Controlled Variable Selection via Mirror Statistics

Xin Xing

Model Setting for Sparse Linear Regression

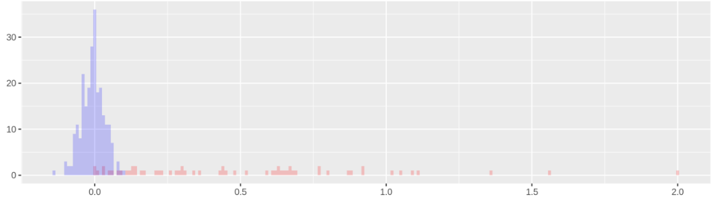
- **Data:** Observations $(y_i, \mathbf{x}_i)_{i=1}^n$ where $y_i \in \mathbb{R}$, $\mathbf{x}_i \in \mathbb{R}^p$
- **Linear Model:** $y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i$
  - $\boldsymbol{\beta} \in \mathbb{R}^p$: Unknown coefficient vector
  - $\varepsilon_i \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$: Noise terms
- **Sparsity Assumption:** Most components of $\boldsymbol{\beta}$ are zero
  - $\|\boldsymbol{\beta}\|_0 = s_1 \ll p$ (only s non-zero coefficients)
  - Active set: $\mathcal{S}_1 = \{j : \beta_j \neq 0\}$ with $|\mathcal{S}_1| = s_1$
- **High-dimensional Setting:** Number of predictors p can be much larger than sample size n

Variable Selection: Null and Non-null Sets

- **Variable Index Set:** $\{1, 2, \ldots, p\}$ for p predictors
- **Null Variable Set:** $\mathcal{S}_0 = \{j : \beta_j = 0\}$
  - Variables with zero coefficients
  - No influence on the response variable
  - To be excluded from the final model
- **Non-null Variable Set:** $\mathcal{S}_1 = \{j : \beta_j \neq 0\}$
  - Variables with non-zero coefficients
  - Contribute to explaining the response variable
- **Partition Property:** $\mathcal{S}_0 \cup \mathcal{S}_1 = \{1, 2, \ldots, p\}$ and $\mathcal{S}_0 \cap \mathcal{S}_1 = \emptyset$

Mirror statistics

- Blue: Mirror statistics for null variables
- Red: Mirror statistics for nonnull variables



Definition: Mirror Statistic

Let Y be a target variable and X_j be the j-th predictor variable. We call M_j a mirror statistic corresponding to the j-th predictor if it satisfies:

- **P1:** When $j \in \mathcal{S}_0$, M_j follows a distribution that is symmetric around zero.
- **P2:** The magnitude of M_j increases with the magnitude of $|\beta_j|$, such that large positive values of M_j indicate strong statistical association between the j-th predictor and the target variable.

Feature Selection with FDP Estimation

- **Selection Rule:** Select variables $\hat{\mathcal{S}} = \{j : M_j > t\}$ for threshold $t > 0$
- **Estimating False Positives:** Leverage symmetry of null mirror statistics
  - Number of selections: $|\hat{\mathcal{S}}| = \#\{j : M_j > t\}$
  - Estimated false positives: $\widehat{FP}(t) = \#\{j : M_j < -t\}$
  - Based on the property: for $j \in \mathcal{S}_0$, $\mathbb{P}(M_j > t) = \mathbb{P}(M_j < -t)$
- **False Discovery Proportion Estimate:**

$$\widehat{FDP}(t) = \frac{\#\{j : M_j < -t\}}{\#\{j : M_j > t\} \vee 1} \tag{1}$$

- **Threshold Selection:** For target FDR level q

$$\hat{t} = \min\{t > 0 : \widehat{FDP}(t) \leq q\} \tag{2}$$

How to Construct Mirror Statistics

- Knockoffs
- Data-splitting

Creating Exchangeable "Fake" Covariates

- **Key Idea:** Generate a set of "fake" covariates $\tilde{\mathbf{X}}$ that:
  - Preserve the covariance structure of original covariates $\mathbf{X}$
  - Are "exchangeable" with the original covariates
- **Exchangeability Conditions:**

$$\tilde{\mathbf{X}}_j^T \tilde{\mathbf{X}}_k = \mathbf{X}_j^T \mathbf{X}_k, \quad \forall j, k \tag{3}$$

$$\tilde{\mathbf{X}}_j^T \mathbf{X}_k = \mathbf{X}_j^T \mathbf{X}_k, \quad \forall j \neq k \tag{4}$$

- **Implication:** This pairwise exchangeability leads to the estimation of false discoveries (FDs)

Knockoffs

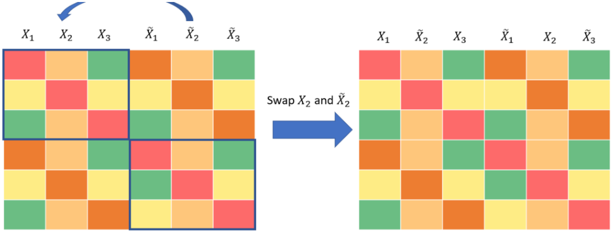
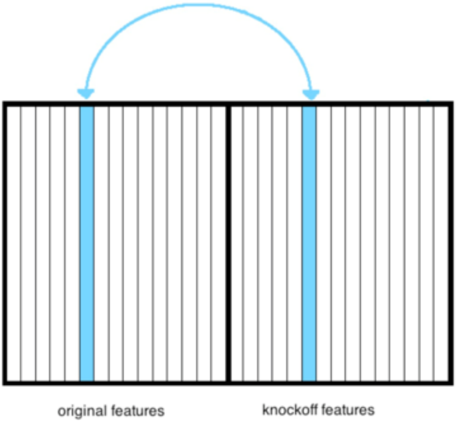


Figure: Swap for null variables

Knockoffs

$$[X\ \tilde{X}]^\top [X\ \tilde{X}] = \begin{pmatrix} \Sigma & \Sigma - diag(s) \\ \Sigma - diag(s) & \Sigma \end{pmatrix} \tag{5}$$



Joint Distribution of Original and Knockoff Variables

- **Original Variables:** $\mathbf{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
- **Gram Matrix of Augmented Design:**

$$[\mathbf{X}\ \tilde{\mathbf{X}}]^\top [\mathbf{X}\ \tilde{\mathbf{X}}] = \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \mathrm{diag}(\mathbf{s}) \\ \boldsymbol{\Sigma} - \mathrm{diag}(\mathbf{s}) & \boldsymbol{\Sigma} \end{pmatrix} \tag{6}$$

- **Joint Distribution:**

$$\begin{pmatrix} \mathbf{X} \\ \tilde{\mathbf{X}} \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{\Sigma} - \mathrm{diag}(\mathbf{s}) \\ \boldsymbol{\Sigma} - \mathrm{diag}(\mathbf{s}) & \boldsymbol{\Sigma} \end{pmatrix} \right) \tag{7}$$

Conditional Distribution Formula for Multivariate Normal

For a partitioned multivariate normal distribution:

$$\begin{pmatrix} \mathbf{Y}_1 \\ \mathbf{Y}_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \right) \tag{8}$$

The conditional distribution follows:

$$\mathbf{Y}_1 | \mathbf{Y}_2 = \mathbf{y}_2 \sim N(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{y}_2 - \boldsymbol{\mu}_2), \tag{9}$$

$$\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}) \tag{10}$$

Applying the Formula to Knockoffs

For our case, we identify:

- $\mathbf{Y}_1 = \tilde{\mathbf{X}}$ and $\mathbf{Y}_2 = \mathbf{X}$
- $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \boldsymbol{\mu}$
- $\boldsymbol{\Sigma}_{11} = \boldsymbol{\Sigma}_{22} = \boldsymbol{\Sigma}$
- $\boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}_{21} = \boldsymbol{\Sigma} - \mathrm{diag}(\mathbf{s})$

Therefore:

$$\tilde{\mathbf{X}}|\mathbf{X} = \mathbf{x} \sim N(\boldsymbol{\mu} + (\boldsymbol{\Sigma} - \mathrm{diag}(\mathbf{s}))\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}), \tag{11}$$

$$\boldsymbol{\Sigma} - (\boldsymbol{\Sigma} - \mathrm{diag}(\mathbf{s}))\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \mathrm{diag}(\mathbf{s}))) \tag{12}$$

Simplifying the Conditional Distribution

Simplifying the Mean:

$$\boldsymbol{\mu} + (\boldsymbol{\Sigma} - \text{diag}(\mathbf{s}))\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = \boldsymbol{\mu} + (\mathbf{I} - \boldsymbol{\Sigma}^{-1}\text{diag}(\mathbf{s}))(\mathbf{x} - \boldsymbol{\mu}) \tag{13}$$

$$= \boldsymbol{\mu} + \mathbf{x} - \boldsymbol{\mu} - \boldsymbol{\Sigma}^{-1}\text{diag}(\mathbf{s})(\mathbf{x} - \boldsymbol{\mu}) \tag{14}$$

$$= \mathbf{x} - \boldsymbol{\Sigma}^{-1}\text{diag}(\mathbf{s})(\mathbf{x} - \boldsymbol{\mu}) \tag{15}$$

For Zero-Mean Case ($\boldsymbol{\mu} = \mathbf{0}$):

$$\mathbb{E}[\tilde{\mathbf{X}}|\mathbf{X} = \mathbf{x}] = \mathbf{x} - \boldsymbol{\Sigma}^{-1}\text{diag}(\mathbf{s})\mathbf{x} \tag{16}$$

$$= \mathbf{x}(\mathbf{I} - \boldsymbol{\Sigma}^{-1}\text{diag}(\mathbf{s})) \tag{17}$$

Simplifying the Conditional Covariance

$$\begin{aligned}
\mathrm{Var}[\tilde{\mathbf{X}}|\mathbf{X}] &= \boldsymbol{\Sigma} - (\boldsymbol{\Sigma} - \mathrm{diag}(\mathbf{s}))\boldsymbol{\Sigma}^{-1}(\boldsymbol{\Sigma} - \mathrm{diag}(\mathbf{s})) \\
&= \boldsymbol{\Sigma} - (\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-1}\mathrm{diag}(\mathbf{s}) \\
&\quad - \mathrm{diag}(\mathbf{s})\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma} + \mathrm{diag}(\mathbf{s})\boldsymbol{\Sigma}^{-1}\mathrm{diag}(\mathbf{s})) \\
&= \boldsymbol{\Sigma} - (\boldsymbol{\Sigma} - \mathrm{diag}(\mathbf{s}) - \mathrm{diag}(\mathbf{s}) + \mathrm{diag}(\mathbf{s})\boldsymbol{\Sigma}^{-1}\mathrm{diag}(\mathbf{s})) \\
&= \boldsymbol{\Sigma} - \boldsymbol{\Sigma} + 2\mathrm{diag}(\mathbf{s}) - \mathrm{diag}(\mathbf{s})\boldsymbol{\Sigma}^{-1}\mathrm{diag}(\mathbf{s}) \\
&= 2\mathrm{diag}(\mathbf{s}) - \mathrm{diag}(\mathbf{s})\boldsymbol{\Sigma}^{-1}\mathrm{diag}(\mathbf{s})
\end{aligned}$$

Constructing SDP Model-X Knockoffs

- **Model-X Assumption:** Distribution of predictors $\mathbf{X} \sim F_X$ is known
- **Construction Steps:**
  1. Compute the covariance matrix $\mathbf{\Sigma} = \text{Cov}(\mathbf{X})$
  2. Solve the semidefinite program (SDP):

$$\max_{\mathbf{s}} \quad \sum_{j=1}^{p} s_j \tag{18}$$

$$\text{subject to} \quad 0 \leq s_j \leq 1, \quad j = 1, \ldots, p \tag{19}$$

$$\mathbf{G} = 2\mathbf{\Sigma} - \text{diag}(\mathbf{s}) \succeq 0 \tag{20}$$

  3. Compute the conditional distribution:

$$\tilde{\mathbf{X}} | \mathbf{X} \sim N(\boldsymbol{\mu}, \mathbf{V}) \tag{21}$$

where $\boldsymbol{\mu} = \mathbf{X}(I - \mathbf{\Sigma}^{-1}\text{diag}(\mathbf{s}))$ and $\mathbf{V} = 2\text{diag}(s) - \text{diag}(\mathbf{s})\mathbf{\Sigma}^{-1}\text{diag}(\mathbf{s})$

Knockoff Construction: Augmented Design Matrix

- **Original Design Matrix:** $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p) \in \mathbb{R}^{n \times p}$
- **Knockoff Variables:** $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2, \ldots, \tilde{\mathbf{X}}_p) \in \mathbb{R}^{n \times p}$
- **Augmented Design Matrix:**

$$[\mathbf{X}, \tilde{\mathbf{X}}] = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_p, \tilde{\mathbf{X}}_1, \tilde{\mathbf{X}}_2, \ldots, \tilde{\mathbf{X}}_p) \tag{22}$$
$$\in \mathbb{R}^{n \times 2p} \tag{23}$$

- **Properties:**
  - Preserves correlation structure of original variables
  - Knockoffs are exchangeable with their original counterparts
  - Used for variable selection with controlled false discovery rate

Lasso with Augmented Design Matrix

- **Objective:** Apply Lasso to the augmented design matrix $[\mathbf{X}, \tilde{\mathbf{X}}]$
- **Lasso Optimization Problem:**

$$\beta^{\lambda} = \arg\min_{\beta \in \mathbb{R}^{2p}} \{\frac{1}{2} \left| \mathbf{y} - [\mathbf{X}, \tilde{\mathbf{X}}] \cdot \beta \right|_2^2 + \lambda |\beta|_1\} \tag{24}$$

- **Parameter Interpretation:**
  - $\beta^{\lambda} = (\beta_1^{\lambda}, \ldots, \beta_p^{\lambda}, \tilde{\beta}_1^{\lambda}, \ldots, \tilde{\beta}_p^{\lambda})$
  - Original variables: β_j^{λ} for $j = 1, \ldots, p$
  - Knockoff variables: $\tilde{\beta}_j^{\lambda}$ for $j = 1, \ldots, p$
- **Feature Importance Measure:**
  - Compare original vs. knockoff coefficients
  - Variables with $|\beta_j^{\lambda}| > |\tilde{\beta}_j^{\lambda}|$ are more likely to be truly relevant

Constructing Mirror Statistics

- **Definition:** The mirror statistic M_j for variable j is defined as:

$$M_j = |\beta_j| - |\tilde{\beta}_j| \tag{25}$$

- **Properties:**
  - $M_j > t$ suggests variable j has stronger association with response than its knockoff
  - Under null (irrelevant variable), M_j is symmetric around zero
- **Key:** For null variables, original and knockoff variables are exchangeable
  - $\mathbb{P}(M_j > t) = \mathbb{P}(M_j < -t)$ for all thresholds t when $j \in \mathcal{S}_0$

Symmetry of Mirror Statistics Under the Null Hypothesis

- **Mirror Statistic:** $M_j = |\beta_j| - |\tilde{\beta}_j|$ for variable j
- **Key Property:** Under $H_0 : j \in \mathcal{S}_0$ (variable j is null), M_j is symmetric around zero
- **Proof of Symmetry:**
  - Under the null hypothesis, X_j has no effect on Y given other variables
  - Perfect exchangeability: $(X_j, \tilde{X}_j)$ can be swapped without affecting joint distribution with Y
  - This implies $P(|\beta_j| - |\tilde{\beta}_j| > t) = P(|\tilde{\beta}_j| - |\beta_j| > t)$
  - Equivalently: $P(M_j > t) = P(M_j < -t)$ for all thresholds t
  - Therefore, $P(M_j \in A) = P(M_j \in -A)$ for any set A
  - This is the definition of a distribution symmetric around zero

Feature Selection with FDP Estimation

- **Selection Rule:** Select variables $\hat{\mathcal{S}} = \{j : M_j > t\}$ for threshold $t > 0$
- **Estimating False Positives:** Leverage symmetry of null mirror statistics
  - Number of selections: $|\hat{\mathcal{S}}| = \#\{j : M_j > t\}$
  - Estimated false positives: $\widehat{FP}(t) = \#\{j : M_j < -t\}$
  - Based on the property: for $j \in \mathcal{S}_0$, $\mathbb{P}(M_j > t) = \mathbb{P}(M_j < -t)$
- **False Discovery Proportion Estimate:**

$$\widehat{FDP}(t) = \frac{\#\{j : M_j < -t\}}{\#\{j : M_j > t\} \vee 1} \tag{26}$$

- **Threshold Selection:** For target FDR level q

$$\hat{t} = \min\{t > 0 : \widehat{FDP}(t) \leq q\} \tag{27}$$

Basic Usage

```
# Default settings
library(knockoff)
result = knockoff.filter(X, y)
# Check FDP
fdp = function(selected) {
  sum(beta[selected] == 0) / max(1, length(selected))
}
fdp(result$selected)
```

Default settings:

- Target FDR $= 0.1$
- Second-order Gaussian knockoffs
- Estimates μ and Σ from data

Custom Knockoff Generation

```
# Using true parameters
gaussian_knockoffs = function(X) {
  create.gaussian(X, mu, Sigma)
}
result = knockoff.filter(X, y,
                         knockoffs=gaussian_knockoffs)
# Using random forests
result = knockoff.filter(X, y,
                         knockoffs=gaussian_knockoffs,
                         statistic=stat.random_forest,
                         fdr=0.2)
```

Specifying Knockoff Types

```
# Full SDP
gaussian_knockoffs = function(X) {
  create.second_order(X, method='sdp', shrink=T)
}
```

Data-Splitting Approach for Mirror Statistics

- **Key idea:** Randomly split the n observations into two groups:
  - Group 1: $(\mathbf{y}^{(1)}, \mathbf{X}^{(1)})$
  - Group 2: $(\mathbf{y}^{(2)}, \mathbf{X}^{(2)})$
- **Procedure:**
  1. Fit the model separately on each group to obtain $\hat{\beta}_j^{(1)}$ and $\hat{\beta}_j^{(2)}$
  2. Compute the mirror statistic:

$$M_j = |\hat{\beta}_j^{(1)} + \hat{\beta}_j^{(2)}| - |\hat{\beta}_j^{(1)} - \hat{\beta}_j^{(2)}| \tag{28}$$

- **Symmetry assumption:** If j is a null variable, then the sampling distribution of $\hat{\beta}_j^{(2)}$ is symmetric about 0
- **Implications:**
  - For null variables: M_j is symmetric around zero
  - Large positive M_j: Consistent signal across splits
  - Large negative M_j: Inconsistent signal across splits (likely noise)
  - No need for knockoff construction