Knockoff filter

Patrick Breheny

April 15, 2025

Introduction

- Today we will discuss one final approach to inference in high-dimensional regression models called the *knockoff filter*
- There are two approaches to the knockoff filter:
 - $\circ~$ In its simplest form, we can generate knockoffs without any assumptions on ${\bf X};$ however this approach only works if ${\bf X}$ is full rank (Barber and Candès 2015)
 - A later paper (Candès et al. 2018) extended this idea to the p>n case, although in order to do so, we need to make some assumptions about ${\bf X}$
- Both approaches are implemented in the R package knockoff

Procedure Constructing the knockoffs Theoretical properties

Step 1: Construct knockoffs

- The basic idea of the knockoff filter is that for each feature \mathbf{x}_j in the original feature matrix, we construct a *knockoff* feature $\tilde{\mathbf{x}}_j$
- We'll go into specifics on constructing knockoffs later; for now, we specify the properties that a knockoff x_j must have:

$$\begin{split} \widetilde{\mathbf{X}}^{\top} \widetilde{\mathbf{X}} &= \mathbf{X}^{\top} \mathbf{X} \\ \widetilde{\mathbf{x}}_{j}^{\top} \mathbf{x}_{k} &= \mathbf{x}_{j}^{\top} \mathbf{x}_{k} & \text{ for all } k \neq j \\ \frac{1}{n} \widetilde{\mathbf{x}}_{j}^{\top} \mathbf{x}_{j} &= 1 - s_{j} & \text{ where } 0 \leq s_{j} \leq 1 \end{split}$$

 In other words, the knockoff matrix X differs from the original matrix X, but has the same correlation structure and the same correlation with the original features

Procedure Constructing the knockoffs Theoretical properties

Step 2: Calculate test statistics

- With the knockoffs constructed, the next step is to fit a (lasso) model to the augmented $n \times 2p$ design matrix $[\mathbf{X} \ \widetilde{\mathbf{X}}]$
- At this point, we need some sort of test statistic that measures whether the original feature is better than the knockoff
- There are actually a variety of statistics we could use here, but in this lecture we'll focus on the point λ along the lasso path at which a feature enters the model, giving us a 2p-dimensional vector $\{Z_1, \ldots, Z_p, \tilde{Z}_1, \ldots, \tilde{Z}_p\}$
- Our test statistic is then

$$W_j = \max(Z_j, \tilde{Z}_j) \cdot \operatorname{sign}(Z_j - \tilde{Z}_j);$$

i.e., W_{j} will be positive if the original feature is selected before the knockoff, and negative if the knockoff is selected first

Procedure Constructing the knockoffs Theoretical properties

Step 3: Estimate false discovery rate

- Now, if we select features such that $W_j \ge t$ for some threshold t, we can use the knockoff features to estimate the false discovery rate
- Specifically, our knockoff estimate of the FDR is:

$$\widehat{\mathrm{FDR}} = \frac{\#\{j: W_j \le -t\}}{\#\{j: W_j \ge t\}},$$

with the understanding that $\widehat{FDR}=1$ if the numerator is larger than the denominator, or if the denominator is zero

• Typically, we would specify the desired FDR q and then choose t to be the smallest value satisfying $\widehat{\mathrm{FDR}}(t) \leq q$

Procedure Constructing the knockoffs Theoretical properties

Illustration: Augmented example data (n = 200, p = 60)



Procedure Constructing the knockoffs Theoretical properties

Power and $\{s_j\}$

- So, how do we actually construct these knockoffs?
- As we will see, the knockoff filter is valid provided that the knockoffs have the correlation structure outlined earlier; its power, however, depends on {s_j}
- For the greatest power, we want the knockoffs to be as different from the original features as possible (i.e, we want the $\{s_j\}$ terms to be as large as possible)

Procedure Constructing the knockoffs Theoretical properties

Nullspace, n, and p

- Let N denote an n × p orthonormal matrix such that N^TX = 0 (in other words, Nα lies within the column null space of X; note that this can be constructed using the QR decomposition)
- Note that the nullspace of \mathbf{X} has dimension $n \operatorname{rank}(\mathbf{X})$
- Thus, for the matrix N to exist, it is not enough for X to be full rank; we also need $n \ge p + \operatorname{rank}(\mathbf{X})$, so $n \ge 2p$ in the full-rank case

Procedure Constructing the knockoffs Theoretical properties

Constructing knockoffs under equal correlation

- So, let's say we have a full rank X with $n \ge 2p$ and thus can construct an orthonormal N with $N^{T}X = 0$
- Furthermore, suppose we require $s_j = s$ for all j and let $\frac{1}{n} \mathbf{C}^{\top} \mathbf{C} = 2s \mathbf{I} s^2 \mathbf{\Sigma}^{-1}$, where $\mathbf{\Sigma} = \frac{1}{n} \mathbf{X}^{\top} \mathbf{X}$
- **Proposition:** The matrix

$$\widetilde{\mathbf{X}} = \mathbf{X}(\mathbf{I} - s\boldsymbol{\Sigma}^{-1}) + \mathbf{N}\mathbf{C}$$

satisfies the requirements of a knockoff matrix

Procedure Constructing the knockoffs Theoretical properties

The non-full rank case

- What if **X** is not full rank?
- It turns out that the maximum value for s is 2 times the minimum eigenvalue of Σ; thus, s_j = s for all j cannot work in the case where X is not full rank
- In this case, we will have to set some of the $s_j = 0$ (meaning no power for those features) and try to maximize the rest as best we can
- In the knockoff package, a semidefinite programming approach is used to determine the values that minimize $\sum_j (1-s_j)$ subject to the constraints (method='sdp'; the earlier approach is method='equi')

Procedure Constructing the knockoffs Theoretical properties

The p < n < 2p case

- Now, what if \mathbf{X} is full rank, but n < 2p?
- In this case, there is an interesting little data augmentation trick that can be used, provided that σ^2 can be estimated accurately
- To get our sample size up to 2p, we can generate 2p nadditional rows of \mathbf{X} that are simply all equal to $\mathbf{0}$ and 2p - nadditional entries for \mathbf{y} that are drawn from a $N(0, \hat{\sigma}^2)$ distribution
- We now have a linear model with p features and 2p observations; the new observations carry no information about β, but are useful for generating knockoffs

Procedure Constructing the knockoffs Theoretical properties

p < n < 2p data augmentation applied to example data



Procedure Constructing the knockoffs Theoretical properties

FDR control

- So does this knockoff procedure actually control the FDR?
- Note quite; instead, Barber and Candès show that it controls a modified version of the FDR:

$$\mathbb{E}\left(\frac{|\mathcal{N}\cap\hat{\mathcal{S}}|}{|\hat{\mathcal{S}}|+q^{-1}}\right) \leq q,$$

where $\hat{\mathcal{S}}$ is the set of features selected by the knockoff filter

- Alternatively, the knockoff filter controls the FDR if we add 1 to the numerator (i.e., to the number of knockoffs selected)
- The modifications have little effect if many features are selected

Procedure Constructing the knockoffs Theoretical properties

Coin flip lemma

- We won't go through the entire proof here, but just present a sketch of the main ideas
- The critical property that knockoffs have is a "coin flipping property": for j ∈ N, we have sign(W_j) [⊥]∼ Bern(1/2)
- This coin flipping property derives from two exchangeability results:
 - $\circ~[X~\widetilde{X}]^{\scriptscriptstyle \top}[X~\widetilde{X}]$ is invariant to any exchange of original and knockoff features
 - The distribution of $[\mathbf{X} \ \widetilde{\mathbf{X}}]^{\top} \mathbf{y}$ is invariant to any exchange of *null* original and knockoff features

Procedure Constructing the knockoffs Theoretical properties

Sketch of proof

• With these lemmas in place, the FDR control proof follows from the inequality

$$FDR \le q \cdot \frac{\#\{j: \beta_j = 0 \text{ and } W_j > t\}}{1 + \#\{j: \beta_j = 0 \text{ and } W_j < -t\}};$$

the coin flipping property ensuring that the expected value of this quantity is below \boldsymbol{q}

• The argument can be extended to a random threshold T through use of martingales and the optional stopping theorem similar to our FDR proof at the beginning of the course

$\mathsf{Modeling}\ \mathbf{X}$

- An obvious shortcoming of the previous approach is that it requires $n \geq p$
- Extending the idea to p > n situations requires us to treat X as random and to model its distribution; Candès et al. refer to these as "model-X knockoffs" or just "MX" knockoffs
- Note that this is an interesting philosophical shift: the classical setup is to assume a very specific distribution for y but assume as little as possible about X, whereas MX knockoffs assume that we know everything about the distribution of X but require no assumptions on the distribution of Y | X

Knockoff properties in the random case

- Recall our exchangeability results from earlier; with these in mind, we can define knockoff conditions in the case where ${\bf X}$ is treated as a random matrix with IID rows
- A knockoff matrix $\widetilde{\mathbf{X}}$ satisfies
 - $\circ~$ The distribution of $[X~\tilde{X}]$ is invariant to any exchange of original and knockoff features
 - $\circ \ \tilde{X} \perp \!\!\!\perp Y | X$
- Note that the second condition is guaranteed if $\widetilde{\mathbf{X}}$ is constructed without looking at \mathbf{y}

Concluding remarks

Gaussian case

- There are special cases in which we actually know something about the distribution of **X**; in general, however, we would likely assume it follows a multivariate normal distribution
- The main challenge here is that now we must estimate Σ , a $p \times p$ covariance matrix, or rather Σ^{-1} , the precision matrix
- We will (time permitting) discuss this problem a bit later in the course; for now, although this is by no means trivial, let us assume that we can estimate Σ well enough to assume that we know $X \sim N(0, \Sigma)$

MX knockoffs in the Gaussian case

• In order to satisfy the knockoff property, let us assume the joint distribution $[X\ \tilde{X}]\sim N({\bf 0},{\bf G})$ where

$$\mathbf{G} = \left[egin{array}{cc} \mathbf{\Sigma} & \mathbf{\Sigma} - \mathbf{S} \\ \mathbf{\Sigma} - \mathbf{S} & \mathbf{\Sigma} \end{array}
ight];$$

here **S** is a diagonal matrix with entries $\{s_j\}$

• Now, we can draw a random $\widetilde{\mathbf{X}}$ from the conditional distribution $\widetilde{X}|X$, which is normal with

$$\mathbb{E}(\tilde{X}|X) = X - \mathbf{S}\mathbf{\Sigma}^{-1}X$$
$$\mathbb{V}(\tilde{X}|X) = 2\mathbf{S} - \mathbf{S}\mathbf{\Sigma}^{-1}\mathbf{S}$$

Concluding remarks

Example data with modeled $\mathbf X$



TCGA data

- I tried applying the MX knockoff approach to the TCGA data using the knockoff package, but this crashed, presumably due to the memory limitations of dealing with a $17,322\times17,322$ matrix
- I even tried running it on our HPC cluster, but this also crashed
- However, it is worth noting that in their paper, Candès et al. applied the MX knockoff filter to a problem with p = 400,000 by taking advantage of a special correlation structure in ${\bf X}$

Remarks: Some advantages

- The knockoff filter also has some nice advantages
- In particular, none of its theory involves any asymptotics, or anything special about the statistic W, or about the lasso, which means:
 - The theory holds exactly in finite dimensions
 - We can use other statistics, such as the lasso coefficient difference: $W_j = |\hat{\beta}_j(\lambda)| |\hat{\beta}_{j+p}(\lambda)|$
 - Perhaps most appealing, we can apply this reasoning to all kinds of other methods – other penalties of course, but also much more ambitious problems: forward selection, random forests, even deep learning

Concluding remarks

Remarks: Some drawbacks

- Result can differ quite a bit depending on the random X one draws; it would seem desirable to aggregate or average these results over the draws, although how exactly to do this is unclear
- Furthermore, scaling the method to high dimensions is not trivial
- Finally, knockoffs appear to be slightly less powerful than some of the other approaches we have discussed

Gaussian mirrors

- A related idea, intended to remedy some of these issues with the knockoff filter, is that of the *Gaussian mirror* (Xing et al., 2023)
- The idea is that for each feature \mathbf{x}_j , we create a pair of "mirror features": $\mathbf{x}_j^+ = \mathbf{x}_j + c_j \mathbf{z}_j$ and $x_j^- = \mathbf{x}\mathbf{j} c_j \mathbf{z}_j$, where c_j is a scalar and $\mathbf{z}_j \sim N(\mathbf{0}, \mathbf{I})$
- The obvious advantages over knockoffs is that we're only perturbing one variable at a time, so
 - $\,\circ\,$ Easier to scale up to high dimensions with p>n
 - $\,\circ\,$ No need to model the joint distribution of all p features

Mirror statistic

- To carry out a test of $H_0: \beta_j^* = 0$, we first construct a new feature matrix \mathbf{X}^j that consists of \mathbf{X}_{-j} plus the mirror features for \mathbf{x}_j and fit the model
- We then construct the *mirror statistic*:

$$M_j = \left|\widehat{\beta}_j^+ + \widehat{\beta}_j^-\right| - \left|\widehat{\beta}_j^+ - \widehat{\beta}_j^-\right|$$

- The first term represents signal while the second represents noise; roughly speaking, in the first term the noise cancels out while in the second term the signal cancels out
- This would then be repeated for all j

FDR for Gaussian mirror

- In the interest of time, I'll skip the details, but it is possible to choose c_j such that the distribution of M_j is symmetric about zero when the null hypothesis is true
 - This is relatively straightforward for OLS
 - Much more complicated for lasso
- Similar to the knockoff filter, we estimate the FDR among selected features to with M_j ≥ t by calculating

$$\widehat{\mathsf{FDR}} = \frac{\#\{j: M_j \le -t\}}{\#\{j: M_j \ge t\}},$$

again with the understanding that $\widehat{\mathrm{FDR}}=1$ if the denominator is zero