Project 1: Iterative Hard Thresholding in GWAS: Generalized Linear Models, Prior Weights, and Double Sparsity

Problem: Given *dense* $\mathbf{y}, \mathbf{X}$ find *sparse* β such that:

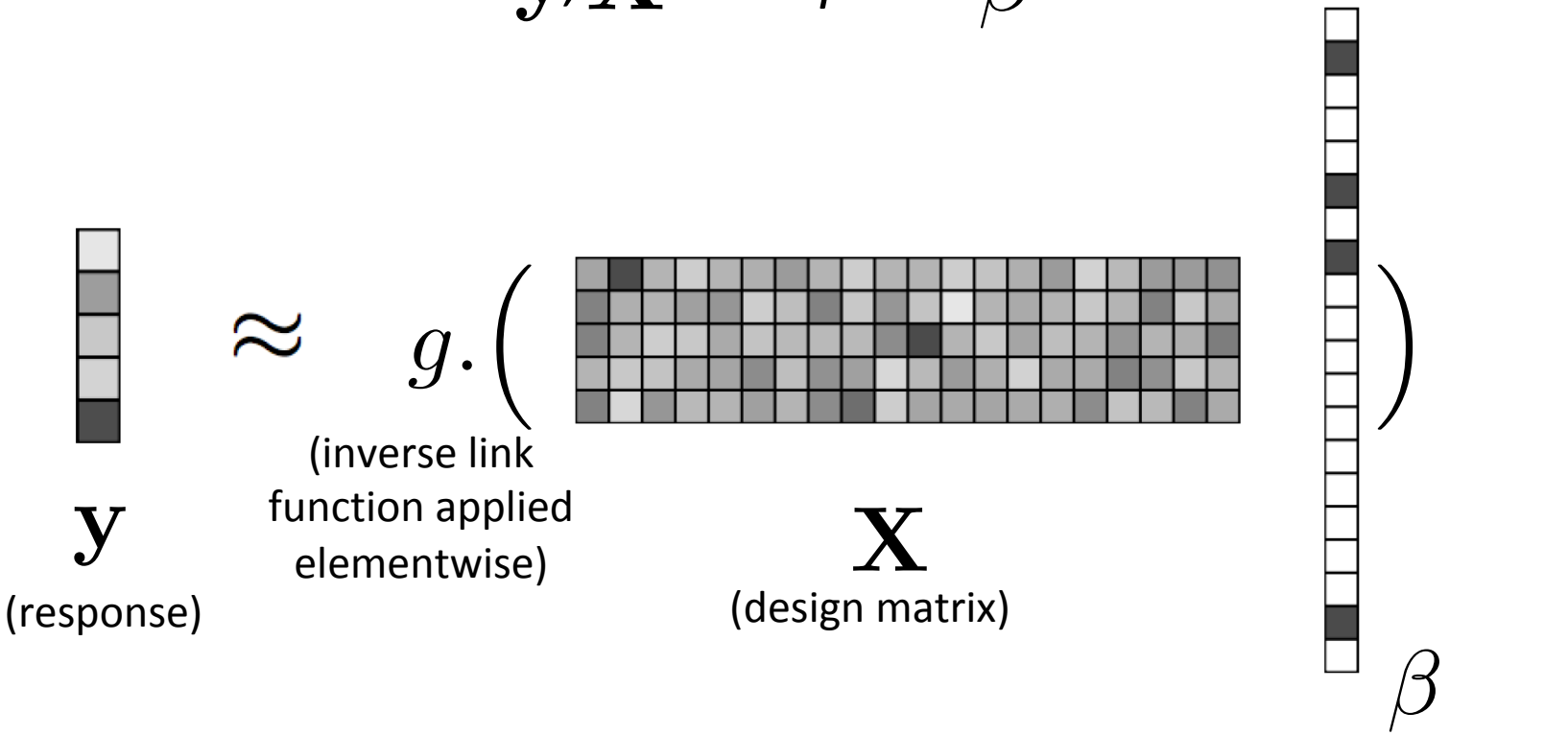


Figure: https://www.cs.cmu.edu/~ggordon/10725-F12/scribes/10725_Lecture2.pdf

IHT vs Lasso vs Marginal testing

	Normal	Logistic	Poisson	Neg Binomial
IHT TF	8.84	6.28	7.2	9.0
IHT FP	0.02	0.1	1.28	0.98
Lasso TP	9.52	8.16	9.28	NA
Lasso FP	31.26	45.76	102.24	NA
Marginal TP	7.18	5.76	9.04 (5.94*)	5.98
Marginal FP	0.06	0.02	1527.9 (0.0*)	0.0

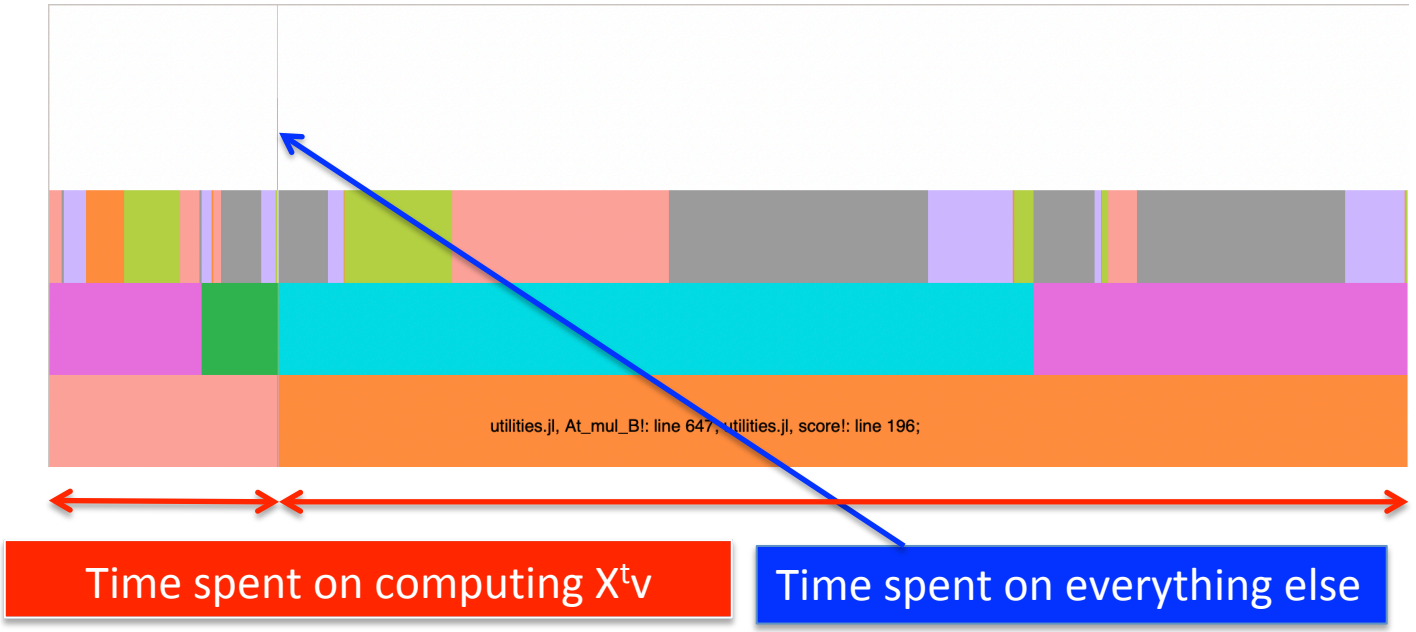
TP = true positives (higher means good)
FP = false positives (lower means good)
* = zero-inflated Poisson regression

Other methods for high dimensional GLM: Elastic net, MCP, Matching pursuit, stability selection, forward stepwise regression…

See our paper: https://www.biorxiv.org/content/10.1101/697755v1

Bottleneck of IHT is gradient computation

IHT iterates according to: $\beta_{n+1} = P_{S_k}\left(\beta_n - s_n \nabla f(\beta_n)\right)$



5k by
30k SNP
Matrix

Question: Can mailman algorithm (which appears to require a lot of preprocessing) help us?

Project 2: Genotype imputation via haplotype reference panels

Problem: Fill missing entries of the inference panel

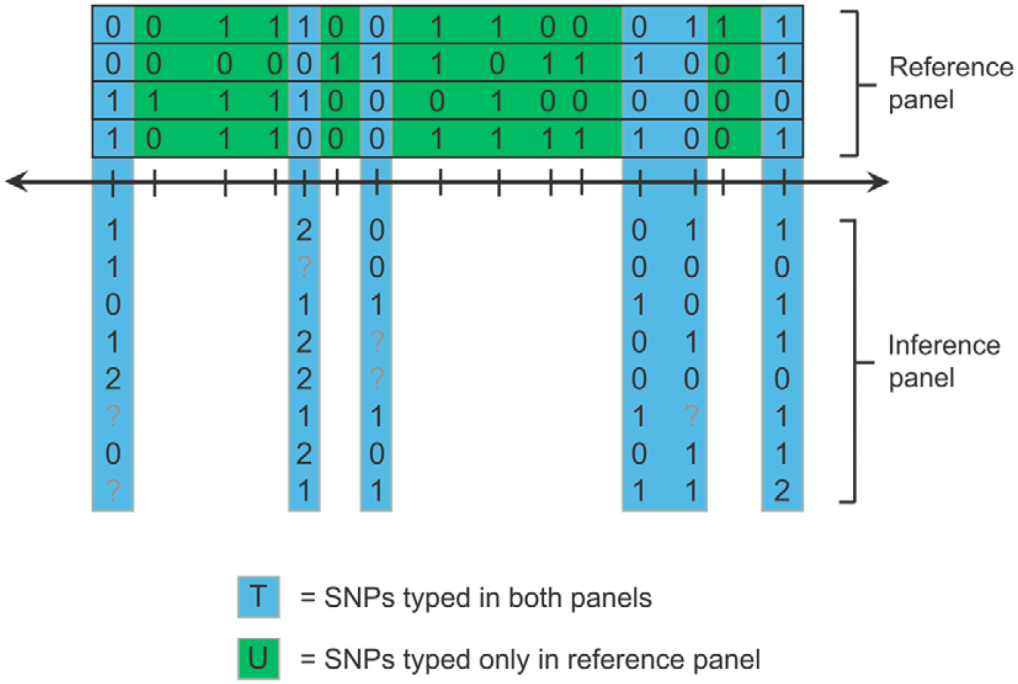


Figure: Howie, Bryan N., Peter Donnelly, and Jonathan Marchini. "A flexible and accurate genotype imputation method for the next generation of genome-wide association studies." PLoS genetics 5.6 (2009): e1000529.

Our method is based on least squares

Our competitors (Minimac4, Impute5, Beagle 5) all use Hidden Markov Models derived in (*)

Preliminary results:

		Java	C++	Julia
	Beagle 5.0 (phased)	**Beagle 5.0 (unphased)**	**Minimac4 (phased)**	**MendelImpute (unphased)**
Error rate (%)	2.9	4.7	1.8	0.3
CPU time (sec)	11 (+2525)	1621	297 (+2525)	5.4

- 36,498 SNPs, 660 reference panels, 665 samples as imputation target
- Phasing data (beagle 4.1) adds 2525 seconds
- MendelImpute does not require pre-phasing
- Timing for MendelImpute does not include data import/export

(*) Li, Na, and Matthew Stephens. "Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data." Genetics 165.4 (2003): 2213-2233.