False discovery rate control for polygenic risk prediction

Benjamin Chu

4/21/2022 CAPE meeting (online)

Methods for Polygenic Risk Scores (PRS)

Goal of PRS: build a model to predict *complex* human traits using genetic (e.g. GWAS) data.

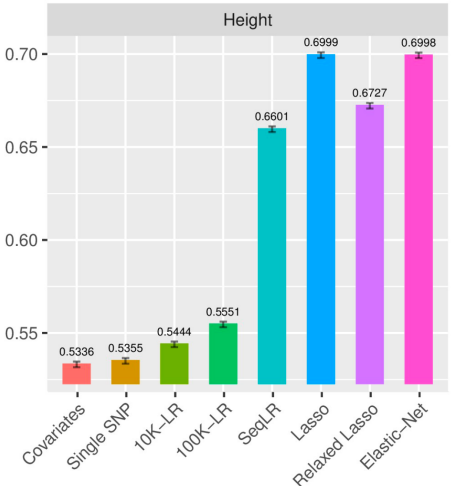
- **Clumping + thresholding** (C+T) [Choi and O'Reilly, 2019]
  1. Remove highly correlated SNPs (before or after selection)
  2. Find significant SNPs passing *loose p-value threshold*
  3. Estimate $\hat{\beta}_j$ from a univariate regression
  4. Predicted phenotype for sample i: $\hat{y}_i = \sum_j x_{ij}\hat{\beta}_j$
- **Penalized regression methods** e.g.

$$\hat{\boldsymbol{\beta}} = \text{minimize} \;\; \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 \qquad \text{(LASSO)}$$

$$\hat{\boldsymbol{\beta}} = \text{minimize} \;\; \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 \text{ s.t. } \|\boldsymbol{\beta}\|_0 \leq k \qquad \text{(IHT)}$$

  where λ and k are sparsity inducing parameters tuned by cross validation.
- Other alternatives: Bayesian methods (BayesR [Moser et al., 2015]), T-Trees (random forest) ...etc

Strength and (possible?) weakness of LASSO for PRS



- LASSO/elastic-net typically outperforms other methods
- The number of non-zero beta for height is 118,322
- How does this model predict in other populations?

Figure: Qian et al. 2020 (PLOS genetics), y-axis measures R^2 predictive performance

Problem: PRSs are population specific

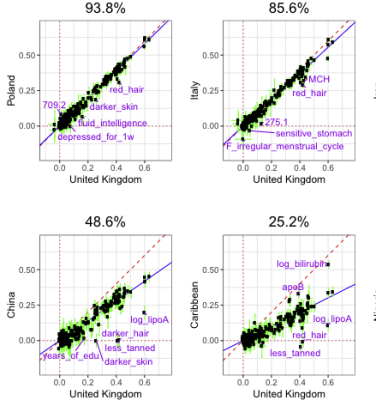


Figure: Lasso study by Prive et al 2022 (AJHG). Prediction accuracy (partial r^2) drops across different group.

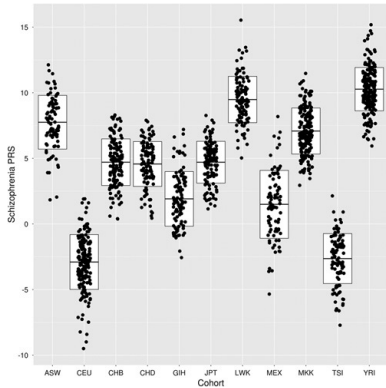


Figure: C+T study by David Curtis 2018 (Psychiatric Genetics). PRS for schizophrenia are highly sensitive to ethnic background

In both studies, a large number of predictors were selected.

PRS are population specific (but why?)

Watch talks[1] by Graham Coop and Jonathan Kaplan

- Unmeasured (true) population differences
  - Due to genetic drift/selection, which may not be surveyed
  - Unmeasured environmental effects
- Estimation bias
  - Unadjusted confounders (e.g. population stratification) can produce excess false positive
  - Imprecise effect size estimation (e.g. non-additive terms, interaction, epistasis...etc)
  - Intrinsic problems with data (e.g. LD and rare alleles) that cannot be modeled well

We will develop methods to address problems related to estimation bias.

[1] https://cehg.stanford.edu/evolgenome-seminars

Idea for solution

1. Some papers[2] suggest controlling the number of **false discoveries** will help prediction
2. Most PRS models include an enormous number of variants
3. Can prediction improve (broadly speaking) if we select a cleaner set of predictors?

Lets explore this idea with **Knockoffs**

[2] For example, [Abramovich et al., 2006] and [Benjamini and Gavrilov, 2009]

The Goals of Statistical Knockoffs

Knockoffs are designed for two purpose:

- Instead of controlling FWER[3], the knockoff procedure controls the FDR

$$FDR = E\left(\frac{\#\text{false positives}}{\#\text{ total discoveries}}\right)$$

  This significantly improves power.
- Knockoff based inference tests *conditional hypotheses*. If G is a SNP or a group of SNPs, we test

$$\mathcal{H}_0 = Y \perp\!\!\!\perp X_G \mid X_{-G}$$

  Conditioning on X_{-G} removes SNPs only marginally associated with the trait due to e.g. linkage disequilibrium, prioritizing causal associations

[3] which is what Bonferroni correction does in GWAS

Properties of Knockoffs

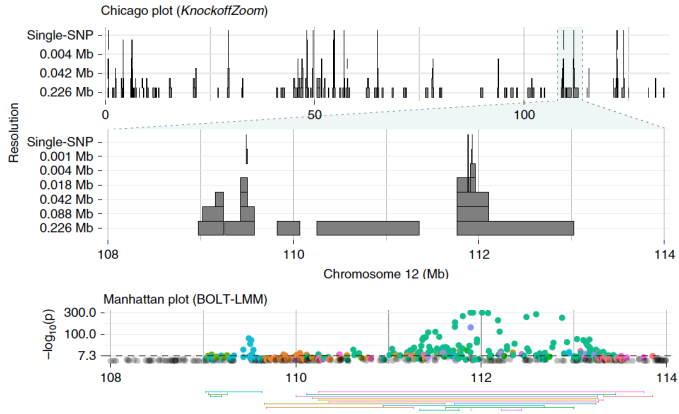
- The knockoff-filter wraps *any* algorithm that provides feature-importance scores and helps us decide which variables to choose.
  - In particular, it does not need valid p-values (thus we can quantify FDR of e.g. LASSO)
  - The selected variables have guaranteed FDR control.
- Knockoffs work for covariates with arbitrary correlation structure, unlike e.g. Benjamini-Hochberg

The Knockoff procedure

1. For each sample $X \in \{0,1,2\}^p$, generate *knockoffs* $\tilde{X} \in \{0,1,2\}^p$ satisfying
   - $Y \perp\!\!\!\perp \tilde{X} \mid X$
   - $(X, \tilde{X}) \overset{d}{=} (X, \tilde{X})_{\text{swap}(S)} \forall S$. E.g. If $S = \{2\}$, then $(X_1, X_2, X_3, \tilde{X}_1, \tilde{X}_2, \tilde{X}_3) \overset{d}{=} (X_1, \tilde{X}_2, X_3, \tilde{X}_1, X_2, \tilde{X}_3)$
2. Compute feature importance statistic on concatenated matrix $\left[\boldsymbol{X}\ \tilde{\boldsymbol{X}}\right]$
3. Compute knockoff scores $\boxed{W_G = \mathit{ImportanceScore}(G) - \mathit{ImportanceScore}(\tilde{G})}$ for all Gs
4. Choose all G such that $W_G \geq \tau$, where τ depends on FDR threshold q via

$$\tau = \min\left\{t > 0 : \frac{1 + \#\{G : W_G \leq -t\}}{\#\{j : W_G \geq t\} \vee 1} \leq q\right\}$$

Knockoffs are successful for GWAS



We previously found Knockoff-Lasso finds **more predictors** and **localizes** better than traditional LMM.

Figure: Sesia et al 2020 (Nature communications)

How to use Knockoffs for prediction?

Recall: We want to do prediction.

After Knockoffs select a set of Gs, there are multiple ways to build a PRS model

1. If Gs are singleton, we can use least squares to fit low-dimensional model on selected Gs
2. If Gs are SNP groups,
   - Option 1: Run a second Lasso routine only allowing non-zero entries in selected Gs
   - Option 2: Run least squares fit on the non-zero entries of the selected Gs

Method summary

Goal: we want to build a PRS model that can predict across ethnically diverse populations.

1. We start with PRS methods that have principled selection *and* estimation procedures, such as LASSO and IHT penalized regression
2. In practice, the best performing PRSs select absurdly many predictors (e.g. $>$ 100k for height), which raises *portability* issues.
3. We will kill (hopefully false positives) predictors by applying the knockoff filter, which will give us a **sparser set with guaranteed FDR control**
4. This is an experimental project that uses Knockoffs in a new way. We do not have finalized results!

Does this procedure work?

Experiment 1: Independent genotypes

- Simulate 6000×50000 matrix $\boldsymbol{X}$ by

$$x_{ij} \sim \text{Binomial}(2, p_j), \quad p_j \sim \text{Uniform}(0, 1)$$

- 5000 samples used as training data, 1000 samples for testing
- Pick $|\mathcal{S}_{\text{causal}}| = 100$ causal SNPs chosen uniformly, $h^2 = 0.5$
- Simulate phenotypes (using standardized $\boldsymbol{X}$):

$$y_i = \sum_{j \in \mathcal{S}_{\text{causal}}} x_{ij}\beta_j + \epsilon_i, \quad \beta_j \sim N\left(0, \sqrt{\frac{h^2}{2|\mathcal{S}_{\text{causal}}|}}\right), \quad \epsilon_i \sim N(0, 1 - h^2).$$

Evaluate performance:

$$R^2 = 1 - \frac{\|\boldsymbol{y} - \boldsymbol{X}_{test}\hat{\boldsymbol{\beta}}\|_2^2}{\|\boldsymbol{y} - \overline{\boldsymbol{y}}\|_2^2} \quad \text{(For predicting in same population)}$$

Partial r^2 (For predicting across populations)

Experiment 1 results: Independent genotypes

	LASSO	LASSO-ko	IHT	IHT-ko
R^2	0.42	0.46	0.45	0.45
power	0.68	0.56	0.55	0.53
FDR	0.83	0.11	0.08	0.04

Table: Single SNP knockoff results with target FDR $= 0.1$, averaged over 100 runs

- R^2 performance: LASSO-ko $\geq$ IHT-ko $=$ IHT $>$ LASSO
- Knockoffs control FDR, pay a small price in power

Experiment 2: UK Biobank genotypes

UK Biobank: $\sim$ 500,000 samples, primarily British but includes numerous other ethnicities. After QC, we retain

Population	Sample size	Description
British	320094	
Irish	1066	
White	15348	
Pakistani	1617	
Bangladeshi	213	
Indian	5188	
White and Asian	722	Mixed
Chinese	1442	
Asian	1695	
Caribbean	3859	
White and Black	912	Mixed
African	3112	

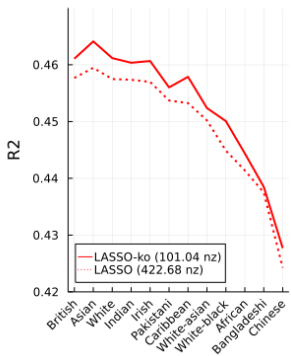
Experiment 2 setup

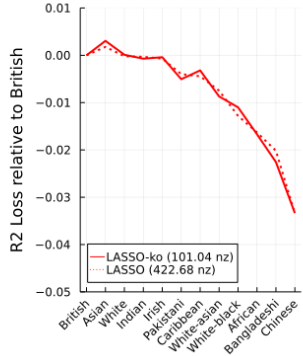
- Train data: chromosome 10 (29,481 SNPs) of UK Biobank with 10,000 British samples (i.e. very small subset)
- Phenotypes: $\boldsymbol{y}$ simulated in the same way as before, i.e.
  - Pick $|\mathcal{S}_{\text{causal}}| = 100, 1000$ causal SNPs chosen uniformly, $h^2 = 0.5$
  - Simulate phenotypes (using standardized $\boldsymbol{X}$):

$$y_i = \sum_{j \in \mathcal{S}_{\text{causal}}} x_{ij}\beta_j + \epsilon_i, \quad \beta_j \sim N\left(0, \sqrt{\frac{h^2}{2|\mathcal{S}_{\text{causal}}|}}\right), \quad \epsilon_i \sim N(0, 1 - h^2).$$

- Test data: all samples in UKB stratified by ethnicity

UKB experiments Lasso results (100/29481 SNPs causal, n=10000)



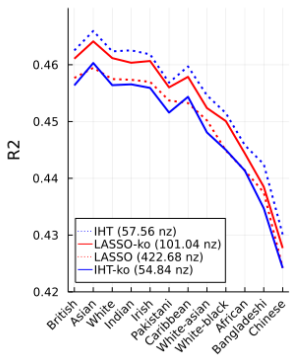


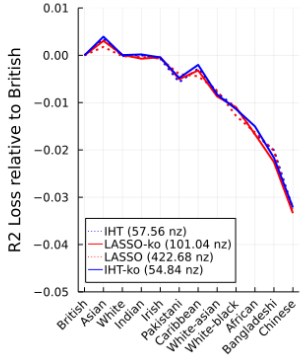
	LASSO	LASSO-ko
# nonzero $\hat{\beta}$	422.7	458.9
# $\hat{\beta}$ selected	422.7	101.0
power (group power)	0.76	0.66 (0.70)
FDR (group FDR)	0.82	0.34 (0.12)

Table: Group knockoff (res5) results with target FDR = 0.1

- Lasso have *better power* but finds a lot of junk
- Knockoffs improves FDR in exchange for power.
- Here, the trade-off improved prediction.

UKB experiments results (100/29481 SNPs causal, n=10000)

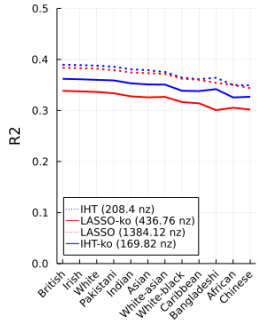


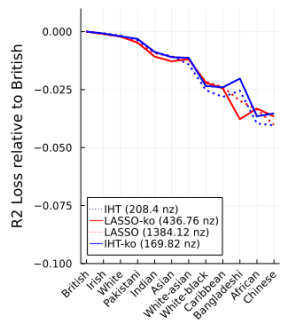


	IHT	IHT-ko	LASSO	LASSO-ko
# nonzero $\hat{\beta}$	57.6	56.6	422.7	458.9
# $\hat{\beta}$ selected	57.6	54.84	422.7	101.0
power (group power)	0.54	0.52 (0.57)	0.76	0.66 (0.70)
FDR (group FDR)	0.05	0.04 (0.01)	0.82	0.34 (0.12)

- Standard IHT has good power and controlled FDR
- If FDR was good to start out with, knockoffs hurt prediction performance

UKB experiments results (500/29481 SNPs causal, n=10000)

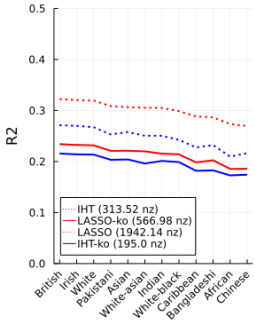


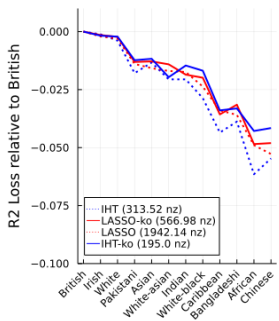


	IHT	IHT-ko	LASSO	LASSO-ko
# nonzero $\hat{\beta}$	208.4	190.42	1384.12	1537.5
# $\hat{\beta}$ selected	208.4	169.82	1384.12	436.76
power (group power)	0.36	0.30 (0.44)	0.55	0.39 (0.57)
FDR (group FDR)	0.17	0.12 (0.02)	0.80	0.55 (0.17)

Table: Group knockoff (res5) results with target FDR = 0.25

UKB experiments results (1000/29481 SNPs causal, n=10000)



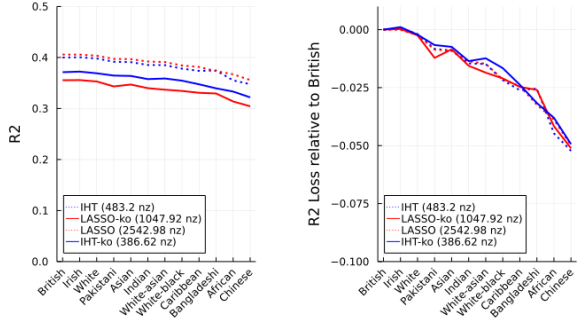


	IHT	IHT-ko	LASSO	LASSO-ko
# nonzero $\hat{\beta}$	313.5	489.1	1942.1	2053.4
# $\hat{\beta}$ selected	313.5	195.0	1942.14	567.0
power (group power)	0.21	0.14 (0.30)	0.43	0.25 (0.48)
FDR (group FDR)	0.28	0.25 (0.03)	0.78	0.57 (0.098)

Table: Group knockoff (res5) results with target group FDR 0.25

- Knockoffs traded too much power for FDR improvements
- Knockoffs predicts worse overall, but its performance degrade less

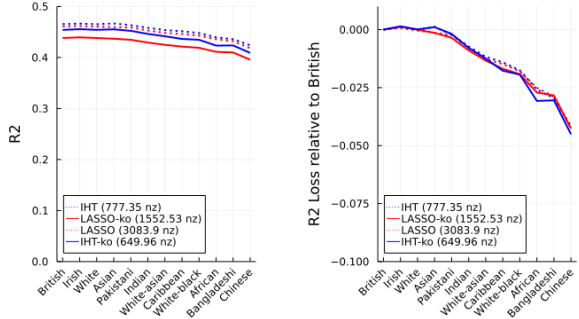
UKB experiments results (1000/29481 causal SNPs, n=20k)



	IHT	IHT-ko	LASSO	LASSO-ko
# nonzero $\hat{\beta}$	483.2	482.72	2542.98	2796.3
# $\hat{\beta}$ selected	483.2	386.62	2542.98	1047.92
power (group power)	0.37	0.32 (0.56)	0.59	0.44 (0.72)
FDR (group FDR)	0.21	0.17 (0.02)	0.77	0.58 (0.12)

Table: Group knockoff (res5) results with target FDR = 0.25

UKB experiments results (1000/29481 causal SNPs, n=50k)



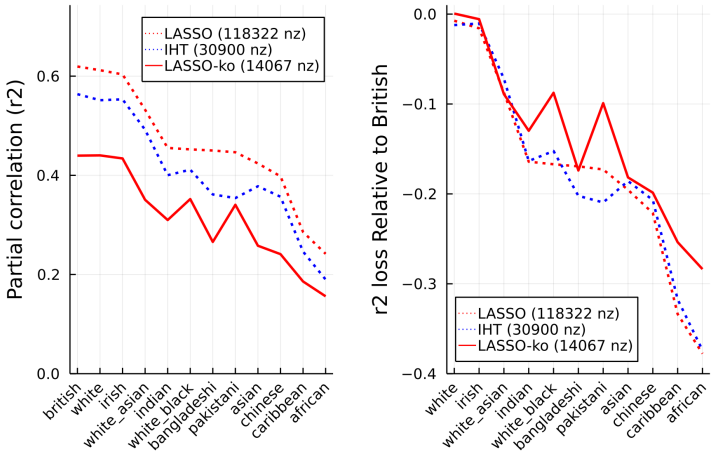
	IHT	IHT-ko	LASSO	LASSO-ko
# nonzero $\hat{\beta}$	777.347	797.122	3083.9	3429.8
# $\hat{\beta}$ selected	777.347	649.959	3083.9	1552.53
power (group power)	0.58	0.53 (0.76)	0.73	0.63 (0.87)
FDR (group FDR)	0.22	0.16 (0.03)	0.76	0.59 (0.15)

Table: Group knockoff (res5) results with target FDR = 0.25

Full UK Biobank analysis

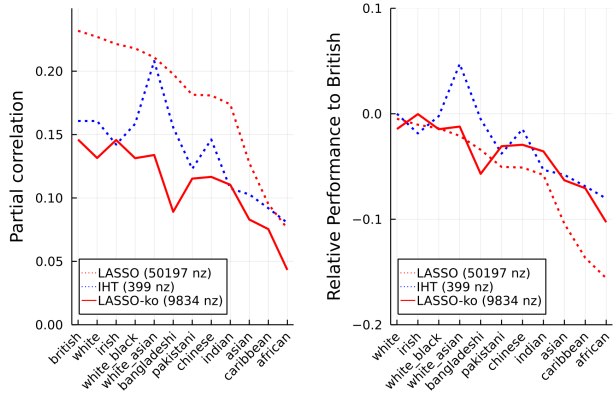
- Fit model on 80% British samples (256,075 samples and 591,513 SNPs)
- Test on 20% British samples (64,019 samples) and all other ethnicities
- Nongenetic covariates: sex, age, PC1-5
- Phenotypes: Height and systopic blood pressure (both continuous)

UKB analysis: Height



For highly polygenic trait, a more polygenic model is needed for better prediction, but knockoff model is more portable across population

UKB analysis: Systolic Blood Pressure



For highly polygenic trait, a more polygenic model is needed for better prediction, but sparser model is more portable across population

Summary

We are exploring a novel application of the knockoff framework for PRS prediction

- The knockoff filter gives us a way to control FDR to a desired level, **in exchange for power**
- This trade-off tends to be more "worth it" when the number of causal variants is low
- Knockoff filter may give worse prediction, but **it suffers less performance degradation** across populations
- We do not have real data example where knockoff models predict better in absolute scale, but we have only looked at very polygenic traits

Interested in Knockoffs?

Main theory papers:

- Barber, Rina Foygel, and Emmanuel J. Candès. "Controlling the false discovery rate via knockoffs." The Annals of Statistics 43, no. 5 (2015): 2055-2085.
- Candes, Emmanuel, Yingying Fan, Lucas Janson, and Jinchi Lv. "Panning for gold:'model-X'knockoffs for high dimensional controlled variable selection." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 80, no. 3 (2018): 551-577.

Genetics application papers:

- Sesia, Matteo, Eugene Katsevich, Stephen Bates, Emmanuel Candès, and Chiara Sabatti. "Multi-resolution localization of causal variants across the genome." Nature communications 11, no. 1 (2020): 1-10.
- Sesia, Matteo, Stephen Bates, Emmanuel Candès, Jonathan Marchini, and Chiara Sabatti. "False discovery rate control in genome-wide association studies with population structure." Proceedings of the National Academy of Sciences 118, no. 40 (2021).

References


Abramovich, F., Benjamini, Y., Donoho, D. L., and Johnstone, I. M. (2006).
Adapting to unknown sparsity by controlling the false discovery rate.
The Annals of Statistics, 34(2):584–653.


Benjamini, Y. and Gavrilov, Y. (2009).
A simple forward selection procedure based on false discovery rate control.
The Annals of Applied Statistics, pages 179–198.


Choi, S. W. and O'Reilly, P. F. (2019).
PRSice-2: Polygenic Risk Score software for biobank-scale data.
Gigascience, 8(7):giz082.


Moser, G., Lee, S. H., Hayes, B. J., Goddard, M. E., Wray, N. R., and Visscher, P. M. (2015).
Simultaneous discovery, estimation and prediction analysis of complex traits using a bayesian mixture model.
PLoS genetics, 11(4):e1004969.