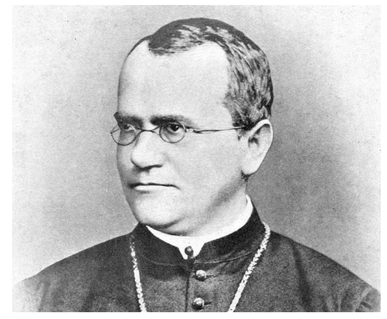
(Group) Iterative Hard Thresholding for GLM in Statistical Genetics



Benjamin Chu, Kevin Keys, Janet Sinsheimer, and Kenneth Lange
Dept. of Biomathematics, University of California, Los Angeles, CA, USA
Contacts: biona001@ucla.edu, kevin.keys@ucsf.edu, jsinshei@ucla.edu, klange@ucla.edu
Git Repository: https://github.com/biona001/IHT.jl

Background:

Individual variations in the DNA sequence are termed **single nucleotide polymorphisms** (SNPs), and they can be identified experimentally via **Genome Wide Association Studies** (GWAS). These studies aim to answer one main question:

Q: Which genetic variants explain variations in a trait?

Picture references: CONVERGE consortium, *Nature* (2015) and Balding et al. *Nature Review Genetics* (2006).

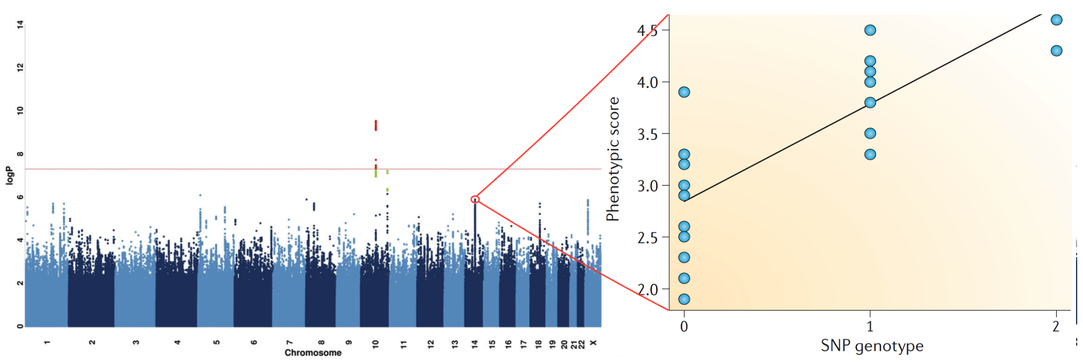


Figure 1: Traditional analysis assigns a p-value for each SNP based on a linear regression that test if slope $\neq 0$

Problems with Traditional Analysis (See Figure 1):

- Individual SNP association testing ignores joint effects of multiple SNPs and suffers high multiple testing burden. **Q: What is the best multivariate model selection method to use?**
- GWAS dataset sizes are growing rapidly (100+ GB). **Q: How to analyze them efficiently?**
- SNPs are sometimes **rare** with **small effect size**. **Q: How to separate *weak signals* from *noise*?**

IHT: ℓ_0 Sparsity without Shrinkage

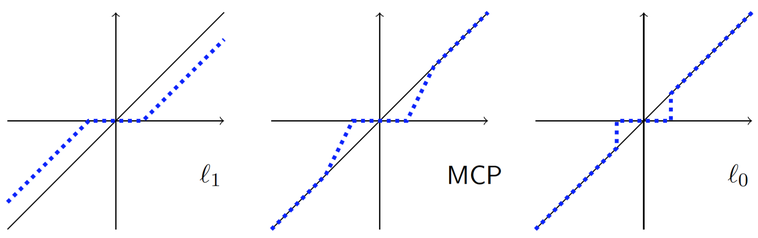


Figure 2: Biological meaning = IHT potentially captures variants with small effect size

Using IHT to find GLM coefficients via MLE

Iterative hard-thresholding (IHT) performs *feature selection* for $\mathbf{X}^{n\times p}$ when $p \gg n$. First we expand the framework of IHT to any generalized linear model. Then we modified the thresholding operator to enforce sparsity on the group-level as well as within groups.

(Group) IHT algorithm

Let $L(\beta)$ be the loglikelihood, $\mathbf{v} = -\nabla L(\beta)$ the negative score (gradient), $J(\beta)$ the expected (Fisher) information matrix, and $S_{R,k}$ a predictor set with at most R active groups and k active predictors per group. We maximize the loglikelihood iteratively via:

$$\beta^{(n+1)} = \mathcal{P}_{S_{R,k}}\left(\beta^{(n)} + s\mathbf{v}\right)$$

$$s = \frac{||\mathbf{v}||_2^2}{\mathbf{v}^T J(\beta^{(n)})\mathbf{v}} = \text{ step size}, \quad \mathcal{P}_{S_{R,k}}(\mathbf{v}) \text{ projects } \mathbf{v} \text{ to } S_{R,k}$$

Results

IHT Reconstruction Results

β_{true}	β_{normal}	β_{logistic}	β_{poisson}
2.15035	2.15076	2.31195	2.15246
1.42043	1.41833	1.51125	1.42261
-1.28871	-1.28929	-1.36258	-1.2846
-1.04068	-1.04139	-1.07631	-1.04235
0.546087	0.548324	0.700991	0.545524
0.360115	0.360985	0.417808	0.358854
0.331856	0.335209	0.388734	0.329764
0.279001	0.278694	0.304497	0.279266
0.103375	0.103152	Not Found	0.100905
0.0344145	0.0363563	Not Found	0.0329963

Simulated result with $n = 5000$ subjects and $p = 100,000$ SNPs. $\beta_{true} \sim N(0,1)$ and responses were simulated via canonical link.

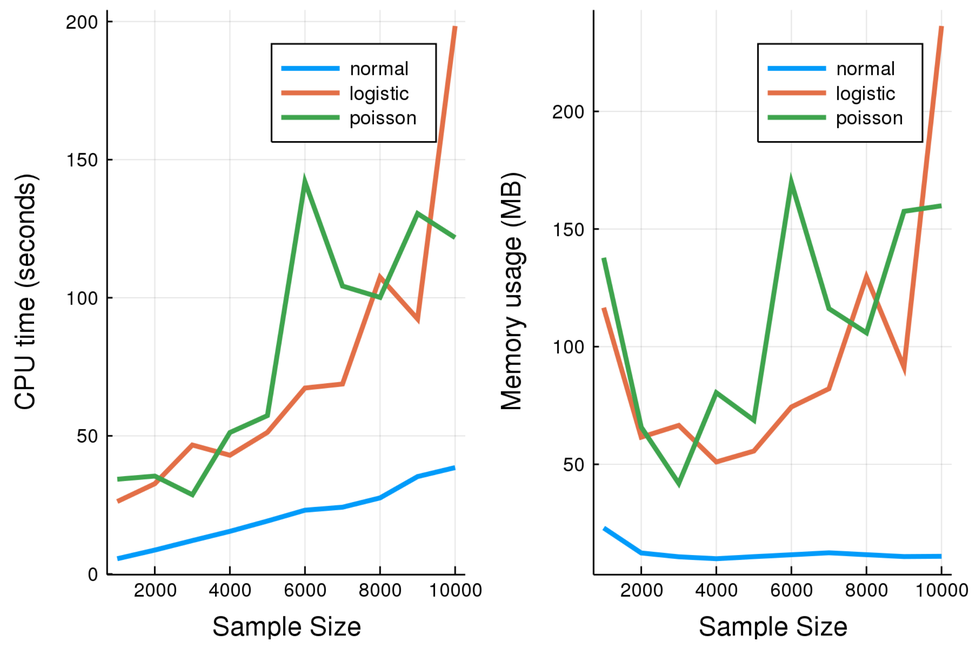


Figure 3: Speed and memory benchmark on 100000 SNPs

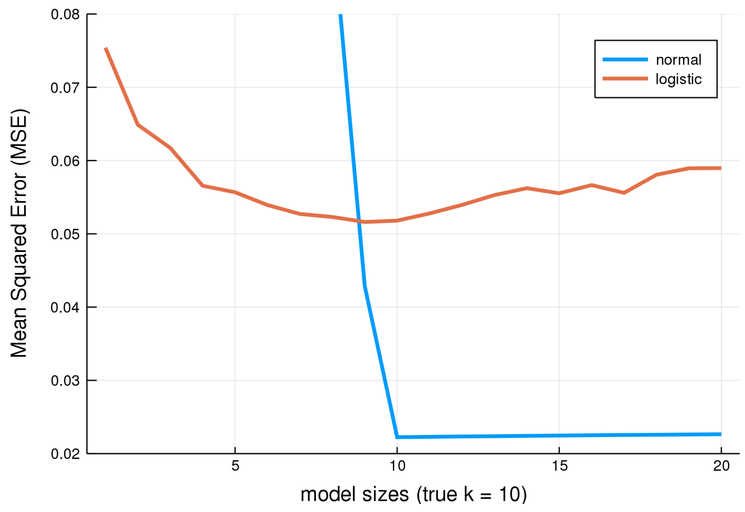


Figure 4: Cross-validation finds $k^{\text{estimate}}_{\text{normal}} = 10$ and $k^{\text{estimate}}_{\text{logistic}} = 9$

Summary of Results

- We applied IHT to perform multivariate model selection on genetics data.
- IHT can recover small effect sizes without shrinkage.
- Computational time increases linearly with data size.
- Memory usage remains relatively constant with data size.
- Cross Validation works well to determine the true model size k_{true}.

Acknowledgements

- This research was supported by NIH Training Grant in Genomic Analysis and Interpretation T32HG002536
- This research was supported by Google Summer of Code 2018 with NumFOCUS, Julia cohort