Documentation implementation

Data simulations:

$X \in \mathbb{R}^{n \times p}$ where the sample size $n \in \{100, 500\}$ and feature size $p \in \left\{500, 5.10^3\right\}$.
$X \sim \mathcal{N}(0_n, \Sigma)$ where 0_n is a zero vector of size n and $\Sigma_{ij} = \frac{1}{2^{|i-j|}}$

Building Y:

- Model 2.a: $Y = 5X_1 + 2\sin(\pi X_2/2) + 2X_3\mathbf{1}\{X_3 > 0\} + 2\exp\{5X_4\} + \varepsilon$
- Model 2.b: $Y = 3X_1 + 3X_2^3 + 3X_3^{-1} + 5\mathbf{1}\{X_4 > 0\} + \varepsilon$
- Model 2.c: $Y = 1 - 5(X_2 + X_3)^3 \exp\left\{-5\left(X_1 + X_4^2\right)\right\} + \varepsilon$
- Model 2.d: $Y = 1 - 5(X_2 + X_3)^{-3} \exp\{1 + 10\sin(\pi X_1/2) + 5X_4\} + \varepsilon$
  With $\varepsilon \sim \mathcal{N}(0, 1)$

Knock off algorithm:

- if $n \in [1, 2, 3]$ do nothing.
- if $p < n/2$ no need for the screening step, we can construct the exact knock-off directly.
  In the other situations:
- if d is incorrectly set, $d = n_2/2 - 1$

Algorithm:

Input:

- $(X, y) \in \mathbb{R}^{n \times p} \times \mathbb{R}^{n \times q}$
- $\alpha \in [0, 1]$
- $p_1 \in [0, 1]$, relative percentage of the data set to be in fold 1. (algorithm parameter instead of n_1 directly)
- $n_1 = \text{int}(n \times p_1)$ (p_1 is given instead of n_1)
- d such that $d < n_2/2$
- An associative measure $\mathcal{T}$, that can be PC^2 (taken from authors code), $HSIC$ or MMD.

Checks before starting:

- if n is large enough no need for splitting or screening, jump to knockoff step.
- Else, split data randomly into two according to n_1 and $n_2 = n - n_1$, use $\left(X^{(1)}, y^{(1)}\right)$ in the screening step and $\left(X^{(2)}, y^{(2)}\right)$ in the knockoff step.

Screening step

1. $\forall j \in [1:p], \hat{\omega_j}^{(1)} = \mathcal{T}(X_j^{(1)}, y^{(1)})$
2. Select top d features, $\widehat{\mathcal{A}}_1 = \left\{j : \widehat{\omega}_j^{(1)} \text{is among the largest } d\right\}$

Knockoff step

1. Keep $\widehat{\mathcal{A}}_1$ features from $X^{(2)}$, named $X^{(2)}_{\widehat{\mathcal{A}}_1}$, build exact knock off with *equicorrelated construction*, code taken from authors code.
2. $\forall j \in \widehat{\mathcal{A}}_1, \widehat{W_j} = \mathcal{T}(X^{(2)}_{\widehat{\mathcal{A}}_1, j}, y^{(2)}) - \mathcal{T}(X^{(2)}_{\widehat{\mathcal{A}}_1, j}, y^{(2)})$
3. $T_\alpha = \min\left\{t \in \mathcal{W} : \frac{1+\#\left\{j:\widehat{W}_j \leq -t\right\}}{\#\left\{j:\widehat{W}_j \geq t\right\}} \leq \alpha\right\}$ where $\mathcal{W} = \left\{\left|\widehat{W}_j\right| : 1 \leq j \leq p\right\} / \{0\}$
4. $\widehat{\mathcal{A}}\left(T_\alpha\right) = \left\{j : \widehat{W}_j \geq T_\alpha, 1 \leq j \leq p\right\}$
5. If $\widehat{\mathcal{A}}\left(T_\alpha\right)$ is empty we return the empty set $\widehat{\mathcal{A}}_1$ ~~or the full set of features.~~

Questions & remarks

1. In the authors code, we notice that the PC function allows to take in input X a matrix and not a feature column as we expected, is there a situation where we would feed multiple feature columns to the association measures $\mathcal{T}$?
2. In 5. of the knockoff step, $\widehat{\mathcal{A}}\left(T_\alpha\right)$ could be empty, if all knockoff variables are better then the originals. Should we return an empty set or the full set?
3. Should we use a "recall" metric? To compare methods and checks if the good features are correctly selected? In the simulation case, this would be to count how many times the first 4 features are selected.
4. How should we control the FDR? After using the procedure to select the appropriate features should we use a classifier, such as a random forest to compute the FDR and estimate it? How does this work in the regression setting?
   --- > $\text{FDR} = \mathbb{E}\left[\frac{\#\left\{j:\beta_j=0 \text{ and } j\in\hat{S}\right\}}{\#\{j:j\in\hat{S}\}\vee 1}\right]$
5. Is it ok to remove the screening step if n is large enough? Even tho we might not be dealing with this case, it is a possible situation, and scikit learn test's this situation.

TODOs:

- [ ] Implement associative measure *HSIC* and *MMD*.
- [ ] Implement the follow up for controlling the FDR rate
- [ ] Find a correct database from UCI