DeepDRK: Deep Dependency Regularized Knockoff for Feature Selection

Hongyu Shen[1], Yici Yan[2], Zhizhen Zhao[1]

[1]Department of Electrical and Computer Engineering, UIUC, IL, USA [2]Department of Statistics, UIUC, IL, USA

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

➢ **Goal:** Select the features associated with the linear response Y, given the covariate design matrix X, with a controlled false discovery rate (FDR) under the Model-X knockoff framework.

➢ **Challenges:** Unknown data distribution and small sample size.

➢ **Approach:** Deep generative models have been used for knockoff generations for non-Gaussian data:

- Deep Knockoff [4], KnockoffGAN [2], sRMMD [3], and DDLK [6]
- Performance declines as the sample size decreases and the data distributions become more complex.

➢ **Our approach:** DeepDRK generates knockoffs with a novel transformer-based generator and a random perturbation technique.

Preliminary

➢ **Core ingredients:** Learned knockoff variables $\tilde{X}$ and knockoff statistics $w_j((X,\tilde{X}),Y)$ for $j \in [p]$.

➢ Two required conditions for the knockoff variables and the knockoff statistics:

- Swap property: $(X,\tilde{X})_{\text{swap}(B)} \overset{d}{=} (X,\tilde{X}), \quad \forall B \subset [p]$;
- Flip-sign property:

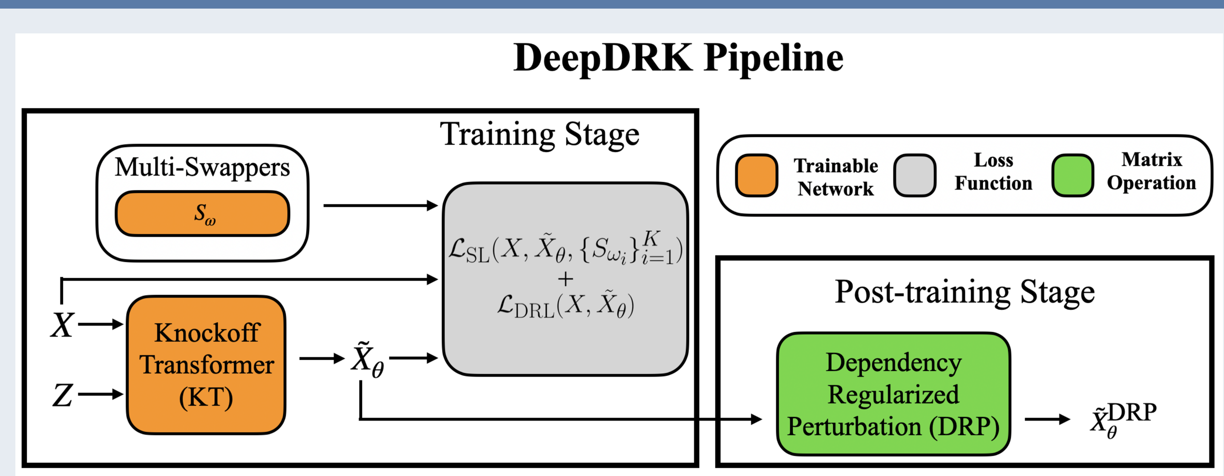
$$w_j\left((X,\tilde{X})_{\text{swap}(B)},Y\right) = \begin{cases} w_j((X,\tilde{X}),Y), & \text{if } j \notin B \\ -w_j((X,\tilde{X}),Y), & \text{if } j \in B \end{cases}$$

;

➢ Feature selection with controlled FDR at nominal level q:

- **Selection rule:** $\mathcal{S} = \{w_j \geq \tau_q\}$;
- **Threshold:** $\tau_q = \min_{t>0}\left\{t : \frac{1+|\{j:w_j\leq -t\}|}{\max(1,|\{j:w_j\geq t\}|)} \leq q\right\}$.

Methodology-Training Stage



➢ The Knockoff Transformer takes X and i.i.d. standard Gaussian random variables Z as the inputs to generate the knockoffs $\tilde{X}_\theta$;

➢ Use K swappers $\{S_{\omega_i}\}_{i=1}^K$ to create adversarial environments for testing the swap property;

➢ The swap loss $\mathcal{L}_{\text{SL}}(X,\tilde{X}_\theta,\{S_{\omega_i}\}_{i=1}^K)$ aims to enforce the swap property;

➢ The dependency regularization loss $\mathcal{L}_{\text{DRL}}(X,\tilde{X}_\theta)$ aims to decorrelate the data X and the knockoff $\tilde{X}_\theta$.

➢ **Training objective:**

$$\min_\theta \max_{\omega_1,\ldots,\omega_K} \left\{\mathcal{L}_{\text{SL}}(X,\tilde{X}_\theta,\{S_{\omega_i}\}_{i=1}^K) + \mathcal{L}_{\text{DRL}}(X,\tilde{X}_\theta)\right\}$$

➢ The **swap loss** includes three terms:

$$\mathcal{L}_{\text{SL}}(X,\tilde{X}_\theta,\{S_{\omega_i}\}_{i=1}^K) = \frac{1}{K}\sum_{i=1}^K \text{SWD}((X,\tilde{X}_\theta),(X,\tilde{X}_\theta)_{S_{\omega_i}})$$
$$+\lambda_1 \cdot \text{REx}(X,\tilde{X}_\theta,\{S_{\omega_i}\}_{i=1}^K) + \lambda_2 \cdot \mathcal{L}_{\text{swapper}}(\{S_{\omega_i}\}_{i=1}^K)$$

- The first term uses sliced Wasserstein distance to measure the distance between the joint distributions of $(X,\tilde{X}_\theta)$ and $(X,\tilde{X}_\theta)_{S_{\omega_i}}$;
- The second term measures the variance of the SWDs under different swap realizations;
- The third term prevents the mode collapse on the parameters ω_i of different swappers;

➢ $\mathcal{L}_{\textbf{DRL}}(X,\tilde{X}_\theta)$ uses the sliced Wasserstein correlation (SWC) to quantitatively measure the dependency between X and $\tilde{X}_\theta$.

Methodology–Post-training Perturbation

➢ Perturb the learned knockoff $\tilde{X}_\theta$:

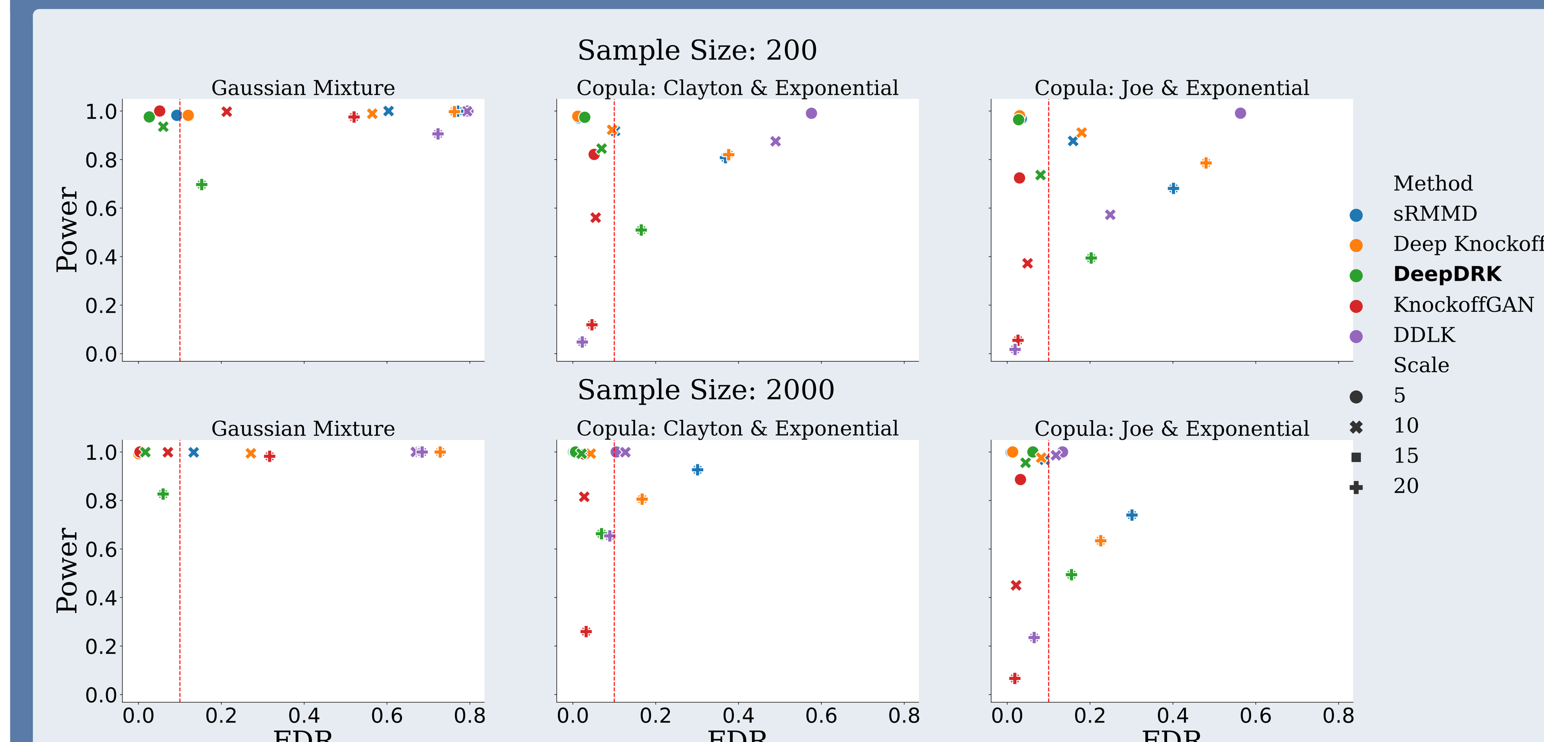
$$\tilde{X}_{\theta,n}^{\text{DRP}} = (1-\alpha_n)\cdot\tilde{X}_\theta + \alpha_n \cdot X_{\text{rp}},$$

where X_{rp} is the random row permutation of the design matrix X.

➢ The perturbation aims to reduce collinearity [5].

➢ As $n \to \infty$, $\alpha_n \to 0$.

Results-Synthetic Data



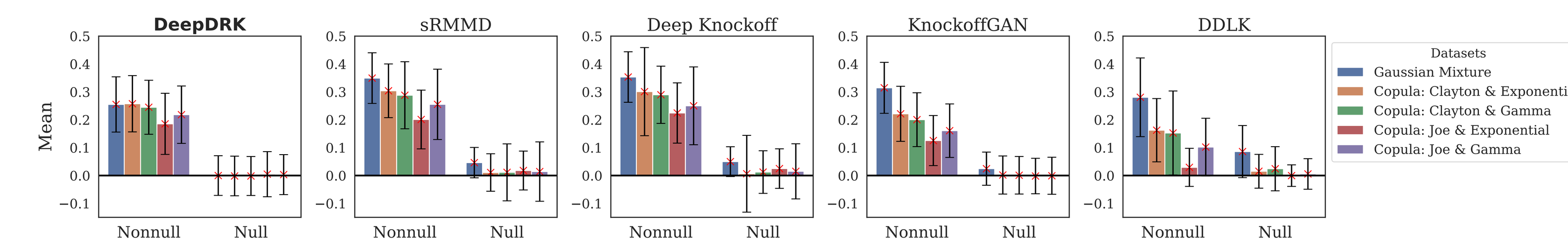
➢ Sample size: $n = 200$ or 2000; data dimension: $p = 100$

➢ Model: $Y \sim \mathcal{N}(X^T\beta, 1)$; feature sparsity: 20

➢ Nonnull $\beta_j \sim \frac{p}{\text{scale}\cdot\sqrt{n}} \cdot \text{Rademacher}(0.5)$;

➢ FDR nominal threshold $q = 0.1$.

Discussion

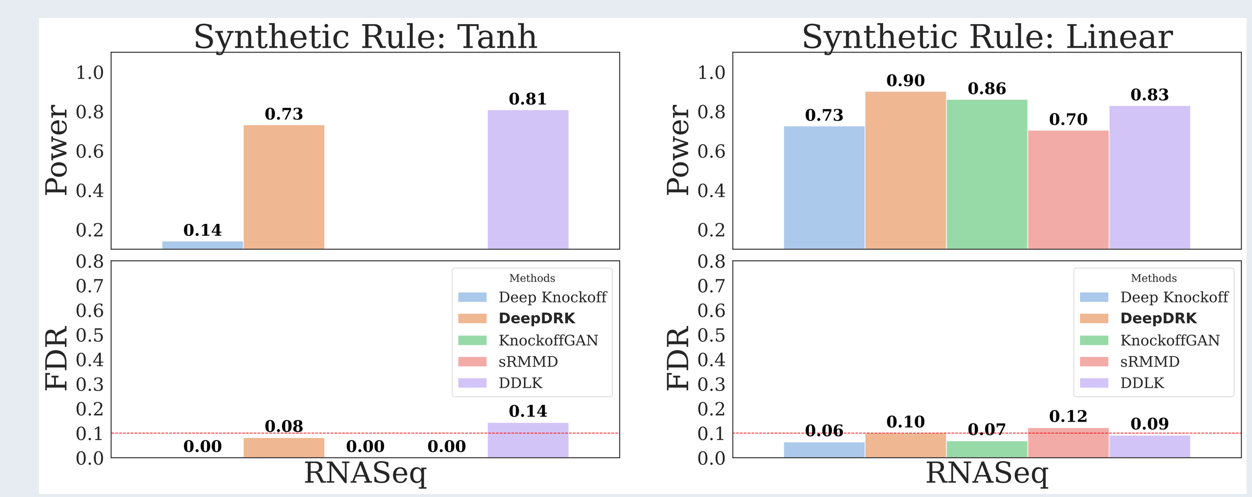


➢ Compare the means and the standard deviations of the knockoff statistics w_j's;

➢ Positive shifts in the null knockoff statistics from baseline models cause:

- smaller thresholds τ_q, as fewer null statistics are remaining on the negative side (lower $|\{j : w_j \leq -t\}|$), where $\tau_q = \min_{t>0}\left\{t : \frac{1+|\{j:w_j\leq -t\}|}{\max(1,|\{j:w_j\geq t\}|)} \leq q\right\}$;
- Increase in the number of false positives given the selection rule $\mathcal{S} = \{w_j \geq \tau_q\}$.

Results–Semi-synthetic Data



➢ X drawn from single-cell RNA sequencing (scRNA-seq) [1] and used to simulate response Y;

➢ $n = 10000$ and $p = 100$.

Conclusion

➢ We developed DeepDRK for feature selection with controlled FDR for non-Gaussian data and limited sample size;

➢ Paper link: https://arxiv.org/pdf/2402.17176v2;

➢ GitHub: https://github.com/nowonder2000/DeepDRK.

Reference

[1] Derek Hansen, Brian Manzo, and Jeffrey Regier. "Normalizing Flows for Knockoff-free Controlled Feature Selection". In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022, pp. 16125–16137.

[2] James Jordon, Jinsung Yoon, and Mihaela van der Schaar. "KnockoffGAN: Generating knockoffs for feature selection using generative adversarial networks". In: *ICLR*. 2018.

[3] Shoaib Bin Masud et al. "Multivariate soft rank via entropy-regularized optimal transport: Sample efficiency and generative modeling". In: *Journal of Machine Learning Research* 24.160 (2023), pp. 1–65.

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

[5] Asher Spector and Lucas Janson. "Powerful knockoffs via minimizing reconstructability". In: *The Annals of Statistics* 50.1 (2022), pp. 252–276.

[6] Mukund Sudarshan, Wesley Tansey, and Rajesh Ranganath. "Deep direct likelihood knockoffs". In: *NeurIPS*. Vol. 33. 2020.