Estimating GLM coefficients for High Dimensional Data, with Application to Statistical Genetics

Benjamin B. Chu[1], Kevin L. Keys[2], Janet Sinsheimer[1], Kenneth Lange[1]

[1]Department of Biomathematics, University of California, Los Angeles

[2]Department of Medicine, University of California, San Francisco

UCLA

INTRODUCTION

➢ **GWASes** (genome wide association studies) tries to find **SNPs** (single nucleotide polymorphisms) ***associated*** with a human trait.

➢ SNP-by-SNP association tests ignore joint effects of SNPs ⇝ **BAD BIOLOGY ASSUMPTION**

➢ LASSO is a multivariate method that exhibits shrinkage and admits too many false positives ⇝ **BAD STATISTICS FOR GWAS**

➢ **Fill the gap with new algorithm: Iterative Hard Thresholding (IHT)**

➢ Project culminates in MendelIHT.jl ⇝ a scalable and open sourced package in Julia. Code Repository: https://github.com/biona001/MendelIHT.jl/tree/develop

METHODS

❖ **Problem:** Given SNP matrix $X \in \{0,1,2\}^{n \times p}$ where $p \gg n$, trait vector $y \in \mathbb{R}^n$, find a statistical model β such that $y \approx X\beta$. Then do it for non-linear case $y \approx g^{-1}(X\beta)$.

❖ **Setup:** Let $L(\beta)$ be the loglikelihood, $\nabla L(\beta)$ the gradient (score), and $J(\beta)$ the expected information matrix. Solve the following:

$$\hat{\beta} = argmax_{\beta:||\beta||_0 \leq k}\ L(\beta).$$

❖ **Algorithm:** In loglikelihood space, move in the positive score direction $v = \nabla L(\beta)$ multiplied by a step length η to increase $\nabla L(\beta)$ along the ray $\beta_{n+1} = \beta_n + \eta v$ prior to projection H:

$$\beta_{n+1} = H\left(\beta_n + \eta_n v_n\right), \quad \text{where } \eta = \frac{v_n^T v_n}{v_n^T J(\beta_n) v_n}$$

❖ H is a hard thresholding operator that projects a point to some sparsity set $S_{J,k}$ with at most J active "groups" and k active predictors per group.

RESULTS

Reconstruction Quality for GWAS data:

β_{true}	β_{Normal}		β_{Logistic}		β_{Poisson}		$\beta_{\text{Negative Binomial}}$	
	est. ± SD	P(found)	est. ± SD	P(found)	est. ± SD	P(found)	est. ± SD	P(found)
0.8	0.801 ± 0.015	1	0.805 ± 0.037	1	0.801 ± 0.007	1	0.81 ± 0.012	1
0.5	0.501 ± 0.013	1	0.499 ± 0.03	1	0.499 ± 0.006	1	0.505 ± 0.012	1
0.25	0.25 ± 0.016	1	0.252 ± 0.032	1	0.25 ± 0.009	1	0.249 ± 0.012	0.56
0.10	0.1 ± 0.015	1	0.157 ± 0.015	0.16	0.1 ± 0.01	0.94	0.0 ± 0.0	0

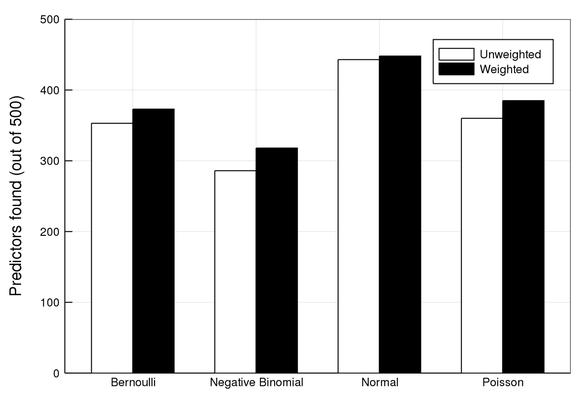
Table 1. Normal, Poisson, and Negative Binomial reconstruction results are unbiased, but logistic results may over-estimates small effect sizes. SD represents sample standard deviation of the *found* predictors, and P(found) indicates the proportion found. Here we ran 100 simulations using n = 5000 samples and p = 100,000 SNPs.

Comparing False Positives/Negatives with LASSO

	IHT			LASSO		
	False Positives	False Negatives	DNC	False Positives	False Negatives	DNC
Normal	0.04	1.28	0	28.54	0.8	0
Bernoulli	0.06	3.68	1	93.66	2.0	0
Poisson	0.5	3.32	1	29.74	3.52	7
Negative Binomial	0.82	5.08	2	NA	NA	NA

Table 2. Fifty independent cross validation studies for each response type shows IHT is superior in limiting the number of false positives. Here we used 1000 samples each with 10,000 SNPs. The number of false positives/ negatives are averaged over 50 runs, and DNC counts the number of runs that did not converge.

Weighted IHT



← Figure 1. Comparison of unweighting and weighting shows weighting can find around 10% more predictors. Here we assumed approximately 1/10 of all SNPs are within a protein coding region, including the 10 true predictors, while others are within introns. We assigned w = 2.0 to all SNPs within coding regions and w = 1.0 for all other SNPs.

Table 3: Comparison of doubly sparse group IHT with non-grouped IHT model selection in a case-control studies. Here there are a total of 250 common variants (maf = 0.4) and 1250 rare variants (maf = 0.005). ↓

Group IHT

	Common Variants Found	Rare Variants Found
Group IHT	185	649
Ungrouped IHT	179	639

Testing Scalability (on 1,000,000 SNPs)

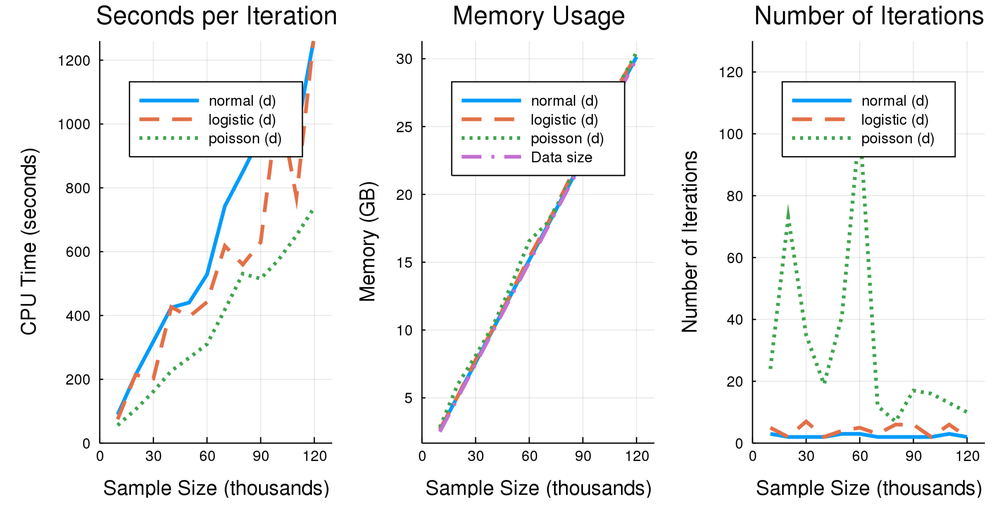


Figure 2. Time per iteration and memory usage scales linearly with data size. Figure shown is running IHT with debiasing (d). The largest problem should fit into a regular desktop computer with 32GB of RAM. Every run is performed on a intel-E5-2670 machine with 63G of RAM and a single 3.3GHz processor.

CONCLUSION

◆ IHT provides a principled way for variable selection and estimating precise coefficient

◆ For GWAS, IHT is much better than LASSO at controlling false positive rates

◆ Our implementation of IHT is scalable to the largest dataset today (e.g. UK Biobank)

ACKNOWLEDGEMENTS

✓ This research issupported by NIH Training Grant in Genomic Analysis and Interpretation T32HG002536

✓ This research issupported by Google Summer of Code 2018 with NumFOCUS and the Julia cohort.