Power Analysis of Knockoff Filters for Correlated Designs

Jingbo Liu, Philippe Rigollet

Massachusetts Institute of Technology

December 2019

Feature Selection Problem

Setting: Observation model

$$\mathbf{Y} = \mathbf{X}\theta + \mathbf{N},$$

where $\mathbf{X}$ is $n \times p$ known. Want: find j for which $|\theta_j| > 0$.

False discovery rate: FDR $= \mathbb{E}\left[\frac{|\mathcal{H}_0 \cap \hat{\mathcal{H}}_1|}{|\hat{\mathcal{H}}_1|}\right]$

- $\mathcal{H}_0$ null; $\mathcal{H}_1$ alternative hypothesis; $\hat{\mathcal{H}}_{0/1}$ declared.

Power: PWR $= \mathbb{E}\left[\frac{|\mathcal{H}_1 \cap \hat{\mathcal{H}}_1|}{|\mathcal{H}_1|}\right]$.

The FDR control problem: Given a budget FDR $\leq q$, maximize PWR.

FDR Control Algorithms: A Review

Benjamini & Hochberg 95

- Simply reject null hypothesis when p-value is small.
- Relevant setting: $\mathbf{X}$ is identity and the null distribution is known.

Simple Lasso

- $\hat{\theta} := \mathrm{argmin}_\theta \left\{ \frac{1}{2n} \|\mathbf{Y} - \mathbf{X}\theta\|_2^2 + \lambda\|\theta\|_1 \right\};$ select $\{j : |\hat{\theta}_j| > T\}$.
- Relevant setting: general $\mathbf{X}$, oracle threshold T.

Knockoff filter [Barber & Candès 15], [Candès, Fan, Janson, Lv 18]

- Relevant setting: general $\mathbf{X}$.

Review of the Knockoff Filter

We consider the setting of random design where the rows of $\mathbf{X}$ are i.i.d. according to known P_{X^p} [Candès, Fan, Janson, Lv 18].

- Design **knockoff mechanism** $P_{\tilde{X}^p|X^p}$ satisfying *exchangeability*

$$(X^p, \tilde{X}^p)_{\mathsf{swap}(S)} \stackrel{\text{dist.}}{=} (X^p, \tilde{X}^p), \quad \forall S \subseteq [p].$$

- Generate knockoff samples $(\tilde{X}_{i,j})_{i=1\ldots n,\, j=1,\ldots,p}$.
- Solve Lasso for $\mathbf{Y}$ and $\underline{\mathbf{X}} := [\mathbf{X}, \tilde{\mathbf{X}}]$.
- Compute *symmetric statistics* $W_j := |\hat{\theta}_j| - |\hat{\theta}_{j+p}|$.
- Find minimum $T > 0$ such that $\frac{|\{W_j \le -T\}|}{|\{W_j \ge T\}| \vee 1} \le q$.
- $\hat{H}_1 := \{1 \le j \le p \colon W_j \ge T\}$.

Big problem:
How to Design the Knockoff Mechanism $P_{\tilde{X}^p|X^p}$?

Let P_{X^p} be Gaussian. Exchangeability imposes that the covariance of $(X^p, \tilde{X}^p)$ has the form

$$\begin{bmatrix} \mathbf{\Sigma} & \mathbf{\Sigma} - \mathrm{diag}(\mathbf{s}) \\ \mathbf{\Sigma} - \mathrm{diag}(\mathbf{s}) & \mathbf{\Sigma} \end{bmatrix}$$

Existing choices in [Barber, Candès 2015][Candès, Fan, Janson, Lv 2018]:

- **SDP-Knockoff**: Minimize $\|\mathbf{s}\|_1$ by solving SDP
- **Equi-Knockoff**: Maximize $s_1 = \cdots = s_p$
- **ASDP-Knockoff**
- ...

No conclusive prior theory to guide our choice

Power Analysis: Previous results

All knockoff mechanisms control FDR; we look for one to maximize PWR.

Sufficient conditions for consistency:

- Fixed $\mathbf{\Sigma}$ and $n \to \infty$ [Fan et al. 2019].
- Fixed n/p, $X_{i,j} \sim \mathcal{N}(0,1)$ i.i.d., and SNR$\to \infty$ [Weinstein, Barber, Candès 2017].

Those sufficient conditions seem to suggest one should maximize $\lambda_{\min}(\underline{\mathbf{\Sigma}})$.

Our Contribution 1:
Necessary Sufficient Condition for Consistency

Suppose that

- $\delta := n/p > 2$ is fixed.
- noise level $\mathrm{var}(N_i)/n = 1$.
- $(\underline{\theta}^{(p)}, \underline{\mathbf{\Sigma}}^{(p)})_{p\geq 1}$ admit a suitable *distributional limits*.
- FDR budget $q \leq \lim_{p\to\infty} \frac{|\mathcal{H}_0|}{p}$.

Let d_{LP} be the *Lévy-Prokhorov metric*, $\underline{\mathbf{P}}^{(p)} := (\underline{\mathbf{\Sigma}}^{(p)})^{-1}$, and

$$\mathsf{ESD}^{(p)} := d_{\mathsf{LP}}\left(\frac{1}{2p}\sum_{j=1}^{2p}\delta_{\underline{\mathbf{P}}_{jj}^{(p)}},\ \delta_0\right), \quad \mathsf{ESD} := \lim_{p\to\infty} \mathsf{ESD}^{(p)}.$$

Main result:

$f(\mathsf{ESD}) \leq 1 - \lim_{p\to\infty} \mathsf{PWR}^{(p)} = F(\mathsf{ESD})$, where f and F are increasing functions that vanish at 0.

The Lévy-Prokhorov metric

Let P, Q be two probability measures on $\mathbb{R}$.

$$d_{\mathsf{LP}}(P,Q) := \inf\{\epsilon > 0\colon P(\mathcal{A}) \le Q(\mathcal{A}^\epsilon) + \epsilon,\, Q(\mathcal{A}) \le P(\mathcal{A}^\epsilon),\, \forall \mathcal{A}\},$$

where $\mathcal{A}^\epsilon$ denotes the ϵ-neighborhood of $\mathcal{A}$.

Remarks:

- d_{LP} metrize the topology of weak convergence.
- Weak convergence is weaker than convergence in norm, in particular ℓ_∞-convergence. Thus $\lambda_{\min}(\underline{\mathbf{\Sigma}})$ large is sufficient for consistency, though not necessary.

Gaussian limit of Debiased Lasso

$$\underline{\hat{\theta}}^u := \underline{\hat{\theta}} + \frac{1}{n - \|\underline{\hat{\theta}}\|_0} \underline{\mathbf{\Sigma}}^{-1} \mathbf{X}^\top (\mathbf{Y} - \underline{\mathbf{X}\hat{\theta}}).$$

Replica Analysis Claim: The empirical distribution of $\left\{ \frac{(\underline{\hat{\theta}}_j^u - \underline{\theta}_j)}{\tau (\underline{\mathbf{\Sigma}}^{-1})_{jj}^{1/2}} \right\}_{j=1}^{2p}$ converges to standard Gaussian (for some τ).

- Rather different than the classical asymptotic normality of MLE!
- [Tanaka 2001] [Guo & Verdú 2005] [Javanmard & Montanari 14]
- Bound τ using only λ and δ when $\delta > 2$.
- Rigorous alternative proofs for: identity $\underline{\mathbf{\Sigma}}$, or sublinear sparsity.

Intuitions and Extensions of the Gaussian Limit Property

- The replica calculation suggests that asymptotically $(\underline{\hat{\theta}}^u - \underline{\theta})$ is distributed as $\tau \mathbf{P}^{1/2}\mathbf{G}$ where $\mathbf{G}$ is standard Gaussian vector.
- [Javanmard & Montanari 14] claims that this is precisely true for the 1-*dimensional projection*: the empirical distribution of the coordinates.
- In our proof of the necessity for consistency, we need an extension to the 2-*dimensional projection*: the empirical distribution of the $(j, j+p)$-th coordinates, $j = 1, \ldots, p$.

Viewpoints on How to Design the Knockoff Mechanism

- **Statistical view:** we should optimize $P_{\tilde{X}^p|X^p}$ to minimize the empirical distribution of the diagonals of $\underline{\mathbf{P}}$ in Lévy-Prokhorov metric.
- **Computational view:** find simple $P_{\tilde{X}^p|X^p}$ with competitive statistical performance.
- **Structure in the data:** e.g. graphical model, sparsity, tree mixtures. . .

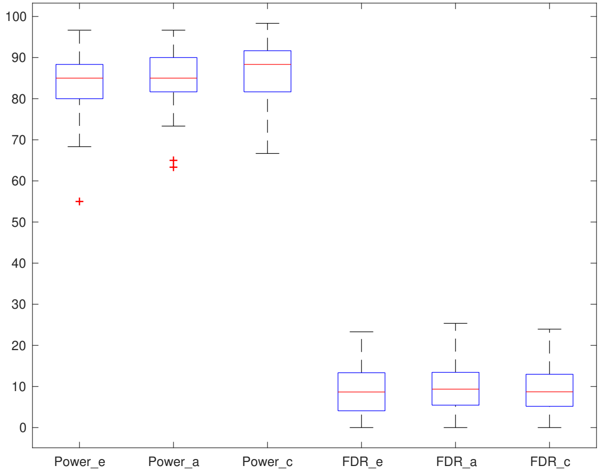
Contribution 2: Knockoff for Tree Models

Conditional independence knockoff:

$P_{X^p \tilde{X}^p}$ in which X_j and $\tilde{X}_j$ are independent conditioned on $X_{\backslash j}$, $\forall j$.

- Well-defined iff X^p is a walk-summable model (including all trees).
- Analytic formula for $\underline{\boldsymbol{\Sigma}}$, hence almost no computational cost.
- Competitive statistical performance against previous computationally involved mechanisms (e.g. SDP).
- Tree models are common in genetics, the common playground of FDR control.

Binary Tree, Equal Correlations



Binary Tree, Random Correlations