Normalizing Flows for Knockoff-free Controlled Feature Selection

Derek Hansen, Brian Manzo, and Jeffrey Regier
University of Michigan Department of Statistics

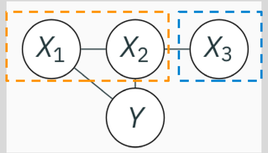
Motivation

Black-box regression and classification models can accurately predict a response based on input features. However, practitioners often need to know *which* specific features drive variation in the response, and they need to do so in a way that limits the number of false discoveries.

Controlled Feature Selection

- Observed features $X_1, X_2, \ldots, X_D$
- Response variable Y
- Black-box model of Y | X
  - (LASSO, random forest, deep neural network)

Unknown association graph between X and Y



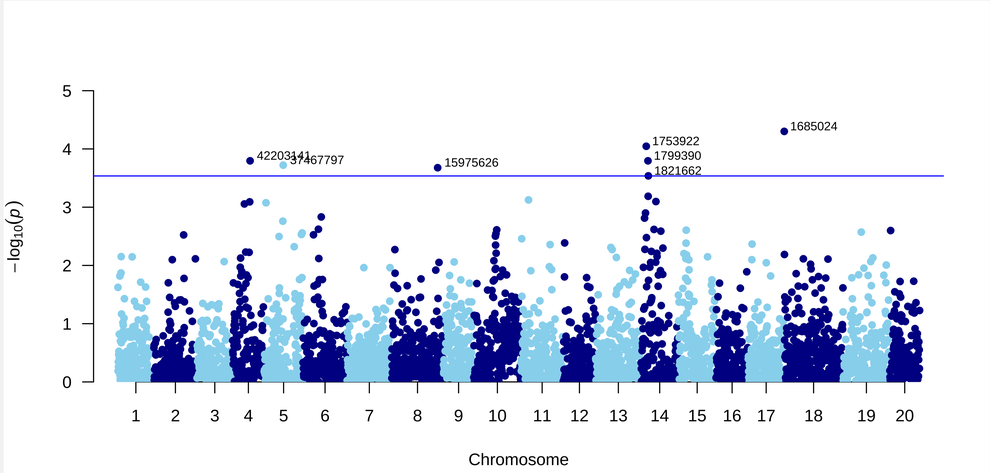
- A **relevant** feature has a direct association with the response Y
- A **null** feature is independent of the response conditional on the other features

Goal:
- Maximize **power**: the proportion of **relevant** features that are selected
- Control the **false discovery rate (FDR)**: the proportion of selected features that are **null**

FlowSelect

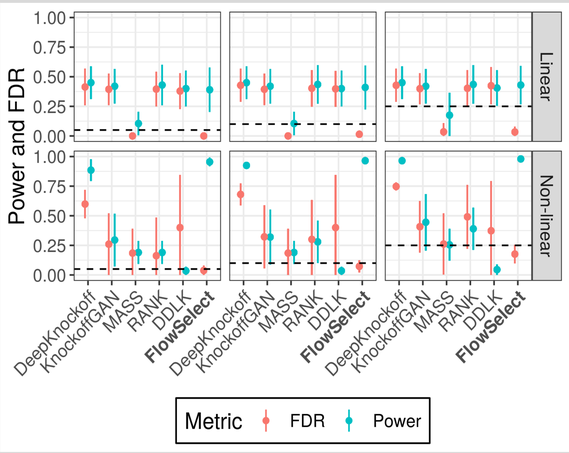
1: Fit probability density of features with normalizing flows → 2: Sample null features using MCMC targeting fitted density → 3: Calculate p-value of feature statistic with true feature vs null features

GWAS Application



- Predict soybean oil content (Y) from SNPs (X), which encode different genetic variants.
- Use FlowSelect to calculate p-values for >4000 SNPs
- Set selection threshold (blue bar) to determine significant SNPs with FDR = 20% (using Benjamini-Hochberg multiple testing correction).

Validation Experiment



- Generate features from mixture of Gaussians
- Two response models (linear, non-linear)
- Evaluate the **observed false discovery rate (FDR)** and **power** of FlowSelect vs empirical knockoff approaches across 20 replicates.

Conclusion

FlowSelect consistently controls FDR and demonstrates higher power than competing empirical knockoff approaches.

Link to paper (arXiV: 2106.01528)