Discovering Conditionally Salient Features with Statistical Guarantees

Jaime Roquero Gimenez, James Zou

Stanford University

Feature Selection

Setting the problem:

- Dataset with d features $X_1, \ldots, X_d$
- Response variable Y
- **Goal**: Find set of important variables $\mathcal{H}_1 \subset \{1, \ldots, d\}$

A variable $j \in \mathcal{H}_0$ is null (i.e. irrelevant for predicting Y) if

$$X_j \perp\!\!\!\perp Y | \boldsymbol{X}_{-j}$$

Otherwise, we say that that $j \in \mathcal{H}_1$ is non-null.

- Construct a procedure that outputs an estimate $\hat{S}$ of $\mathcal{H}_1$
- False Discovery Rate control as statistical guarantee:

$$\text{FDR} = \mathbb{E}\Big[\frac{|\hat{S} \cap \mathcal{H}_0|}{|\hat{S}| \vee 1}\Big]$$

Feature Selection

Setting the problem:

- Dataset with d features $X_1, \dots, X_d$
- Response variable Y
- **Goal**: Find set of important variables $\mathcal{H}_1 \subset \{1, \dots, d\}$

A variable $j \in \mathcal{H}_0$ is null (i.e. irrelevant for predicting Y) if

$$X_j \perp\!\!\!\perp Y | \boldsymbol{X}_{-j}$$

Otherwise, we say that that $j \in \mathcal{H}_1$ is non-null.

- Construct a procedure that outputs an estimate $\hat{S}$ of $\mathcal{H}_1$
- False Discovery Rate control as statistical guarantee:

$$\mathrm{FDR} = \mathbb{E}\Big[\frac{|\hat{S} \cap \mathcal{H}_0|}{|\hat{S}| \vee 1}\Big]$$

Feature Selection

Setting the problem:

- Dataset with d features $X_1, \dots, X_d$
- Response variable Y
- **Goal**: Find set of important variables $\mathcal{H}_1 \subset \{1, \dots, d\}$

A variable $j \in \mathcal{H}_0$ is null (i.e. irrelevant for predicting Y) if

$$X_j \perp\!\!\!\perp Y | \boldsymbol{X}_{-j}$$

Otherwise, we say that that $j \in \mathcal{H}_1$ is non-null.

- Construct a procedure that outputs an estimate $\hat{S}$ of $\mathcal{H}_1$
- False Discovery Rate control as statistical guarantee:

$$\mathrm{FDR} = \mathbb{E}\Big[\frac{|\hat{S} \cap \mathcal{H}_0|}{|\hat{S}| \vee 1}\Big]$$

Feature Selection in Linear Model

Fit a linear model to the data:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \cdots + \beta_d X_d + \epsilon$$

Which variables are important? Those whose corresponding coefficients are non-zero.

$$\beta_1, \beta_3 \neq 0 \Rightarrow 1, 3 \in \mathcal{H}_1$$
$$\beta_2 = \beta_4 = \cdots = \beta_d = 0 \Rightarrow 2, 4, \ldots, d \in \mathcal{H}_0$$

In this model, non-null features are **global non-nulls**. We have $\mathcal{H}_1 = \{1, 3\}$, **regardless of the value of X**

Feature Selection in Linear Model

Fit a linear model to the data:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \cdots + \beta_d X_d + \epsilon$$

Which variables are important? Those whose corresponding coefficients are non-zero.

$$\beta_1, \beta_3 \neq 0 \Rightarrow 1, 3 \in \mathcal{H}_1$$
$$\beta_2 = \beta_4 = \cdots = \beta_d = 0 \Rightarrow 2, 4, \ldots, d \in \mathcal{H}_0$$

In this model, non-null features are **global non-nulls**. We have $\mathcal{H}_1 = \{1, 3\}$, **regardless of the value of X**

Feature Selection in Linear Model

Fit a linear model to the data:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \cdots + \beta_d X_d + \epsilon$$

Which variables are important? Those whose corresponding coefficients are non-zero.

$$\beta_1, \beta_3 \neq 0 \Rightarrow 1, 3 \in \mathcal{H}_1$$
$$\beta_2 = \beta_4 = \cdots = \beta_d = 0 \Rightarrow 2, 4, \ldots, d \in \mathcal{H}_0$$

In this model, non-null features are **global non-nulls**. We have $\mathcal{H}_1 = \{1, 3\}$, **regardless of the value of X**

Global vs. Local non-nulls

What if a feature is non-null **depending on the value of other features**?

$$\begin{cases} Y = X_2 + \epsilon & \text{if } X_1 > c \\ Y = X_3 + \epsilon & \text{if } X_1 \leq c \end{cases}$$

$$\text{"} \Rightarrow \text{"} \begin{cases} \mathcal{H}_1 = \{1,2\} & \text{if } X_1 > c \\ \mathcal{H}_1 = \{1,3\} & \text{if } X_1 \leq c \end{cases}$$

From a *global* perspective, $\mathcal{H}_1 = \{1, 2, 3\}$.
Can we generate a procedure that selects non-null features **locally**, while retaining statistical guarantees? Potentially yes if model *interactions* in parametric models of $Y|\boldsymbol{X}$. What if such models are not available?

Global vs. Local non-nulls

What if a feature is non-null **depending on the value of other features**?

$$\begin{cases} Y = X_2 + \epsilon & \text{if } X_1 > c \\ Y = X_3 + \epsilon & \text{if } X_1 \leq c \end{cases}$$

$$\text{"} \Rightarrow \text{"} \begin{cases} \mathcal{H}_1 = \{1, 2\} & \text{if } X_1 > c \\ \mathcal{H}_1 = \{1, 3\} & \text{if } X_1 \leq c \end{cases}$$

From a *global* perspective, $\mathcal{H}_1 = \{1, 2, 3\}$.

Can we generate a procedure that selects non-null features **locally**, while retaining statistical guarantees? Potentially yes if model *interactions* in parametric models of $Y|\boldsymbol{X}$. What if such models are not available?

Global vs. Local non-nulls

What if a feature is non-null **depending on the value of other features**?

$$\begin{cases} Y = X_2 + \epsilon & \text{if } X_1 > c \\ Y = X_3 + \epsilon & \text{if } X_1 \leq c \end{cases}$$

$$\text{“} \Rightarrow \text{”} \begin{cases} \mathcal{H}_1 = \{1, 2\} & \text{if } X_1 > c \\ \mathcal{H}_1 = \{1, 3\} & \text{if } X_1 \leq c \end{cases}$$

From a *global* perspective, $\mathcal{H}_1 = \{1, 2, 3\}$.
Can we generate a procedure that selects non-null features **locally**, while retaining statistical guarantees? Potentially yes if model *interactions* in parametric models of $Y|\boldsymbol{X}$. What if such models are not available?

Local Definition of Null Variable

A variable $j \in \mathcal{H}_0$ is null if

$$X_j \perp\!\!\!\perp Y | \boldsymbol{X}_{-j}$$

We define / construct:

- the sets of *local nulls* $\mathcal{H}_0(\boldsymbol{x})$, *local non-nulls* $\mathcal{H}_1(\boldsymbol{x})$ at points in feature space
- a procedure to return a *local estimate* $\hat{S}(\boldsymbol{x})$ of the local non-nulls
- a generalization of FDR to a *local FDR*

How to retain *FDR control* in a *local* setting, *without using a parametric model* for $Y|\boldsymbol{X}$?

Local Definition of Null Variable

A variable $j \in \mathcal{H}_0(\boldsymbol{x})$ is a local null at $\boldsymbol{X} = \boldsymbol{x}$ if

$$X_j \perp\!\!\!\perp Y | \boldsymbol{X}_{-j} = \boldsymbol{x}_{-j}$$

We define / construct:

- the sets of *local nulls* $\mathcal{H}_0(\boldsymbol{x})$, *local non-nulls* $\mathcal{H}_1(\boldsymbol{x})$ at points in feature space
- a procedure to return a *local estimate* $\hat{S}(\boldsymbol{x})$ of the local non-nulls
- a generalization of FDR to a *local FDR*

How to retain *FDR control* in a *local* setting, *without using a parametric model* for $Y|\boldsymbol{X}$?

Local Definition of Null Variable

A variable $j \in \mathcal{H}_0(\boldsymbol{x})$ is a local null at $\boldsymbol{X} = \boldsymbol{x}$ if

$$X_j \perp\!\!\!\perp Y | \boldsymbol{X}_{-j} = \boldsymbol{x}_{-j}$$

We define / construct:

- the sets of *local nulls* $\mathcal{H}_0(\boldsymbol{x})$, *local non-nulls* $\mathcal{H}_1(\boldsymbol{x})$ at points in feature space
- a procedure to return a *local estimate* $\hat{S}(\boldsymbol{x})$ of the local non-nulls
- a generalization of FDR to a *local FDR*

How to retain *FDR control* in a *local* setting, *without using a parametric model* for $Y|\boldsymbol{X}$?

Knockoff Procedure

Most feature selection procedures construct scores T_j for each feature:

$$\begin{gathered} X_1, X_2, \ldots, X_d, \quad Y \\ \downarrow \\ T_1, T_2, \ldots, T_d \end{gathered}$$

Then scores are ranked and some cutoff leads to $\hat{S}$.

- *Need a statistical model* to have statistical guarantees on FDR
- If high-dimensional setting, statistical assumptions may fail.
- If wanted to do local feature selection, subsetting data could limit the power and break assumptions based on asymptotic behavior.

These limitations make local feature selection a hard problem for usual methods.

Knockoff Procedure

Most feature selection procedures construct scores T_j for each feature:

$$X_1, X_2, \ldots, X_d, \quad Y$$
$$\downarrow$$
$$T_1, T_2, \ldots, T_d$$

Then scores are ranked and some cutoff leads to $\hat{S}$.

- *Need a statistical model* to have statistical guarantees on FDR
- If high-dimensional setting, statistical assumptions may fail.
- If wanted to do local feature selection, subsetting data could limit the power and break assumptions based on asymptotic behavior.

These limitations make local feature selection a hard problem for usual methods.

Knockoff Procedure

The knockoff procedure generates a new, synthetic dataset $\tilde{\boldsymbol{X}}$, and constructs scores as previously:

$$X_1, X_2, \ldots, X_d, \tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_d, \quad Y$$
$$\downarrow \qquad \downarrow$$
$$T_1, T_2, \ldots, T_d, \ \tilde{T}_1, \tilde{T}_2, \ldots, \tilde{T}_d$$

Ranking the differences $W_j = T_j - \tilde{T}_j$ allows to select features with FDR control.

Does not require modeling $Y|\boldsymbol{X}$ for FDR control. Statistical guarantees *only depend on the validity of the process to generate* $\tilde{\boldsymbol{X}}$.

Knockoff Procedure

The knockoff procedure generates a new, synthetic dataset $\tilde{\boldsymbol{X}}$, and constructs scores as previously:

$$\begin{array}{ccc} X_1, X_2, \ldots, X_d, & \tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_d, & Y \\ \downarrow & \downarrow & \\ T_1, T_2, \ldots, T_d, & \tilde{T}_1, \tilde{T}_2, \ldots, \tilde{T}_d & \end{array}$$

Ranking the differences $W_j = T_j - \tilde{T}_j$ allows to select features with FDR control.
Does not require modeling $Y|\boldsymbol{X}$ for FDR control. Statistical guarantees *only depend on the validity of the process to generate* $\tilde{\boldsymbol{X}}$.

Localize the Knockoff Procedure

Our work generalizes the Knockoff procedure to tackle local feature selection:

- Generalize the distributional properties of the knockoff variables $\tilde{\boldsymbol{X}}$ to the local setting, without additional constraints.
- Generalize the construction of the scores to capture local dependence.

By generating $\tilde{X}$ as in the usual knockoff procedure, using the whole dataset, the statistical guarantees hold for the localized procedure.

Localize the Knockoff Procedure

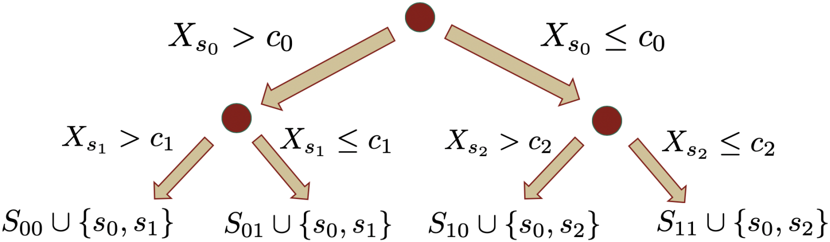
Our work generalizes the Knockoff procedure to tackle local feature selection:

- Generalize the distributional properties of the knockoff variables $\tilde{\boldsymbol{X}}$ to the local setting, without additional constraints.
- Generalize the construction of the scores to capture local dependence.

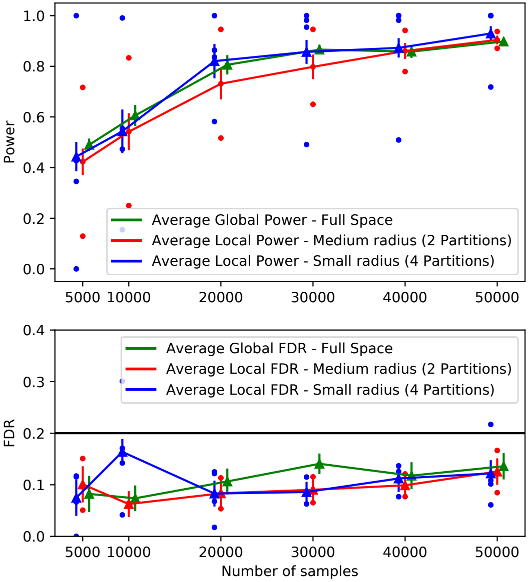
By generating $\tilde{\boldsymbol{X}}$ as in the usual knockoff procedure, using the whole dataset, the statistical guarantees hold for the localized procedure.

Example: Switch variable model

Three switch features $X_{s_0}, X_{s_1}, X_{s_2}$ and four different sets of local non-nulls $S_{00}, S_{01}, S_{10}, S_{11}$. Y has a linear response in $X_{S_{ij}}$.



Local FDR control



Thank you