Controlled Variable Selection: Knockoff Filters

Yibin Xiong

May 27, 2022

Table of Contents

1. Knockoffs
2. Model-X Knockoffs
3. Knockoffs for HMM
4. Deep Knockoffs

Controlled Variable Selection Problem

▷ Too many covariates

We want to find out which explanatory variables are "truly" associated with the response variable, in the sense of statistical significance.

i.e. $\forall j$ test $H_0^{(j)} : Y \perp\!\!\!\perp X_j \,|\, X_{-j}$. This is to find the *Markov blanket*, which retains information with minimal number of variables.

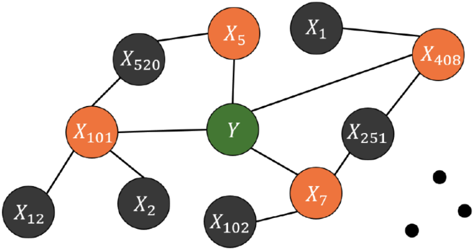


Figure: Markov blanket $\mathcal{S} : Y \perp\!\!\!\perp \{X_j\}_{j \notin \mathcal{S}} \mid \mathcal{S}$

Controlled Variable Selection Problem

Now we have a multiple hypothesis testing problem

$\triangleright$ Avoid many false positives: control the **false discovery rate** (FDR)

Let $\hat{S}$ be the set of covariates we selected and $\mathcal{H}_0$ be the set of “null” variables, which are defined by the individual null hypotheses.

$$\mathsf{FDR} := \mathbb{E}\left[\mathsf{FDP}\right] = \mathbb{E}\left[\frac{|\hat{S} \cap \mathcal{H}_0|}{\max\{|\hat{S}|, 1\}}\right]$$

Controlled Variable Selection Problem

Now we have a multiple hypothesis testing problem

$\triangleright$ Avoid many false positives: control the **false discovery rate** (FDR)

Let $\hat{S}$ be the set of covariates we selected and $\mathcal{H}_0$ be the set of "null" variables, which are defined by the individual null hypotheses.

$$\mathrm{FDR} := \mathbb{E}\left[\mathrm{FDP}\right] = \mathbb{E}\left[\frac{|\hat{S} \cap \mathcal{H}_0|}{\max\{|\hat{S}|, 1\}}\right]$$

$\circ$ More liberal than controlling the family-wise error rate $\mathbb{P}\left(|\hat{S} \cap \mathcal{H}_0| \geq 1\right)$

$\circ$ Making a false discovery, for example in genome-wide association studies, does not have high real-world costs.

Idea of Knockoffs

$\triangleright$ A "copy" of covariate X_j that *preserves the correlation structures* and serves as a *negative control* (i.e. no correlation with the response Y)

- "Geometric" assumptions [1]

$$\begin{aligned} \tilde{X}_j^T \tilde{X}_k &= X_j^T X_k = 1 \ \forall j, k \\ X_j^T \tilde{X}_k &= X_j^T X_k \qquad \forall j \neq k \\ X_j^T \tilde{X}_j &= 1 - s_j \quad \text{for some } s_j \text{ close to 1} \end{aligned}$$

In matrix form, this is $\tilde{X}^T \tilde{X} = X^T X = \Sigma$, $X^T \tilde{X} = \Sigma - \text{diag}\{s\}$

Idea of Knockoffs

▷ A "copy" of covariate X_j that *preserves the correlation structures* and serves as a *negative control* (i.e. no correlation with the response Y)

- "Geometric" assumptions [1]

$$\begin{aligned} \tilde{X}_j^T \tilde{X}_k &= X_j^T X_k = 1 \ \forall j, k \\ X_j^T \tilde{X}_k &= X_j^T X_k \qquad \forall j \neq k \\ X_j^T \tilde{X}_j &= 1 - s_j \quad \text{for some } s_j \text{ close to } 1 \end{aligned}$$

In matrix form, this is $\tilde{X}^T \tilde{X} = X^T X = \Sigma$, $X^T \tilde{X} = \Sigma - \mathrm{diag}\{s\}$

▷ Compute a statistics that indicates relative importance of the covariate versus its knockoff.

▷ Select variables sequentially until the empirical estimate of FDR hits a threshold.

Main Example: Low-dimensional Linear Regression Model

Assume $n \geq 2p$ and $y = X\beta + z$, where $z \sim \mathcal{N}(0, \sigma^2 I)$.

▷ Step 1: Construct the knockoffs

We want

$$\begin{bmatrix} X & \tilde{X} \end{bmatrix}^T \begin{bmatrix} X & \tilde{X} \end{bmatrix} = \begin{bmatrix} \Sigma & \Sigma - \mathrm{diag}\{s\} \\ \Sigma - \mathrm{diag}\{s\} & \Sigma \end{bmatrix} \succeq 0$$

By Schur complement calculation, we can construct

$$\tilde{X} := X(I - \Sigma^{-1}\mathrm{diag}\{s\}) + \tilde{U}C$$

where $\tilde{U}$ orthogonal, $\tilde{U}^T X = 0$, and $C^T C = 2\mathrm{diag}\{s\} - \mathrm{diag}\{s\}\Sigma^{-1}\mathrm{diag}\{s\}$.

[1] $\wedge$ denotes min and $\vee$ denotes max

Main Example: Low-dimensional Linear Regression Model

Assume $n \geq 2p$ and $y = X\beta + z$, where $z \sim \mathcal{N}(0, \sigma^2 I)$.

▷ Step 1: Construct the knockoffs

We want

$$\begin{bmatrix} X & \tilde{X} \end{bmatrix}^T \begin{bmatrix} X & \tilde{X} \end{bmatrix} = \begin{bmatrix} \Sigma & \Sigma - \mathrm{diag}\{s\} \\ \Sigma - \mathrm{diag}\{s\} & \Sigma \end{bmatrix} \succeq 0$$

By Schur complement calculation, we can construct

$$\tilde{X} := X(I - \Sigma^{-1}\mathrm{diag}\{s\}) + \tilde{U}C$$

where $\tilde{U}$ orthogonal, $\tilde{U}^T X = 0$, and $C^T C = 2\mathrm{diag}\{s\} - \mathrm{diag}\{s\}\Sigma^{-1}\mathrm{diag}\{s\}$.

How to choose s?

- Equi-correlated knockoffs: $\forall j : s_j := 2\lambda_{\min}(\Sigma) \wedge 1$ [1]
- SDP knockoffs: $\min_{s_j} \sum_j 1 - s_j \quad \text{s.t. } 0 \leq s_j \leq 1,\ \mathrm{diag}\{s\} \preceq 2\Sigma$

[1] $\wedge$ denotes min and $\vee$ denotes max

Main Example: Low-dimensional Linear Regression Model

▷ Step 2: Compute the statistics for each pair of original and knockoff variables

We consider LASSO coefficients that enable us to screen out the important variables:

$$\hat{\beta}(\lambda) = \arg\min_{b} \frac{1}{2} \|y - \begin{bmatrix} X & \tilde{X} \end{bmatrix} b\|_2^2 + \lambda \|b\|_1$$

As λ gets larger, we focus more on the penalization and the solution tends to be more sparse and leaves only the important variables.

Main Example: Low-dimensional Linear Regression Model

▷ Step 2: Compute the statistics for each pair of original and knockoff variables

We consider LASSO coefficients that enable us to screen out the important variables:

$$\hat{\beta}(\lambda) = \arg\min_b \frac{1}{2} \|y - \begin{bmatrix} X & \tilde{X} \end{bmatrix} b\|_2^2 + \lambda \|b\|_1$$

As λ gets larger, we focus more on the penalization and the solution tends to be more sparse and leaves only the important variables.

We define $Z_j := \sup\{\lambda : \hat{\beta}_j(\lambda) \neq 0\}$ as an importance score for each covariate and its knockoff. Then define

$$W_j := Z_j \vee \tilde{Z}_j \cdot \begin{cases} +1, & Z_j > \tilde{Z}_j \\ -1, & Z_j < \tilde{Z}_j \\ 0, & Z_j = \tilde{Z}_j \end{cases}$$

Main Example: Low-dimensional Linear Regression Model

▷ Step 3: Define a data-dependent threshold for the statistics

We specify a level q for controlling the FDR (e.g. $q = 0.05$).

- knockoff:

$$T := \min\left\{ t : \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q \right\}$$

- knockoff+:

$$T := \min\left\{ t : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q \right\}$$

Main Example: Low-dimensional Linear Regression Model

▷ Step 3: Define a data-dependent threshold for the statistics

We specify a level q for controlling the FDR (e.g. $q = 0.05$).

- knockoff:

$$T := \min\left\{t : \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q\right\}$$

- knockoff+:

$$T := \min\left\{t : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq q\right\}$$

▷ Step 4: Sequential selection

Arrange the covariates by descending order of $|W_j|$, and select the covariates that correspond to positive W_j until the quantity exceeds q.

Theoretical Guarantees: Assumptions

▷ What we assumed

- $n \geq 2p$ (can adapt to $n \geq p$)
- linear regression model $F_{Y|X} = \mathcal{N}(X\beta, \sigma^2 I)$, used when computing Z_j and showing exchangebility results
- geometric assumptions
- sufficiency of the statistics:

  W ONLY depends on the Gram matrix and the feature-response inner product

  $$W = f(\begin{bmatrix} X & \tilde{X} \end{bmatrix}^T \begin{bmatrix} X & \tilde{X} \end{bmatrix}, \begin{bmatrix} X & \tilde{X} \end{bmatrix}^T y)$$
- antisymmetry of the statistics: $W_j(Z_j, \tilde{Z}_j) = -W_j(\tilde{Z}_j, Z_j)$

  swapping X_j and $\tilde{X}_j$ in the augmented design matrix changes the sign of W_j

  $$\forall \, \text{set } S : \forall j \in S : W_j\left(\begin{bmatrix} X & \tilde{X} \end{bmatrix}_{\text{swap}(S)}, y\right) = -W_j\left(\begin{bmatrix} X & \tilde{X} \end{bmatrix}, y\right)$$

Theoretical Guarantees: Exchangeability

Exchangeability result

For any subset S of the *null* variables,

$$\begin{bmatrix} X & \tilde{X} \end{bmatrix}_{\mathrm{swap}(S)}^T \begin{bmatrix} X & \tilde{X} \end{bmatrix}_{\mathrm{swap}(S)} = \begin{bmatrix} X & \tilde{X} \end{bmatrix}^T \begin{bmatrix} X & \tilde{X} \end{bmatrix} \tag{1}$$

$$\begin{bmatrix} X & \tilde{X} \end{bmatrix}_{\mathrm{swap}(S)}^T y \overset{d}{=} \begin{bmatrix} X & \tilde{X} \end{bmatrix}^T y \tag{2}$$

∘ (1) is a direct consequence of the geometric assumptions, i.e. by our construction.

∘ Proof of (2) uses the geometric assumption and linear homoskedastic model.

∘ Exchangeability, combined with antisymmetry of the statistics, is the key for showing i.i.d. signs of null statistics because $M \overset{d}{=} M' \Rightarrow f(M) \overset{d}{=} f(M')$.

Theoretical Guarantees: A Key Property

i.i.d. signs of null statistics

For any null variable j, conditional on $|W| = (|W_1|, \ldots, |W_p|)$,

$$W_j \overset{d}{=} -W_j \tag{3}$$

$$\#\{j : \beta_j = 0, W_j \leq -t\} \overset{d}{=} \#\{j : \beta_j = 0, W_j \geq t\} \tag{4}$$

∘ This makes intuitive sense since we expect W_j to be a large positive value if X_j is a signal and W_j to be close to 0 if X_j is a null variable. For a null X_j, It should be equally very unlikely for W_j to be a positive large number or to be a negative large number.

∘ This gives us a good proxy for the number of false positives and the empirical FDR

$$\frac{\#\{j : \beta_j = 0, W_j \geq t\}}{\#\{j : W_j \geq t\} \vee 1} \approx \frac{\#\{j : \beta_j = 0, W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1} \leq \frac{\#\{j : W_j \leq -t\}}{\#\{j : W_j \geq t\} \vee 1}$$

Theoretical Guarantees: FDR Control

For knockoff+, we can guarantee that the FDR is at most q.

◦ Since T is a stopping time, by optional stopping theorem of supermartingales

$$\mathbb{E}\left[\frac{\#\{j:\beta_j=0, W_j\geq T\}}{1+\#\{j:\beta_j=0, W_j\leq -T\}}\right]\leq \mathbb{E}\left[\frac{\#\{j:\beta_j=0, W_j\geq 0\}}{1+\#\{j:\beta_j=0, W_j\leq 0\}}\right]<1$$

◦ Main inequalities

$$\begin{aligned}\mathrm{FDR} &= \mathbb{E}\left[\frac{\#\{j:\beta_j=0, W_j\geq T\}}{\#\{j:W_j\geq T\}\vee 1}\right]\\ &= \mathbb{E}\left[\frac{\#\{j:\beta_j=0, W_j\geq T\}}{1+\#\{j:\beta_j=0, W_j\leq -T\}}\cdot\frac{1+\#\{j:\beta_j=0, W_j\leq -T\}}{\#\{j:W_j\geq T\}\vee 1}\right]\\ &\leq \mathbb{E}\left[\frac{\#\{j:\beta_j=0, W_j\geq T\}}{1+\#\{j:\beta_j=0, W_j\leq -T\}}\right]\cdot q\\ &\leq q\end{aligned}$$

Table of Contents

1. Knockoffs
2. Model-X Knockoffs
3. Knockoffs for HMM
4. Deep Knockoffs

Model-X Knockoffs

▷ Assumptions [2]:

- Exchangeability: $(X, \tilde{X})_{\text{swap}(S)} \overset{d}{=} (X, \tilde{X})$
- Conditional independence: $\tilde{X} \perp\!\!\!\perp Y \mid X$
- Distribution F_X, used for constructing $\tilde{X}$

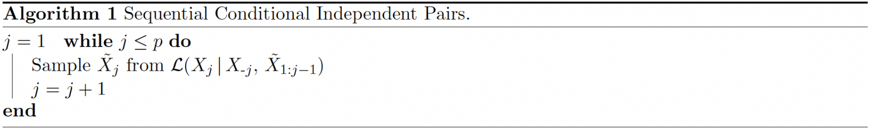
Model-X Knockoffs

- Assumptions [2]:
  - Exchangeability: $(X, \tilde{X})_{\text{swap}(S)} \overset{d}{=} (X, \tilde{X})$
  - Conditional independence: $\tilde{X} \perp\!\!\!\perp Y \mid X$
  - Distribution F_X, used for constructing $\tilde{X}$
- Improvements over vanilla knockoffs
  - Doesn't assume $F_{Y|X}$, which we don't exactly know but have to use models to approximate
  - Unsupervised data can be used
  - No geometric assumption, can be applied to *high-dimensional* settings
  - No sufficiency condition for W_j
  - Not just for homoscedastic linear model, but any regression and classification models

Constructing Knockoffs [2]

▷ If the components of X are independent, then we can sample $\tilde{X}$ independently from F_X *component-wise*.

▷ If the components are correlated, use SCIP algorithm:



Algorithm 1 Sequential Conditional Independent Pairs.

```
$j = 1$   while $j \leq p$ do
    Sample $\tilde{X}_j$ from $\mathcal{L}(X_j \,|\, X_{\text{-}j}, \tilde{X}_{1:j-1})$
    $j = j + 1$
end
```

* SCIP can be take costly runtime in practice as some conditional distribution $\mathcal{L}(X_j \,|\, X_{-j}, \tilde{X}_{1:j-1})$ can be complicated and we need to compute a conditional distribution in each iteration.

Exchangeability Results

Exchangeability of $Y \mid (X, \tilde{X})$

For any subset S of the *null* variables,

$$Y \mid (X, \tilde{X})_{\mathrm{swap}(S)} \overset{d}{=} Y \mid (X, \tilde{X}) \tag{5}$$

Exchangeability Results

Exchangeability of $Y \mid (X, \tilde{X})$

For any subset S of the *null* variables,

$$Y \mid (X, \tilde{X})_{\mathrm{swap}(S)} \overset{d}{=} Y \mid (X, \tilde{X}) \tag{5}$$

▷ A quick proof:

- $p_{Y|(X,\tilde{X})_{\mathrm{swap}(S)}}(y|(x,\tilde{x})) = p_{Y|(X,\tilde{X})}(y|(x,\tilde{x})_{\mathrm{swap}(S)}) = p_{Y|X}(y|x')$, where $x_i' = \tilde{x}_i$ if $i \in S$ and $x_i' = x_i$ otherwise. (by conditional indep.)
- $\forall j \in S$, since $Y \perp\!\!\!\perp X_j \mid X_{-j}$,

$$p_{Y|X}(y|x_j', x_{-j}') = p_{Y|X}(y|\tilde{x}_j, x_{-j}') = p_{Y|X}(y|x_j, x_{-j}')$$

This shows that $Y \mid (X, \tilde{X})_{\mathrm{swap}(S)} \overset{d}{=} Y \mid (X, \tilde{X})_{\mathrm{swap}(S \setminus \{j\})}$. Repeat this process for all j gives us the result.

Theoretical Guarantees

▷ This and exchangeability of $(X, \tilde{X})$ implies the exchangeability of the *joint* distribution.

▷ Then any antisymmetric statistics in the form $W_j = w_j(\begin{bmatrix} X & \tilde{X} \end{bmatrix}, y)$ will have i.i.d. signs for the nulls. FDR control follows from this property.

▷ No sufficiency assumption! The choice of Z_j can be more flexible and even produced by ML models.

- LASSO coefficient difference: $Z_j = |\hat{\beta}_j|$, $W_j = |\hat{\beta}_j| - |\hat{\beta}_{j+p}|$
  - * Can use cross-validation to choose regularization parameter λ
  - * Can apply to GLMs
- $Z_j =$ feature importance score in random forests
- Data adaptive: apply random forests and LASSO, choose Z_j to be the feature importance measure corresponding to the model with smaller cross-validation error

Comparison with Other FDR Control Methods

- ▷ Canonical: produce p-values + **Benjamini Hochberg (BHq)** method

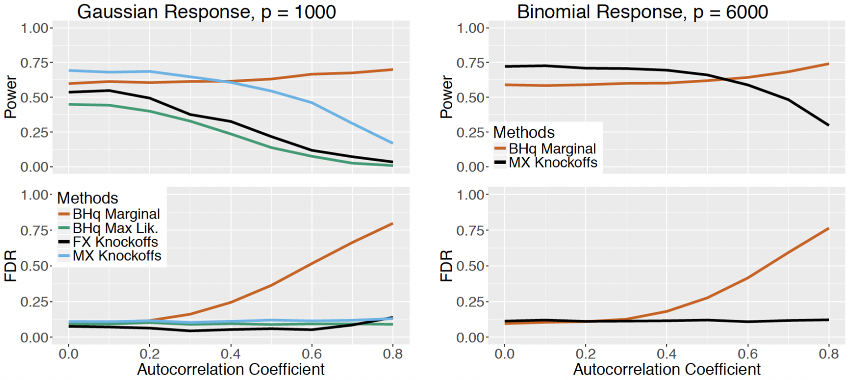


Table of Contents

1 Knockoffs

2 Model-X Knockoffs

3 Knockoffs for HMM

4 Deep Knockoffs

Motivations

- Variable selection problem: Identify which SNPs, out of millions of candidates, influence the risk of a particular disease or a trait
- Genome data particularly fit the assumptions of Model-X knockoff (know F_X and a lot of unsupervised data)
- Apply hidden Markov model (HMM) to deal with *linkage disequilibrium*, i.e. dependence between nearby alleles on a genome.
- HMM has relatively simple close-forms for SCIP algorithm
- Computationally efficient algorithm for constructing knockoffs, $\mathcal{O}(np)$

SCIP for Markov Chains [4]

chain generates the jth knockoff variable $\tilde{X}_j$ *by sampling from*

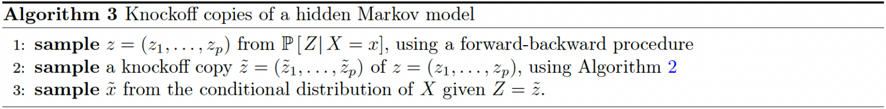
$$\mathbb{P}\left[\tilde{X}_j = \tilde{x}_j \,\middle|\, X_{-j} = x_{-j}, \tilde{X}_{1:(j-1)} = \tilde{x}_{1:(j-1)}\right] = \begin{cases} \dfrac{q_1(\tilde{x}_1)\, Q_2(x_2|\tilde{x}_1)}{\mathcal{N}_1(x_2)}, & j = 1, \\ \dfrac{Q_j(\tilde{x}_j|x_{j-1})\, Q_j(\tilde{x}_j|\tilde{x}_{j-1})\, Q_{j+1}(x_{j+1}|\tilde{x}_j)}{\mathcal{N}_{j-1}(\tilde{x}_j)\, \mathcal{N}_j(x_{j+1})}, & 1 < j < p, \\ \dfrac{Q_p(\tilde{x}_p|x_{p-1})\, Q_p(\tilde{x}_p|\tilde{x}_{p-1})}{\mathcal{N}_{p-1}(\tilde{x}_p)\, \mathcal{N}_p(1)}, & j = p, \end{cases} \tag{4}$$

with the normalization functions $\mathcal{N}_j : \mathcal{X} \mapsto \mathbb{R}_+$ *defined recursively as*

$$\mathcal{N}_j(k) = \begin{cases} \sum_{l \in \mathcal{X}} q_1(l)\, Q_2(k|l), & j = 1, \\ \sum_{l \in \mathcal{X}} \dfrac{Q_j(l|x_{j-1})\, Q_j(l|\tilde{x}_{j-1})\, Q_{j+1}(k|l)}{\mathcal{N}_{j-1}(l)}, & 1 < j < p, \\ \sum_{l \in \mathcal{X}} \dfrac{Q_p(l|x_{p-1})\, Q_p(l|\tilde{x}_{p-1})}{\mathcal{N}_{p-1}(l)}, & j = p. \end{cases} \tag{5}$$

SCIP for HMM [4]

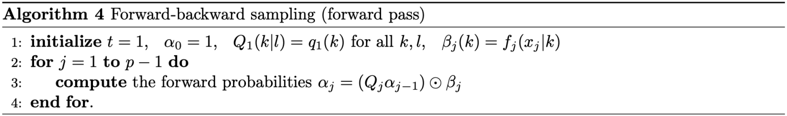
▷ General structure (with param. (π, A, B) specified for genetics context)



Algorithm 3 Knockoff copies of a hidden Markov model

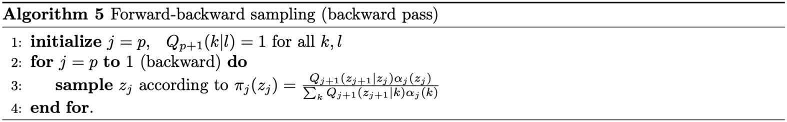
```
sample $z = (z_1, \dots, z_p)$ from $\mathbb{P}[Z | X = x]$, using a forward-backward procedure
sample a knockoff copy $\tilde{z} = (\tilde{z}_1, \dots, \tilde{z}_p)$ of $z = (z_1, \dots, z_p)$, using Algorithm 2
sample $\tilde{x}$ from the conditional distribution of $X$ given $Z = \tilde{z}$.
```

▷ Sample a path of hidden states: "soft" version of Viterbi algorithm Instead of taking the arg max for backward pass in Viterbi algorithm, we sample from a probability distribution to get the previous state.



Algorithm 4 Forward-backward sampling (forward pass)

```
initialize $t = 1$,   $\alpha_0 = 1$,   $Q_1(k|l) = q_1(k)$ for all $k, l$,   $\beta_j(k) = f_j(x_j|k)$
for $j = 1$ to $p - 1$ do
    compute the forward probabilities $\alpha_j = (Q_j \alpha_{j-1}) \odot \beta_j$
end for.
```



Algorithm 5 Forward-backward sampling (backward pass)

```
initialize $j = p$,   $Q_{p+1}(k|l) = 1$ for all $k, l$
for $j = p$ to 1 (backward) do
    sample $z_j$ according to $\pi_j(z_j) = \frac{Q_{j+1}(z_{j+1}|z_j)\alpha_j(z_j)}{\sum_k Q_{j+1}(z_{j+1}|k)\alpha_j(k)}$
end for.
```

Table of Contents

1 Knockoffs

2 Model-X Knockoffs

3 Knockoffs for HMM

4 Deep Knockoffs

Approximate Knockoff Constructions

$\triangleright$ We want $(X, \tilde{X})$ and $(X, \tilde{X})_{\mathsf{swap}(S)}$ to have the same *first two moments*.

This is equivalent to

$$\mathbb{E}\left[X\right] = \mathbb{E}\left[\tilde{X}\right], \ \mathrm{Cov}(X, \tilde{X}) = \begin{bmatrix} \Sigma & \Sigma - \mathrm{diag}\{s\} \\ \Sigma - \mathrm{diag}\{s\} & \Sigma \end{bmatrix}$$

* If $X \sim \mathcal{N}(0, \Sigma)$, then matching the first two moments is equivalent to matching the distributions.

Approximate Knockoff Constructions

▷ We want $(X, \tilde{X})$ and $(X, \tilde{X})_{\mathsf{swap}(S)}$ to have the same *first two moments*.

This is equivalent to

$$\mathbb{E}[X] = \mathbb{E}\left[\tilde{X}\right], \ \mathrm{Cov}(X, \tilde{X}) = \begin{bmatrix} \Sigma & \Sigma - \mathrm{diag}\{s\} \\ \Sigma - \mathrm{diag}\{s\} & \Sigma \end{bmatrix}$$

* If $X \sim \mathcal{N}(0, \Sigma)$, then matching the first two moments is equivalent to matching the distributions.

▷ Moment matching ⇒ "worst" function matching

Define the **maximum mean discrepancy** (MMD) between two distributions as

$$\mathcal{D}_{\mathrm{MMD}}(P_X, P_Z) := \sup_{\|f\|_{\mathcal{H}_K}=1} \left|\mathbb{E}_{X\sim P_X}[f(X)] - \mathbb{E}_{Z\sim P_Z}[f(Z)]\right|$$

MMD Estimation (General)

By Cauchy-Schwarz and reproducing property of kernel,

$$\mathcal{D}_{\mathrm{MMD}}(P_X, P_Z)^2 = \mathbb{E}_{X,X' \overset{i.i.d}{\sim} P_X}[k(X, X')] - 2\mathbb{E}_{X \sim P_X, Z \sim P_Z}[k(X, Z)] + \mathbb{E}_{Z,Z' \overset{i.i.d}{\sim} P_Z}[k(Z, Z')]$$

An unbiased estimate of this quantity is

$$\hat{\mathcal{D}}_{\mathrm{MMD}}(\mathbf{X}, \mathbf{Z})^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} \left[k(X^i, X^j) + k(Z^i, Z^j) \right] - \frac{2}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(X^i, Z^j)$$

where $\mathbf{X}, \mathbf{Z} \in \mathbb{R}^{n \times p}$ are design matrices.

Deep Knockoff Machine [3]

▷ Generate $\tilde{X}$ using deep neural networks with X and random noise V

$$\tilde{X} := f_\theta(X, V)$$

▷ Ensure $\hat{\mathcal{D}}_{\text{MMD}}((X, \tilde{X}), (X, \tilde{X})_{\text{swap}(j)})^2$ for every j is small

We randomly partition X into $X', X'' \in \mathbb{R}^{\frac{n}{2} \times p}$ to obtain *unbiased* estimate. Define the loss function as

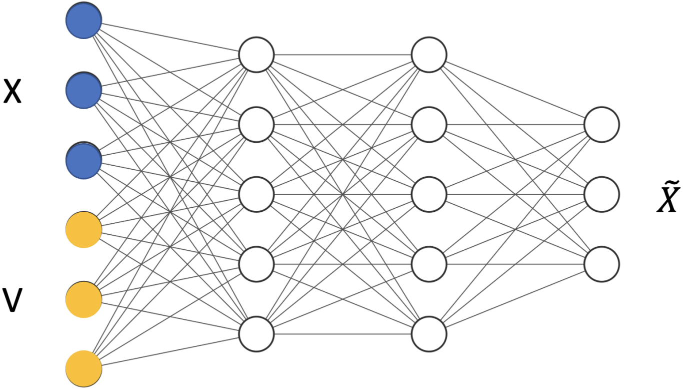
$$J_{\text{MMD}}(X, \tilde{X}) = \hat{\mathcal{D}}_{\text{MMD}}((X', \tilde{X}'), (\tilde{X}'', X''))^2 + \hat{\mathcal{D}}_{\text{MMD}}((X', \tilde{X}'), (X'', \tilde{X}'')_{\text{swap}(S)})^2$$

where S is a uniformly random subset of $[p]$ where each index have $1/2$ probability of being in S.

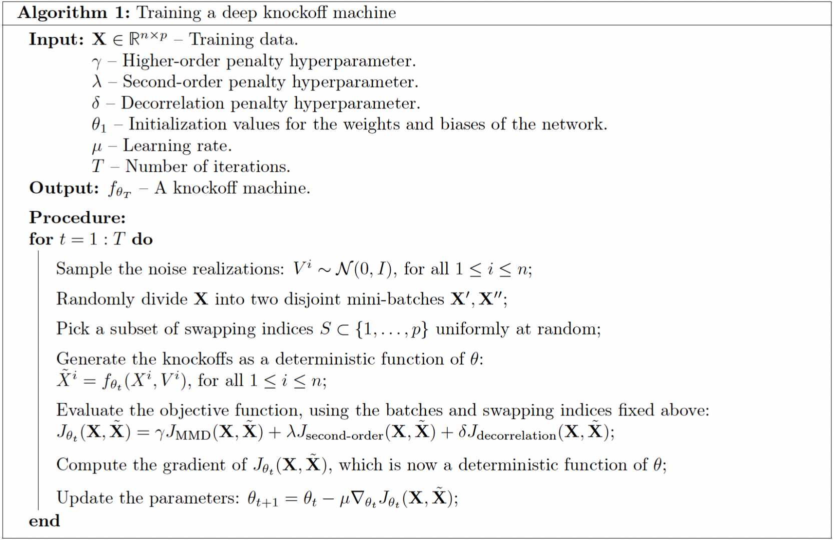
▷ Minimize this loss function using stochastic gradient descent

▷ When the loss is small enough, we can use the DNN to produce knockoffs and select variables using Model-X knockoff procedures

Deep Knockoff Machine: NN Architecture



Deep Knockoff Machine: Algorithm [3]



Algorithm 1: Training a deep knockoff machine

```
Input: $\mathbf{X} \in \mathbb{R}^{n \times p}$ – Training data.
       $\gamma$ – Higher-order penalty hyperparameter.
       $\lambda$ – Second-order penalty hyperparameter.
       $\delta$ – Decorrelation penalty hyperparameter.
       $\theta_1$ – Initialization values for the weights and biases of the network.
       $\mu$ – Learning rate.
       $T$ – Number of iterations.
Output: $f_{\theta_T}$ – A knockoff machine.

Procedure:
for $t = 1 : T$ do
    Sample the noise realizations: $V^i \sim \mathcal{N}(0, I)$, for all $1 \leq i \leq n$;
    Randomly divide $\mathbf{X}$ into two disjoint mini-batches $\mathbf{X}', \mathbf{X}''$;
    Pick a subset of swapping indices $S \subset \{1, \ldots, p\}$ uniformly at random;
    Generate the knockoffs as a deterministic function of $\theta$:
    $\tilde{X}^i = f_{\theta_t}(X^i, V^i)$, for all $1 \leq i \leq n$;
    Evaluate the objective function, using the batches and swapping indices fixed above:
    $J_{\theta_t}(\mathbf{X}, \tilde{\mathbf{X}}) = \gamma J_{\text{MMD}}(\mathbf{X}, \tilde{\mathbf{X}}) + \lambda J_{\text{second-order}}(\mathbf{X}, \tilde{\mathbf{X}}) + \delta J_{\text{decorrelation}}(\mathbf{X}, \tilde{\mathbf{X}})$;
    Compute the gradient of $J_{\theta_t}(\mathbf{X}, \tilde{\mathbf{X}})$, which is now a deterministic function of $\theta$;
    Update the parameters: $\theta_{t+1} = \theta_t - \mu \nabla_{\theta_t} J_{\theta_t}(\mathbf{X}, \tilde{\mathbf{X}})$;
end
```

References

[1] Rina Foygel Barber and Emmanuel J Candès. “Controlling the false discovery rate via knockoffs”. In: *The Annals of Statistics* 43.5 (2015), pp. 2055–2085.

[2] Emmanuel Candes et al. “Panning for gold:‘model-X’knockoffs for high dimensional controlled variable selection”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 80.3 (2018), pp. 551–577.

[3] Yaniv Romano, Matteo Sesia, and Emmanuel Candès. “Deep knockoffs”. In: *Journal of the American Statistical Association* 115.532 (2020), pp. 1861–1872.

[4] Matteo Sesia, Chiara Sabatti, and Emmanuel J Candès. “Gene hunting with hidden Markov model knockoffs”. In: *Biometrika* 106.1 (2019), pp. 1–18.

Recommended resources: Prof. Sesia’s website, Prof. Janson’s talk, Prof. Candès’ talk, Philp Anderson’s video on FDR introduction