$

Lemma 1 is a consequence of the fact that, under the working assumptions, $(\boldsymbol{\beta}, \widehat{\boldsymbol{\beta}}(\lambda))$ is in some (limited) sense asymptotically distributed as $(\boldsymbol{\beta}, \eta_{\alpha\tau}(\boldsymbol{\beta} + \tau\mathbf{W}))$, where $\mathbf{W} \sim \mathcal{N}_{\mathbf{p}}(\mathbf{0}, \mathbf{I})$ independently of $\boldsymbol{\beta}$; here, the soft-thresholding operation acts on each component of a vector.

It follows immediately from Lemma 1 that for a fixed $\lambda > 0$, the limits of FDP and TPP are

$$\mathrm{FDP}(\lambda) \xrightarrow{\mathbb{P}} \frac{2(1-\epsilon)\Phi(-\alpha)}{2(1-\epsilon)\Phi(-\alpha) + \epsilon\mathbb{P}(|\Pi^* + \tau W| > \alpha\tau)} \tag{5}$$

and

$$\mathrm{TPP}(\lambda) \xrightarrow{\mathbb{P}} \mathbb{P}(|\Pi^* + \tau W| > \alpha\tau). \tag{6}$$

2.2 The knockoff procedure

Let $\widetilde{\mathbf{X}} \in \mathbb{R}^{n\times r}$ be a matrix with i.i.d. entries drawn from G independently of $\boldsymbol{X}$ and of all other variables, where r is a fixed positive integer. Next, let

$$\mathbb{X} := [\boldsymbol{X}\ \widetilde{\mathbf{X}}] \in \mathbb{R}^{n\times(p+r)}$$

denote the augmented design matrix, and consider the Lasso solution for the augmented design,

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\mathbf{b}\in\mathbb{R}^{p+r}}{\arg\min}\ \frac{1}{2}\|\boldsymbol{y} - \mathbb{X}\mathbf{b}\|^2 + \lambda\|\mathbf{b}\|_1. \tag{7}$$

For simplicity we keep the notation $\widehat{\boldsymbol{\beta}}(\lambda)$ for the solution corresponding to the augmented design, although it is of course different from (2) (for one thing, (7) has $p+r$ components versus p for (2)). Let

$$\mathcal{H} = \{1, ..., p\}, \quad \mathcal{H}_0 = \{j \in \mathcal{H} : \beta_j = 0\}, \quad \mathcal{K}_0 = \{p+1, ..., p+r\}$$

be the sets of indices corresponding to the original variables, the null variables and the knockoff variables, respectively. Note that, because we are conditioning on $\boldsymbol{\beta}$, the set $\mathcal{H}_0$ is nonrandom. Now define the statistics

$$T_j = \sup\{\lambda : \widehat{\beta}_j(\lambda) \neq 0\}, \quad j = 1, ..., p+r. \tag{8}$$

Informally, T_j measures how early the jth variable enters the Lasso path (the larger, the earlier).

Let

$$V_0(\lambda) = |\{j \in \mathcal{H}_0 : T_j \geq \lambda\}|, \quad V_1(\lambda) = |\{j \in \mathcal{K}_0 : T_j \geq \lambda\}|$$

be the number of true null and fake null variables, respectively, which enter the Lasso path "before" time λ. Also, let

$$R(\lambda) = |\{j \in \mathcal{H} : T_j \geq \lambda\}|$$

be the total number of original variables entering the Lasso path before "time" λ. The class of procedures we propose (henceforth, knockoff procedures) reject the null hypotheses corresponding to

$$\{j \in \mathcal{H} : T_j \geq \widehat{\lambda}\},$$

where

$$\widehat{\lambda} = \inf\left\{\lambda \in \Lambda : \frac{(1 + V_1(\lambda)) \cdot \frac{|\mathcal{H}|\widehat{\pi}_0}{1+|\mathcal{K}_0|}}{R(\lambda)} \leq q\right\} \tag{9}$$

for an estimate $\widehat{\pi}_0$ of $\pi_0 := |\mathcal{H}_0|/|\mathcal{H}|$ and for a set Λ associated with $\widehat{\pi}_0$.

Theorem 2 (FDR control for knockoff procedure with $\widehat{\pi}_0 = 1$). *Let $T_j,\ j = 1, ..., p+r$ be test statistics with a joint distribution that is invariant to permutations in the set $\mathcal{H}_0 \cup \mathcal{K}_0$ (that is, reordering the statistics with indices in the extended null set, leaves the joint distribution unchanged). Set $V_0(t) = |\{j \in \mathcal{H}_0 : T_j \geq t\}|$, $V_1(t) = |\{j \in \mathcal{K}_0 : T_j \geq t\}|$, and $R(t) = |\{j \in \mathcal{H} : T_j \geq t\}|$. Then the procedure that rejects $\mathcal{H}_0^j$ if $T_j \geq \tau$, where*

$$\tau = \inf \left\{ t \in \mathbb{R} : \frac{(1 + V_1(t)) \cdot \frac{|\mathcal{H}|}{1+|\mathcal{K}_0|}}{1 \vee R(t)} \leq q \right\} \tag{10}$$

controls the FDR *at level q. In fact,*

$$\mathbb{E}\left[\frac{V_0(\tau)}{1 \vee R(\tau)}\right] \leq q\frac{|\mathcal{H}_0|}{|\mathcal{H}|}.$$

2.3 Power Analysis of the knockoff procedure

Taking $r = \rho p$ for a constant $\rho > 0$, we have that $n/(p + r) \to \delta' := \delta/(1 + \rho)$ as $n, p \to \infty$. Furthermore, because $\mathbb{E}Y$ is related to $\mathbb{X}$ through

$$\mathbb{E}Y = \mathbb{X}\begin{bmatrix}\boldsymbol{\beta} \\ \mathbf{0}\end{bmatrix},$$

where $\mathbf{0} = (0, ..., 0)^T \in \mathbb{R}^r$, we have that the empirical distribution of the coefficients corresponding to the *augmented* design converges exactly to

$$\Pi' = \begin{cases} \Pi^*, & \text{w.p. } \epsilon', \\ 0, & \text{w.p. } 1 - \epsilon', \end{cases} \tag{11}$$

where $\epsilon' := \epsilon/(1 + \rho) = \lim\{p\epsilon/(p + \rho p)\}$.

Recalling the definitions of $\mathcal{H}, \mathcal{H}_0$ and $\mathcal{K}_0$ from Section **??**, a counterpart of Lemma 1 asserts that

$$\frac{|\{j \in \mathcal{H} : \widehat{\beta}_j(\lambda) \neq 0, \beta_j = 0\}|}{p} \longrightarrow 2(1 - \epsilon)\Phi(-\alpha'), \tag{12}$$

$$\frac{|\{j \in \mathcal{H} : \widehat{\beta}_j(\lambda) \neq 0, \beta_j \neq 0\}|}{p} \longrightarrow \epsilon\mathbb{P}(|\Pi^* + \tau' W| > \alpha'\tau') \tag{13}$$

and

$$\frac{|\{j \in \mathcal{K}_0 : \widehat{\beta}_j(\lambda) \neq 0\}|}{r} \longrightarrow 2\Phi(-\alpha'), \tag{14}$$

where $\widehat{\beta}_j(\lambda)$ is the solution to (7); W is a $\mathcal{N}(0, 1)$ variable independent of Π; and $\tau' > 0,\ \alpha' > \max\{\alpha'_0, 0\}$ is the unique solution to (4) when δ is replaced by δ' and ϵ is replaced by ϵ'. Hence,

$$\begin{aligned} \mathrm{FDP}(\lambda) &= \frac{|\{j \in \mathcal{H}_0 : T_j \geq \lambda\}|}{1 \vee |\{j \in \mathcal{H} : T_j \geq \lambda\}|} \\ &= \frac{|\{j \in \mathcal{H}_0 : \widehat{\beta}_j(\lambda') \neq 0 \text{ for some } \lambda' \geq \lambda\}|}{1 \vee |\{j \in \mathcal{H} : \widehat{\beta}_j(\lambda') \neq 0 \text{ for some } \lambda' \geq \lambda\}|} \\ &\approx \frac{|\{j \in \mathcal{H}_0 : \widehat{\beta}_j(\lambda) \neq 0\}|}{1 \vee |\{j \in \mathcal{H} : \widehat{\beta}_j(\lambda) \neq 0\}|} \\ &\longrightarrow \frac{2(1 - \epsilon)\Phi(-\alpha')}{2(1 - \epsilon)\Phi(-\alpha') + \epsilon\mathbb{P}(|\Pi^* + \tau' W| > \alpha'\tau')} \equiv \mathrm{FDP}^{\infty}_{\mathrm{aug}}(t). \end{aligned} \tag{15}$$

With the equation we are able to know the optimal λ for the knockoff setting and we can compare that with the λ given by the knockoff procedure and also make comparisons of the power TPP. The simulations results show that the two procedures give almost the same power. This in some sense proves the optimality of the knockoff procedure.

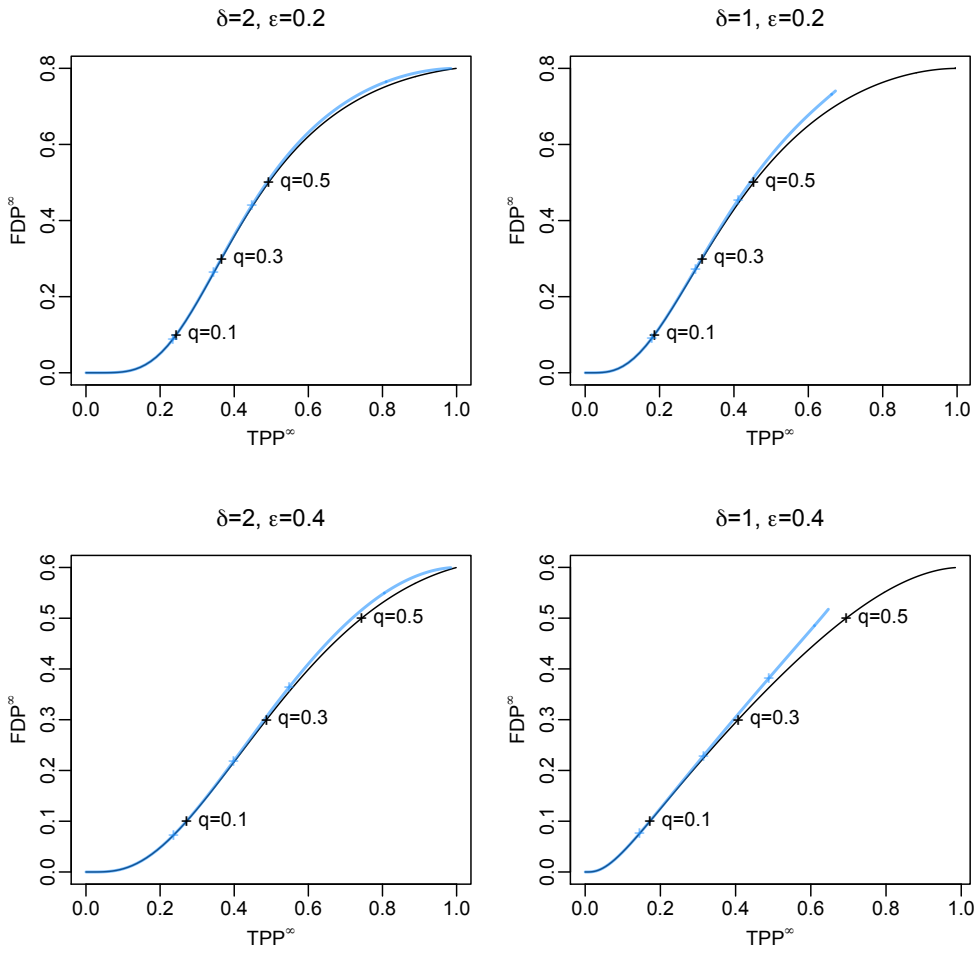

Figure 1: Tradeoff diagrams for $\Pi^* \sim \exp(1)$. $\sigma = 0.5$; $\rho = 1$. Different panels correspond to different combinations of δ and ϵ, as appears in the title of each of the plots. The black (resp. light blue) curve corresponds to the oracle (resp. knockoffs). Markers denote the pair (tpp, fdp) of asymptotic power and type I error, attained for three example values of q: black for oracle, light blue for knockoff. The knockoff procedure loses a little bit of power due to the estimate of $1 - \epsilon$ (proportion of true nulls).

3 Remarks

In this section, I will give the remarks I summarized to the whole paper.

3.1 Success of this paper

This paper bridges the two excellent results from knockoff procedure and the oracle LASSO diagram, and make up for each other's shortcoming. The knockoff procedure may not be convinced by people without power guarantee, and the oracle result may not be practically useful enough with the strong assumptions. And this paper shows us that under the strong assumptions, where we are able to compare the power of the knockoff procedure with the oracle optimal one asymptotically, the power of the two procedures are almost the same!

3.2 Shortcomings of this paper

Still, despite that we proved the power of the knockoff procedure under the specific settings, we don't know how much we can rely on the results given by the strong assumptions: Do we still have the optimal power when the assumptions break? This is especially important, since the knockoff procedures being used are mostly around finite sample and usually doesn't have the Gaussian iid, and the proportion assumption. Can the strong assumptions be seen commonly enough in practice? It will be interesting if we are able to find enough examples in practice that satisfy such assumption setting, or to work out other powerful oracle results under other assumptions. But still, this paper in some sense gives us guidance that the knockoff procedure works well.

3.3 Importance of the iid assumption

Recall that the knockoff procedure for our iid setting works on statistics

$$T_j = \sup\{\lambda : \widehat{\beta}_j(\lambda) \neq 0\}, \quad j = 1, ..., p + r.$$

And decides to reject hypothesis j with $T_j \geq \hat{\lambda}$. This T_j statistics actually differs from the statistics they used in the firstly proposed knockoff procedure, and the reason why this statistics can work is because that here we are having the iid X_{ij} design, so that for all null hypothesis $j \in \mathcal{H}_0 : T_j$ distributed the same. But if we don't have the guarantee of the iid X_{ij} then for $j \in \mathcal{H}_0$, T_j may not have the same distribution.

3.4 The estimation of $\hat{\pi_0}$

In the paper we see that we can use $\hat{\pi_0} = 1$ as an overestimator, which can also control FDR, with a bit loss of the TPP. The more powerful way is to estimate $\hat{\pi_0}$ via "censor" technique in multiple hypothesis testing, which was initially proposed by (Barber, and Li, 2017 SABHA), the idea is to choose some intermediate t_0, and the knockoff procedure decides the threshold τ via those hypothesis with $T_j > t_0$, and estimate the null proportion via those hypothesis with $T_j < t_0$. The reason why we need this intermediate t_0 is to avoid overfitting and the violation of the FDR control.

4 Extension

Direction 1: other knockoff procedures As mentioned, actually, the original knockoff statistics in (Barber and Candes, 2015) was the Lasso coefficient-difference (LCD) statistic.

$$W_j = |\widehat{\beta}_j(\lambda)| - |\widehat{\beta}_{p+j}(\lambda)|$$

But the knockoff statistics in this paper is just the

$$T_j = \sup\{\lambda : \widehat{\beta}_j(\lambda) \neq 0\}$$

I explained that it's not necessary to use the Lasso coefficient-difference (LCD) statistic because of the benefit of our iid design setting. But using the LCD, we can relax the iid assumptions. And in fact, somebody did that work already. "A Power Analysis for Knockoffs with the Lasso Coefficient-Difference Statistic Asaf Weinstein, Weijie J. Su, Ma, lgorzata Bogdan, Rina F. Barber, Emmanuel J. Cand'es". (2020)

I guess there still exists some variation of the knockoff procedures. Perhaps changing the model a bit, or changing the statistics like W_j.

Direction 2: a more flexible sequential testing procedure So I think one of the hot topics in multiple hypothesis testing is to utilize the "weights", to emphasize the different significance across hypothesis, it's pretty much like the adaptive LASSO. But I think it will be fun to have some "group structure" into the procedure. For example, let's divide all the hypothesis into groups, (usually scientists believe that some hypothesis are grouped together) and then compute the LASSO statistics or run the procedure separately in each group, etc.
And in fact the discussion 3.4 of the estimation $\hat{\pi}_0$ is one special case of such extension.

5 References

- Multiple testing with the structure adaptive Benjamini-Hochberg algorithm (Ang Li and Rina Foygel Barber 09.13.17)
- Controlling the false discovery rate via knockoffs(Barber and Candes, 2015)
- False discoveries occur early on the lasso path(Weijie Su and Ma, lgorzata Bogdan† and Emmanuel Cande's).
- A Power Analysis for Knockoffs with the Lasso Coefficient-Difference Statistic. (Asaf Weinstein , Weijie J. Su , Malgorzata Bogdan, Rina F. Barber, Emmanuel J. Cand'es)
- The LASSO risk for gaussian matrices (Mohsen Bayati, and Andrea Montanari)
- A Power and Prediction Analysis for Knockoffs with Lasso Statistics (Asaf Weinsteina, Rina Barberb, and Emmanuel Cande's)