STATS 300C: Theory of Statistics **Spring 2022**

Lecture 10 — Apr 28, 2022

Lecturer: Prof. Emmanuel Candès *Editor: Parth Nobel, Scribe: Ran Xie*

Warning: *These notes may contain factual and/or typographic errors. They are based on Emmanuel Candès's course from 2018 and 2022, and scribe notes written by Emmanuel Candès, Michael Celentano, Zijun Gao, Shuangning Li, and Chenyang Zhong.*

Agenda:

1. Model-X Knockoffs
2. FDR control
3. Statistical modeling: what do we know?
4. Applications in genetics

10.1 Setup

Recall the *control variable selection problem* introduced in the last lecture: we observe n i.i.d. pairs $\{(X^{(i)}, Y^{(i)})\}_{i=1}^n$ each with joint distribution $P_{Y,X}$, where X is a p-dimensional vector of potential control variables and Y is a response variable, and we want to discover which control variables X_j are predictive of Y given all the other control variables $X_{-j} \coloneqq \{X_i\}_{i=1}^p \setminus \{X_j\}$. Specifically, we want to test the set of null hypotheses $H_j \colon Y \perp\!\!\!\perp X_j \mid X_{-j}$ for $j \in [p] \coloneqq \{1, \dots, p\}$; we refer to X_j such that H_j is true as a "null" control variable and a variable X_j such that H_j is not true as a "non-null" control variable. Further, we let $\mathcal{H}_0 \coloneqq \{j \in [p] \colon Y \perp\!\!\!\perp X_j \mid X_{-j}\}$ denote the set of indices corresponding to null control variables, we let $\mathcal{H}_0^c \coloneqq [p] \setminus \mathcal{H}_0$ denote the set of indices corresponding to non-null control variables, and we let $X_{\mathcal{H}_0} \coloneqq \{X_j \colon j \in \mathcal{H}_0\}$ and $X_{\mathcal{H}_0^c} \coloneqq \{X_j \colon j \notin \mathcal{H}_0\}$ denote the sets of null and non-null control variables, respectively. We note that under mild regularity conditions, $X_{\mathcal{H}_0^c}$ forms a Markov blanket for Y, i.e. $X_{\mathcal{H}_0^c}$ is the smallest set of control variables such that $Y \perp\!\!\!\perp X_{\mathcal{H}_0} \mid X_{\mathcal{H}_0^c}$; see [3] for details. An illustration of a Markov blanket for Y is given in Figure 10.1.

Recall from last lecture that many statistical procedures and machine learning algorithms that model the conditional distribution $P_{Y|X}$ (or certain features of it like $\mathbb{E}[Y|X]$) report "feature importance scores" $T_j(Y, X)$ that measure how predictive X_j is of Y given the other control variables X_{-j}. Some examples include the magnitude of the (potentially regularized) regression coefficient $|\hat{\beta}_j|$ corresponding to variable X_j in a (generalized) linear model, or the percentage of trees in a random forest with splits on X_j. We provide an example set of such feature importance scores in Figure 10.2a.

Unfortunately, it is not clear whether if any importance score is high, it is actually predictive of Y. Consider the following simple example:

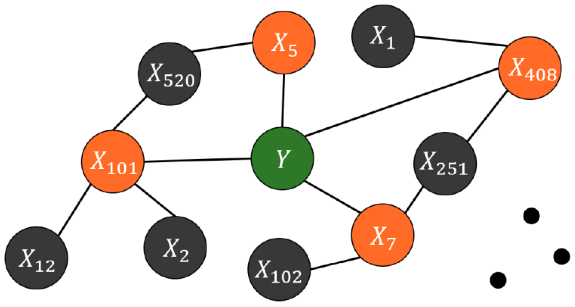
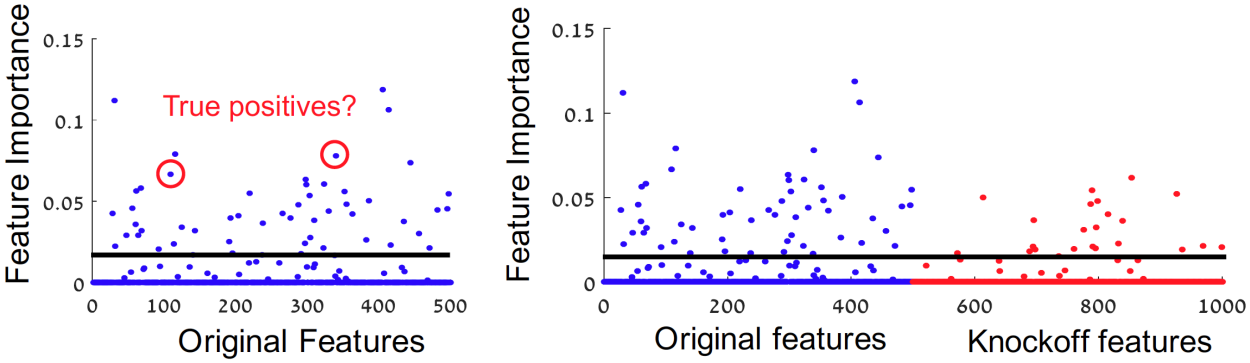


Figure 10.1: An example of a dependency graph for (Y, X), where edges denote dependence between variables; the orange-highlighted nodes form a Markov blanket for Y because all paths to Y must pass through one of these nodes.



(a) Importance Scores for Features

(b) Importance Scores for Features and Knockoffs

Figure 10.2: An example of importance scores for a set of 500 null features with feature indices on the x-axis, along with the importance scores for appropriately generated knockoffs copies of the original features.

Example 1. Suppose there are two control variables X_1 and X_2, $Y = X_2 + \epsilon$ for $\epsilon \sim \mathcal{N}(0, 1)$ independent of X_1 and X_2, and $\text{Cov}(X_1, X_2) = 0.5$

If $T_1(Y, (X_1, X_2))$ is large in the context of Example 1, it could just be because the importance score generation procedure picked up on the fact that X_1 is predictive of Y because it is correlated with X_2, even though $Y \perp\!\!\!\perp X_1 \mid X_2$. As we discussed in the last lecture then, if we want to use $T_j(Y, X)$ as a black-box measure of variable importance that allows us to select control variables in a statistically valid fashion, it is important to benchmark it against some comparable importance score generated under the null H_j to assess whether $T_j(Y, X)$ is actually informative about the veracity of H_j. This perspective motivates both the Conditional Randomization Testing and Knockoffs frameworks, with appropriately-generated knockoffs illustrated in Figure 10.2b.

When we introduced the Conditional Randomization Testing framework last lecture, we noted that although it provides valid p-values for testing H_j, there are two principal limitations for this approach. First of all, the p-values constructed in CRT are not independent, while the Knockoffs framework provides independent evidence for each test, i.e. independent coarse p-values. Secondly, CRT is computationally intense if Bonferroni threshold is low,

since it requires recomputing the feature importance score (an expensive procedure if based on a complex machine learning algorithm) for each control variable by sampling from the (assumed known) distribution $X_j \mid X_{-j}$ for each $j \in \{1, \dots p\}$. The knockoffs approach, first introduced by [1], aims to circumvent this computational complexity by generating a single set of p "knockoff" control variables $\tilde{X}$ for each observation that are independent of Y but otherwise "mimic" the control variables X if they were null and then comparing the importance scores $T_j(Y, (X, \tilde{X}))$ of X_j and $T_{p+j}(Y, (X, \tilde{X}))$ of its knockoff $\tilde{X}_j$, where the importance scores here are computed by feeding the concatenated set of $2p$ true *and* knockoff control variables $(X, \tilde{X})$ into the importance score function. If the importance scores are large for both X_j and $\tilde{X}_j$, then spurious correlation between Y and X_1 is a more plausible explanation for large $T_j(Y, (X, \tilde{X}))$ than H_j being false.

10.2 The Model-X Knockoffs Framework

10.2.1 Introducing Model-X Knockoffs

Intuitively, in the construction of a knockoff version (e.g. fake SNP) for each variable (e.g. SNP), we wish for the comparison between the two to be fair for nulls, and utilize clever filter (SeqStep) to control FDR in finite samples. More formally, we seek to construct a set of knockoff control variables with the following properties:

Definition 1 (Model-X Knockoffs)**.** In the setting described in Section 10.1, a set of variables $\tilde{X}$ are considered *model-X knockoffs* if they satisfy the following two properties:

1. *Pairwise Exchangeability*: for any index j, we have that

$$X_j, X_{-j}, \tilde{X}_j, \tilde{X}_{-j} \stackrel{d}{=} \tilde{X}_j, X_{-j}, X_j, \tilde{X}_{-j}$$

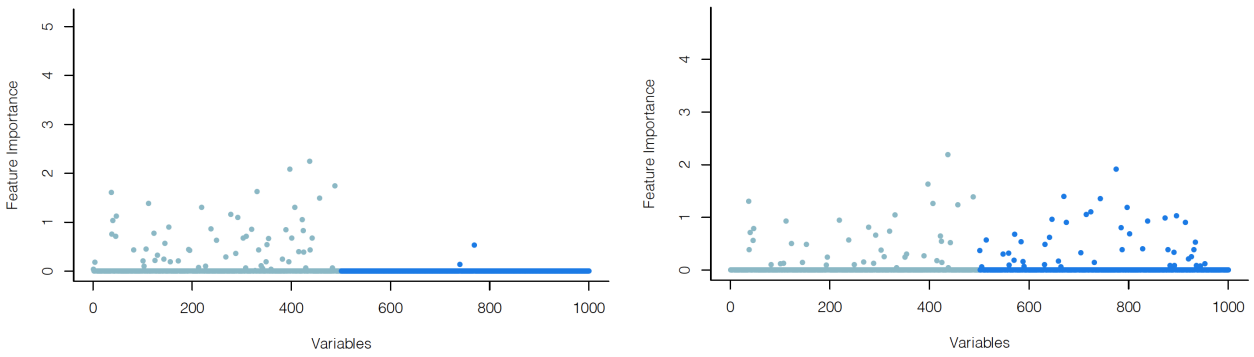
2. *Response Independence*: $\tilde{X}$ are constructed such that $Y \perp\!\!\!\perp \tilde{X}_j \mid X_{-j}$.

Importantly, it is not sufficient to just choose a permutation of the rows of X. Let $\tilde{X}$ be a permutation of the rows of matrix X, then Pairwise Exchangeability does not hold because the correlation structure between X_j (the jth column of X) and X_{-j} is not preserved when replacing X_j with $\tilde{X}_j$. If P_X is exactly known (or approximately known; see [2]), it is possible to construct knockoff variables $\tilde{X}$ that satisfy the properties in Definition 1; we will discuss methods for doing so in Lecture 11. We illustrate the contrast between permuted control variables and appropriately-generated knockoff variables in Figure 10.3.

Once we construct knockoffs $\tilde{X}$, we can construct feature importance scores $Z_j := T_j(Y, (X, \tilde{X}))$ for the real X_j and $\tilde{Z}_j := T_{p+j}(Y, (X, \tilde{X}))$ for its knockoff $\tilde{X}_j$. We can then show the following exchangeability result, which we state here without proof (as mentioned in class, completing the proof might appear in a future homework):

Lemma 1. For any null index $j \in \mathcal{H}_0$, Z_j and $\tilde{Z}_j$ are exchangeable, i.e. $(Z_j, \tilde{Z}_j) \stackrel{d}{=} (\tilde{Z}_j, Z_j)$.

We illustrate what exchangeability of importance scores looks like in Figure 10.4.



(a) Permuted Control Variables (b) Knockoffs

Figure 10.3: Plots of LASSO coefficient magnitudes estimated from data containing 500 null variables colored light blue concatenated with either 500 permuted control variables colored dark blue (Figure 10.3a) or knockoff variables (Figure 10.3b). Clearly, the importance scores for the permuted control variables are not pairwise exchangeable with the original control variables' importance scores, while knockoffs' importance scores are.

Why is it important to use both X and $\tilde{X}$ as control variables inputted into $T_j(Y, (X, \tilde{X}))$? It is instructive to consider Example 1. If we just compared $T_j(Y, (X_1, X_2))$ to $T_j(Y, (\tilde{X}_1, \tilde{X}_2))$ to determine variable importance, then since X_1 is correlated with X_2, it may be that $T_1(Y, (X_1, X_2))$ would be large since the null variable X_1 is correlated with the non-null variable X_2, while $T_1(Y, (\tilde{X}_1, \tilde{X}_2))$ would be small since $Y \perp\!\!\!\perp (\tilde{X}_1, \tilde{X}_2)$ by construction; as such, we would spuriously "reject" H_1.

10.2.2 The Knockoffs Procedure

To use the variable importance scores Z_j and their corresponding knockoff importance scores $\tilde{Z}_j$ to test H_j, we construct *knockoffs-adjusted scores* $W_j := w_j(Z_j, \tilde{Z}_j)$ via some *anti-symmetric* function $w_j := \mathbb{R}^2 \to \mathbb{R}$, i.e. a function such that $w_j(Z_j, \tilde{Z}_j) = -w_j(\tilde{Z}_j, Z_j)$, e.g. $w_j(Z_j, \tilde{Z}_j) = Z_j - \tilde{Z}_j$.

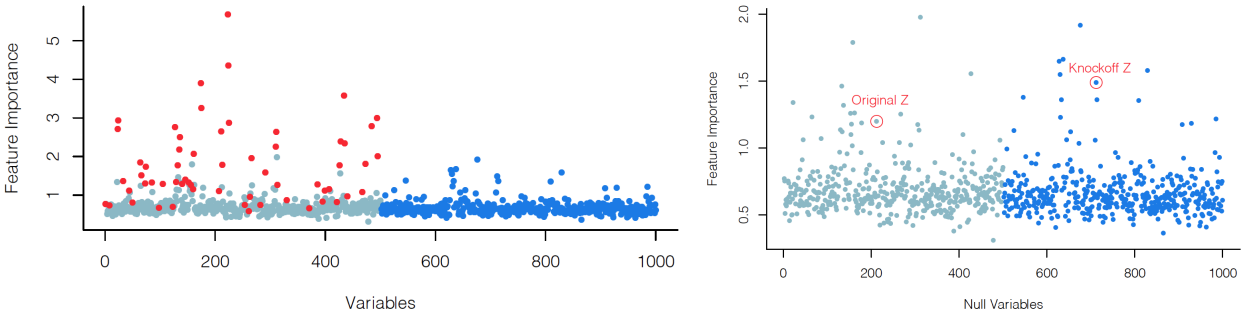
Lemma 2. For any null index $j \in \mathcal{H}_0$, the distribution of W_j is symmetric, so $\Pr(\text{sign}(W_j) = 1) = \Pr(\text{sign}(W_j) = -1) = \frac{1}{2}$, i.e. $\text{sign}(W_j)$ is a Rademacher random variable. Further, conditional on $|W| := (|W_j|)_{j=1}^p$, $\text{sign}(W_j) \overset{i.i.d.}{\sim} \text{Rad}$.

Proof. We only show the first fact. Consider any measurable set $A \subseteq \mathbb{R}$; because $(Z_j, \tilde{Z}_j) \overset{d}{=} (\tilde{Z}_j, Z_j)$ by Lemma 1,

$$\Pr(w_j(Z_j, \tilde{Z}_j) \in A) = \Pr((Z_j, \tilde{Z}_j) \in w_j^{-1}(A)) = \Pr((\tilde{Z}_j, Z_j) \in w_j^{-1}(A)) = \Pr(-w_j(Z_j, \tilde{Z}_j) \in A),$$

which is exactly the definition of symmetry of W_j. □

With this lemma in hand, we can construct a procedure for testing the hypotheses $\{H_j\}_{j=1}^p$ while controlling the FDR. To do so, we take a perspective akin to to the empirical process viewpoint on the BH procedure discussed in Lecture 7, in which we line up



(a) Importance Scores for Variables and Knockoffs (b) Null Variables and Their Knockoffs

Figure 10.4: Figure 10.4a shows random forest-based importance scores estimated from data containing 500 variables (light blue scores are correspond to null variables, red scores correspond to non-null variables) concatenated with 500 knockoff control variables with scores colored in dark blue. Figure 10.4b shows just the null variables and their corresponding knockoffs; the distributions of scores appear to be pairwise exchangeable, i.e. statistically indistinguishable.

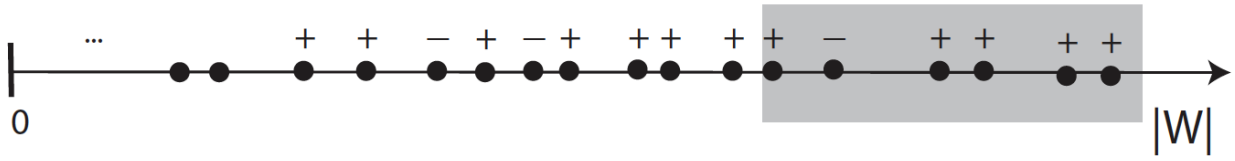


Figure 10.5: An illustration of the `SeqStep` procedure, where dots correspond to magnitudes of knockoffs-adjusted importance scores, pluses and minuses indicate the signs of each score, and the shaded region highlights the scores above a particular magnitude threshold t, as discussed below. Here, the conservative estimate of $\mathrm{FDP}(t)$, $\widehat{\mathrm{FDP}}(t) = 1/6$ since there is a single negative score in the set of six scores with magnitudes larger than t.

the W_j's in order of their magnitude $|W_j|$ and seek to reject H_j for all j such that $W_j \geq t$ for some threshold t. Since W_j is symmetrically distributed under the null, it is equally likely that $W_j \geq t$ and $W_j \leq -t$. Heuristically then, if we let $\mathcal{S}^+(t) := \{j \colon W_j \geq t\}$ and $\mathcal{S}^-(t) := \{j \colon W_j \leq -t\}$, it is natural to construct the following "conservative" estimate $\widehat{\mathrm{FDP}}(t)$ of $\mathrm{FDP}(t)$:

$$FDP(t) = \frac{|\{j \in \mathcal{H}_0 \colon W_j \geq t\}|}{|\mathcal{S}^+(t)| \vee 1} \approx \frac{|\{j \in \mathcal{H}_0 \colon W_j \leq -t\}|}{|\mathcal{S}^+(t)| \vee 1} \leq \frac{|\mathcal{S}^-(t)|}{|\mathcal{S}^+(t)| \vee 1} =: \widehat{\mathrm{FDP}}(t).$$

Now we can use this estimate $\widehat{\mathrm{FDP}}(t)$ to select control variables via a step-down rule. This is called the `SeqStep` procedure [1]: Fist, order the W_j's from largest to smallest in magnitude. Then, starting with the largest $|W_j|$, step downward, count the number of negative vs. positive W_j's we encounter and reject each H_j corresponding to a positive W_j until the last $|W_j|$ such that the ratio of the number of negative W_j's counted to the number of positive W_j's counted is less than the desired FDR level q. We illustrate this procedure in Figure 10.5.

More formally, we can define the set of selected variables $\hat{\mathcal{S}}$, or equivalently, the set of rejected hypotheses, as follows (note that we use a slightly more conservative estimate of $\mathrm{FDP}(t)$ in the definition below):

$$\hat{\mathcal{S}} := \{j : W_j \geq \tau_q\}, \quad \tau_q := \min\left\{t : \widehat{\mathrm{FDP}}(t) := \frac{1 + |\mathcal{S}^-(t)|}{|\mathcal{S}^+(t)| \vee 1} \leq q\right\}. \tag{10.1}$$

As it happens, it can be shown that this procedure controls FDR:

Theorem 1 ([1])**.** The knockoffs procedure defined in (10.1) controls FDR.

We defer a proof of Theorem 1 to Section 10.5.

Remark 1. In fact, even if P_X is not known exactly and knockoffs are only exchangeable with respect to an estimate Q_X of P_X, the same knockoffs procedure still controls FDR; see [2] for details.

10.3 Statistical modeling: what do we know?

Before proving Theorem 1, we discuss the model-X knockoffs procedure in context with classical variable selection methods. Note that some important benefits of the knockoffs framework are that, by Theorem 1, it always controls FDR in finite samples regardless of the dimension p of X, the distribution $Y \mid X$, and the black-box model of $P_{Y|X}$ used to compute importance scores Z_j. In contrast, classical approaches to the control variable selection problem often require a strong model of $P_{Y|X}$, e.g. a (generalized) linear model, so inference is typically only valid if at least some of those assumptions (e.g. linearity of some function of the conditional expectation $\mathbb{E}[Y|X]$) hold and often only if $n \to \infty$. By shifting the burden of knowledge from $P_{Y|X}$ (e.g. the distribution of phenotypes given genotypes), which is almost never known, to P_X, which is much more commonly known (e.g. the distribution of genotypes), the model-X knockoffs procedure makes it much easier to do credible statistical inference in a variety of potentially high-dimensional settings while leveraging state-of-the-art statistical learning tools for measuring variable importance. Note also that instead of treating observations of X as fixed as in classical models, and making inference conditional on the observed value of X, the Knockoffs frameworks treat observations of X as random, which is often more appropriate in 'big data' applications, e.g. SNPs of subjects randomly sampled.

10.4 Applications in genetics

There are some applications of knockoff method in genetics:

- Sesia, Katsevich, Bates, Candès, Sabatti (2020) "Multi-resolution localization of causal variants across the genome," Nature Communications [5]
- Sesia, Bates, Candès, Marchini, Sabatti (2021) "Controlling the false discovery rate in GWAS with population structure" [4]

In the application of the knockoff frameworks to genetics studies, the key idea is to model the randomness we already understand (P_X) and make inferences about the unknown $P_{Y|X}$. First of all, in accordance with the exchangeability property established in Section 10.2.2, knockoffs preserve important information and properties of the original data. Figure 10.6 displays the similarity in population structure, kinship and linkage disequilibrium between the original and knockoff version of the UKBiobank dataset, which contains 592,000 Single Nucleotide Polymorphisms (SNPs) genotyped on 489,000 individuals. Moreover, Knockoffs sometimes speak more directly to scientific questions of interest. Figure 10.7 compares knockoff analysis to marginal analysis. Since the nearby location of genome would be highly correlated, it's difficult to tell which one is important. A possible solution is to group nearby variables as a set G and look at a simpler problem: whether Y is conditional independent of G given everything else, i.e. $Y \perp\!\!\!\perp X_G | X_{-G}$. Hence, we could have a knockoff discovery scheme that runs at various resolutions: from the smallest resolution to the highest resolution (individual snips). Finally, Knockoffs frameworks display highly desirable locus discovery performance. Figure 10.8 shows that Knockoff GWAS is as powerful as oracle based on BOLT-LMM, controls the FDR, and makes more precise discoveries.

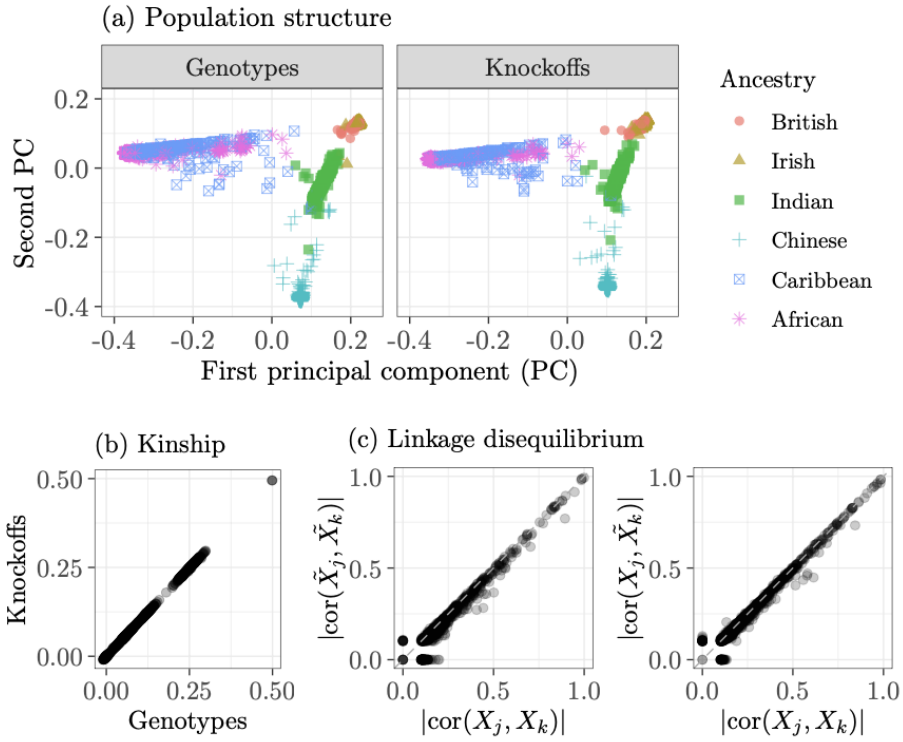
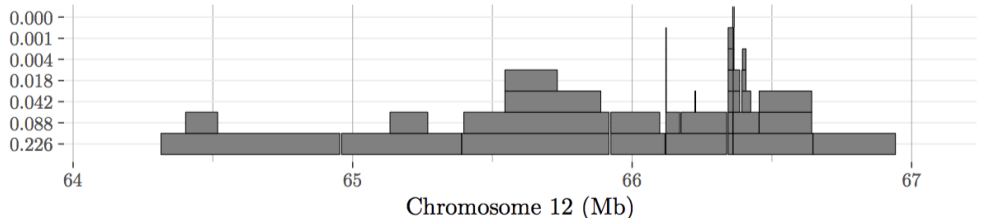
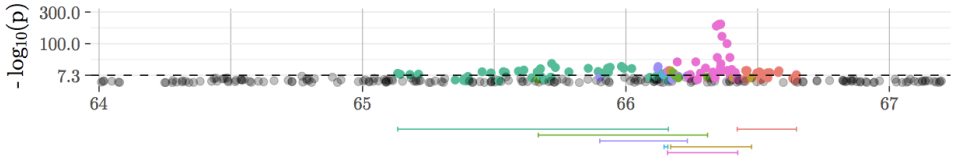


Figure 10.6: Comparison between original and knockoff version of the UKBiobank data.



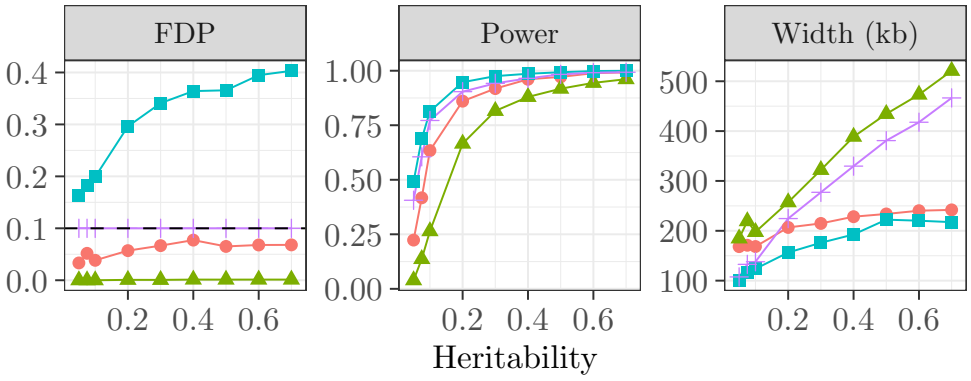
(a) Findings of Knockoff analysis.



(b) Findings of Marginal analysis.

Figure 10.7: Findings of Knockoff analysis (Figure 10.7a) is sometimes referred to as Manhattan plot (where y-axis is Resolution (kb)) and Marginal analysis (Figure 10.7b) Chicago plot.

(a) Locus discovery at low resolution (208 kb)



(b) Distinct discoveries at resolution with most findings

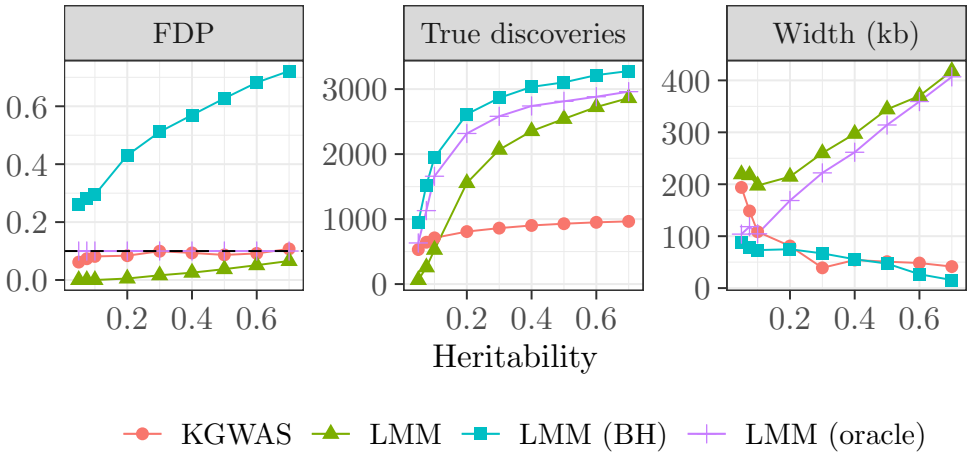


Figure 10.8: Locus discovery performance of Knockoff GWAS, LMM, LMM(BH), and LMM(oracle).

10.5 Proof Sketch of Theorem 1

We conclude by proving Theorem 1 in a similar manner to the empirical process viewpoint-based proof that the BH procedure controls FDR. First, since we reject all H_j such that $j \in \mathcal{S}^+(\tau_q)$, we can write the FDP using the threshold τ_q defined in (10.1) as follows:

$$\begin{aligned}
\mathrm{FDP}(\tau_q) &= \frac{|\mathcal{H}_0 \cap \mathcal{S}^+(\tau_q)|}{|\mathcal{S}^+(\tau_q)| \vee 1} \\
&= \frac{|\mathcal{H}_0 \cap \mathcal{S}^+(\tau_q)|}{1 + |\mathcal{H}_0 \cap \mathcal{S}^-(\tau_q)|} \cdot \frac{1 + |\mathcal{H}_0 \cap \mathcal{S}^-(\tau_q)|}{|\mathcal{S}^+(\tau_q)| \vee 1} \\
&\leq \frac{|\mathcal{H}_0 \cap \mathcal{S}^+(\tau_q)|}{1 + |\mathcal{H}_0 \cap \mathcal{S}^-(\tau_q)|} \cdot \frac{1 + |\mathcal{S}^-(\tau_q)|}{|\mathcal{S}^+(\tau_q)| \vee 1} \\
&\leq q \cdot \frac{|\mathcal{H}_0 \cap \mathcal{S}^+(\tau_q)|}{1 + |\mathcal{H}_0 \cap \mathcal{S}^-(\tau_q)|}. \qquad \text{(by (10.1))}
\end{aligned}$$

As such, let $V^+(\tau_q) := |\mathcal{H}_0 \cap \mathcal{S}^+(\tau_q)|$ and $V^-(\tau_q) := |\mathcal{H}_0 \cap \mathcal{S}^-(\tau_q)|$. Then, to show that $\mathrm{FDR}(\tau_q) = \mathbb{E}[\mathrm{FDP}(\tau_q)] \leq q$, it suffices to show that

$$\mathbb{E}\left[\frac{|\mathcal{H}_0 \cap \mathcal{S}^+(\tau_q)|}{1 + |\mathcal{H}_0 \cap \mathcal{S}^-(\tau_q)|}\right] = \mathbb{E}\left[\frac{V^+(\tau_q)}{1 + V^-(\tau_q)}\right] \leq 1. \tag{10.2}$$

Next, akin to the argument used in the empirical process perspective-based proof that the BH procedure controls FDR, we will argue that $V^+(t)/(1+V^-(t))$ is a supermartingale with resepct to the filtration $\mathcal{F}_t := \{\sigma(V^\pm(u))\}_{u \leq t}$ with t increasing from 0 (instead of decreasing from 1 as in Lecture 7) so that we can apply Doob's Optional Stopping Theorem. Consider any $s \geq t$, and note that, conditional on $V^+(s) + V^-(s)$, $V^+(s)$ has a hypergeometric distribution. Then it is straightforward to show that

$$\mathbb{E}\left[\frac{V^+(s)}{1 + V^-(s)} \;\middle|\; V^\pm(t), V^+(s) + V^-(s)\right] \leq \frac{V^+(t)}{1 + V^-(t)},$$

which is exactly what is required for $V^+(t)/(1 + V^-(t))$ to be a supermartingale. As a prelude to applying Doob's Optional Stopping Theorem, recall that if $Y \sim \mathrm{Bin}(n_0, 1/2)$, $\mathbb{E}[Y/(1 + n_0 + Y)] \leq 1$. Since by Lemma 2, $V^+(0) \sim \mathrm{Bin}(|\mathcal{H}_0|, 1/2)$, then by combining Doob's Optional Stopping Theorem with this fact, we have that

$$\mathrm{FDR} \leq q\mathbb{E}\left[\frac{V^+(\tau_q)}{1 + V^-(\tau_q)}\right] \leq q\mathbb{E}\left[\frac{V^+(0)}{1 + |\mathcal{H}_0| - V^+(0)}\right] \leq q,$$

as required.

Bibliography

[1] Rina Foygel Barber and Emmanuel J Candès. Controlling the false discovery rate via knockoffs. *Annals of Statistics*, 43(5):2055–2085, 2015.

[2] Rina Foygel Barber, Emmanuel J Candès, and Richard J Samworth. Robust inference with knockoffs. *arXiv preprint arXiv:1801.03896*, 2018.

[3] Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: "model-x" knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577, 2018.

[4] Matteo Sesia, Stephen Bates, Emmanuel Candès, Jonathan Marchini, and Chiara Sabatti. Controlling the false discovery rate in gwas with population structure. *bioRxiv*, 2020.

[5] Matteo Sesia, Eugene Katsevich, Stephen Bates, Emmanuel Candès, and Chiara Sabatti. Multi-resolution localization of causal variants across the genome. *Nature Communications*, 11(1):1093, 2020.