

A MULTIPLE REGRESSION APPROACH FOR GWAS AND HIGH DIMENSIONAL INFERENCE

Benjamin Chu

Computational Medicine, Biomathematics

1/31/2020

OpenMendel platform for Statistical Genetics

Visit us: github.com/OpenMendel

OPENMENDEL Option	Description
MendelAimSelection.jl	Selects the most informative SNPs for predicting ancestry
MendelEstimateFrequencies.jl	Estimates allele frequencies from pedigree data
MendelGameteCompetition.jl	Tests for association under the gamete competition model
MendelGeneticCounseling.jl	Computes risks in genetic counseling problems
MendelGWAS.jl	Tests for association in genome-wide data
MendelIHT.jl	GWAS using Iterative Hard Thresholding (forthcoming)
MendelImpute.jl	Genotype imputation (forthcoming)
MendelKinship.jl	Computes kinship and other identity coefficients
MendelLocationScores.jl	Maps a trait via the method of location scores

Our focus today

MendelGeneDropping.jl	Simulates genotypes based on pedigrees
MendelSearch.jl	Optimization routines
MendelTraitSimulate.jl	Trait simulation using GLM and GLMM (forthcoming)
SnArrays.jl	Utilities for handling compressed storage of biallelic SNP data
VCFTools.jl	Utilities for handling compressed storage of sequence data
VarianceComponentModels.jl	Utilities for fitting and testing variance components models

Outline

- How simple scientific questions lead to a “high dimensional” data
- Pros and cons of “p-value methods” vs multiple regression methods.
- Maybe convince you to use Iterative hard thresholding for your high dimensional data

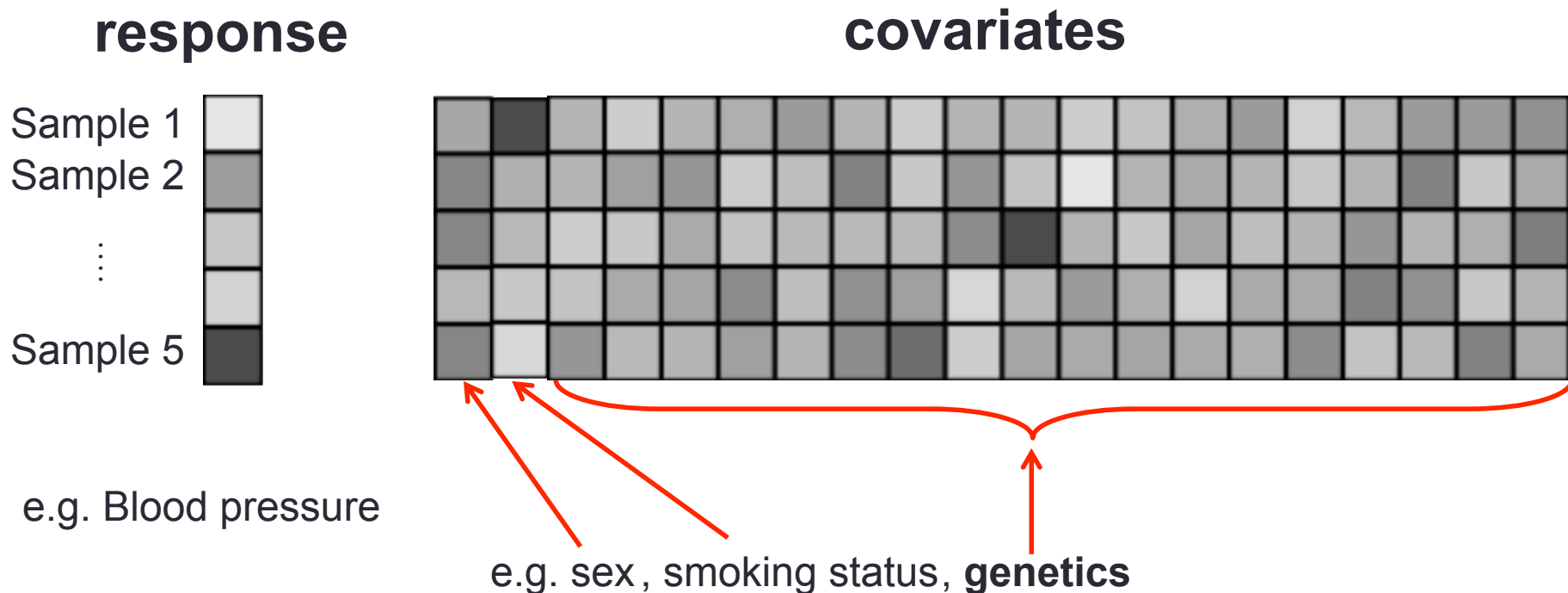
Package: github.com/biona001/MendellHT.jl

Slides: github.com/biona001/public-talks

Paper: <https://www.biorxiv.org/content/10.1101/697755v2>

What is high dimensional data?

- Dataset with more **covariates** than samples.
 - Covariates = independent variables, features, predictors, regressors...



Examples

- **Imaging:** 1000 images, each with 1 million pixel
- **Netflix:** 200,000 users and ~500,000 movies (as of 2010)
- **Genome wide association study (GWAS):** 500,000 samples and 10+ million single nucleotide polymorphisms (SNPs)

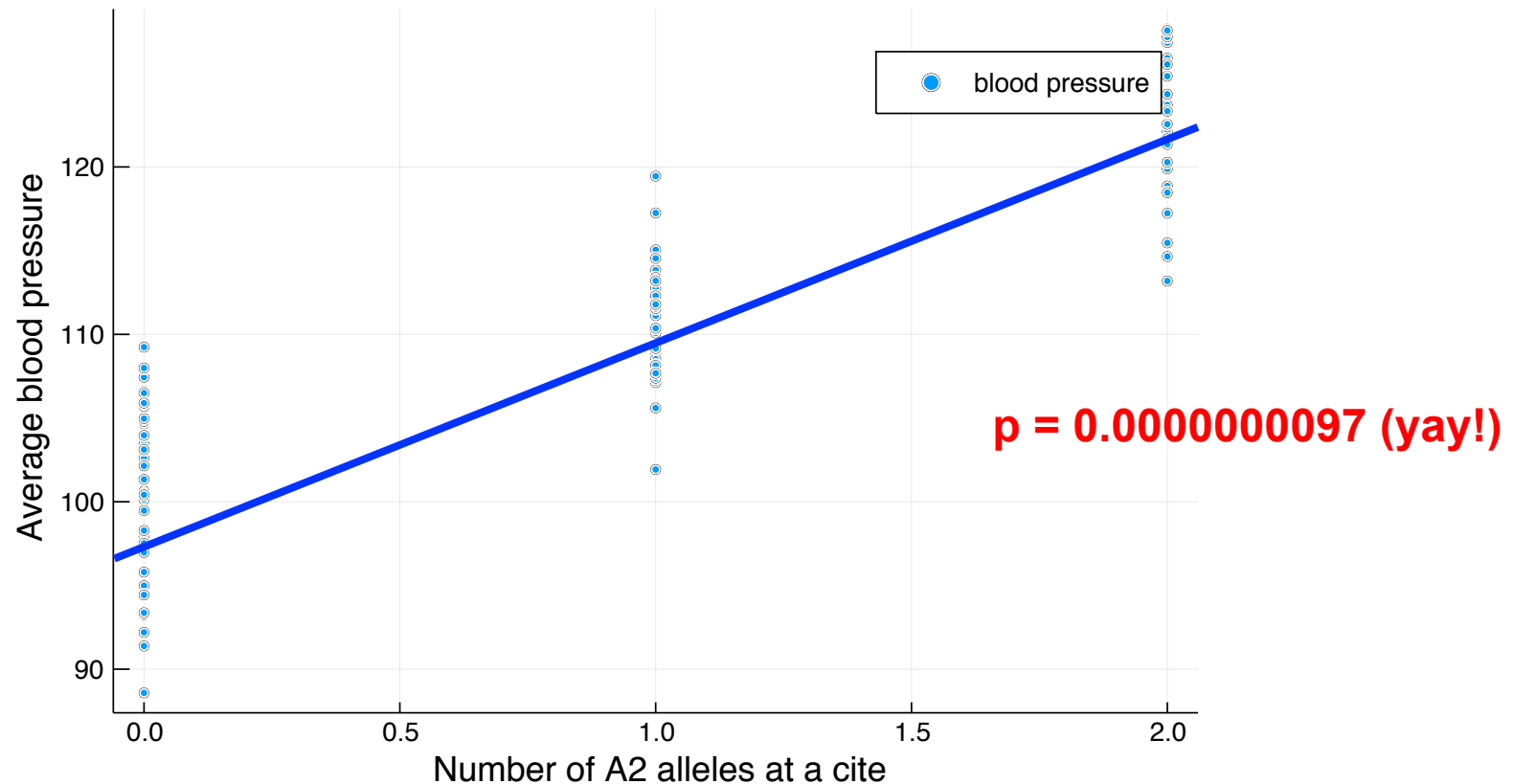
To learn anything from these data, we need to:

1. Identify which covariates are causal
2. Figure out *how much* of an effect they exert



Find causal covariates by p-values (cont.)

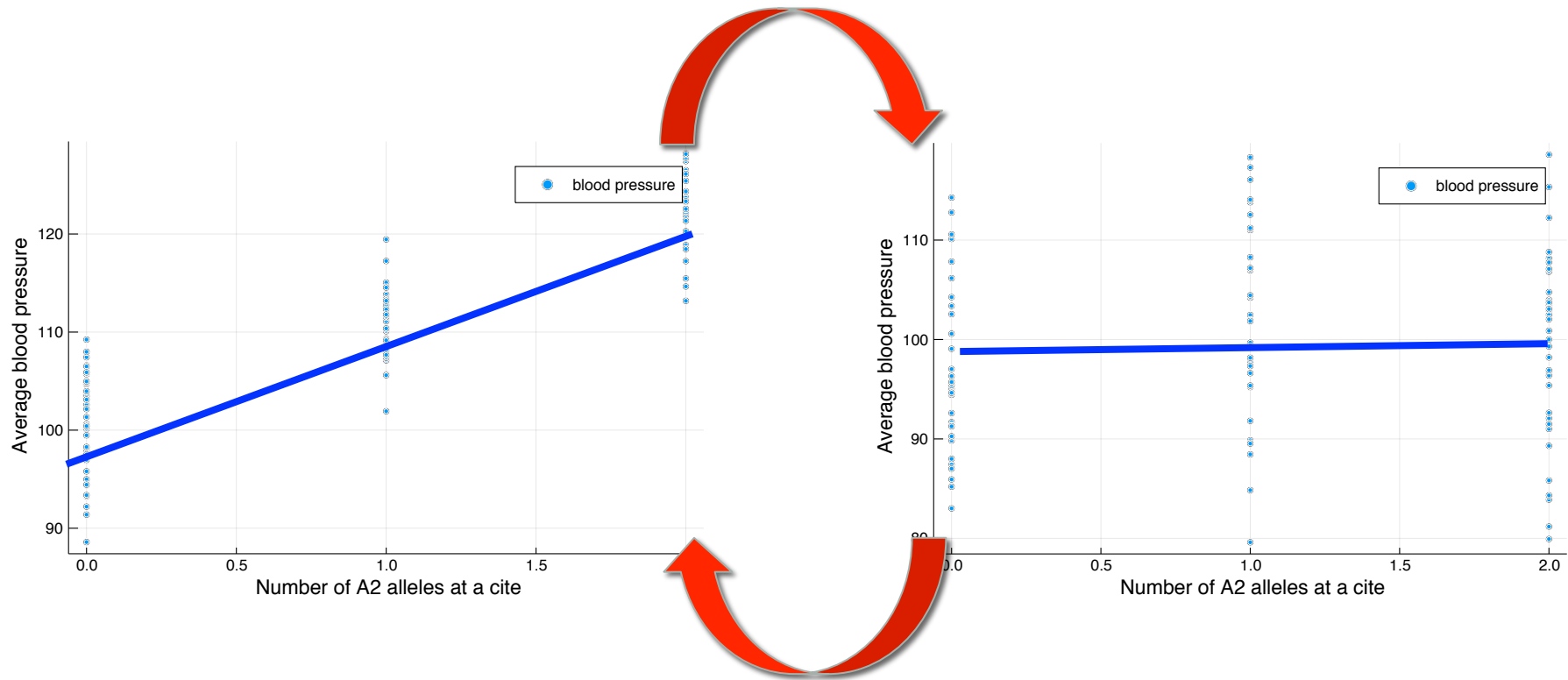
Given , fit a (separate) linear regression **for each** covariate



Pros and cons of p-values for GWAS

- Pros
 - Simple to compute (least squares or likelihood ratio tests)
 - Easy to interpret (low p value = good)
 - Low computational complexity (1 pass for each SNP)
 - Low memory requirement (since only genotype vectors can be loaded one by one)
- Cons:
 - Assumes all covariates have **independent effects**. **Thus does not control for confounders.**

What if you don't control for confounders?



Adjusting for confounder x2 can reverse significance

Multiple regression control for confounding

Ideally, every individual is **identical** except 1 thing. Multiple regression captures this effect.

$$\text{Model 1: blood pressure} = \beta_0 + \begin{array}{|c|} \hline \text{ } \\ \hline \end{array} \beta_1 + e$$

$$\text{Model 2: blood pressure} = \beta_0 + \begin{array}{|c|} \hline \text{ } \\ \hline \end{array} \beta_1 + x_2\beta_2 + e$$

Message of the talk: Multiple regressions measure covariate effects **as if other covariates are fixed**, solving the problem of ***unbalanced data***

Did people attempt multiple regression GWAS?

2009: lasso

Genome-wide association analysis by lasso penalized logistic regression

Tong Tong Wu¹, Yi Fang Chen², Trevor Hastie^{2,3}, Eric Sobel⁴ and Kenneth Lange^{4,5,*}

2011: stability
selection lasso


Stability Selection for Genome-Wide Association

David H. Alexander^{1*} and Kenneth Lange²

- **Their conclusion: less power than marginal association**
- We propose IHT as new multiple regression method for GWAS, but we need to understand why lasso didn't work.

Case study: Lasso vs Marginal testing

	Quantitative trait	Binary trait
Lasso TP	9.52	8.16
Lasso FP	31.26	45.76
Marginal TP	7.18	5.76
Marginal FP	0.06	0.02

 = true positives (higher means good)

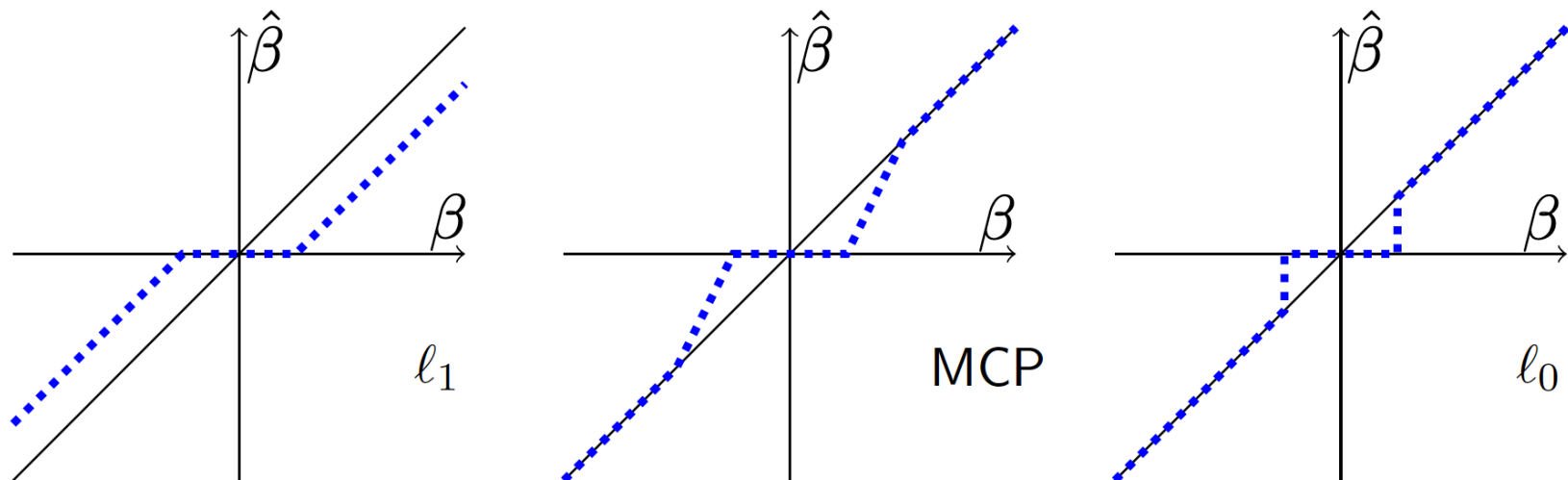
 = false positives (lower means good)

Simulation with 10 causal SNPs.

Conclusion: False positive rate very high for lasso.

Lasso's high false positives are due to shrinkage

Lasso's shrinkage leaves much trait variance unexplained, which are filled by false positives.



Lasso (left): All parameters are shrunk towards 0 (biased for all)
 MCP (mid): Unbiased for *large* effects but biased for small effects.
 IHT (right): Unbiased for all non-zero effects

Case study: parameter estimates in IHT vs lasso

Normal (quantitative trait)

True beta	0.5	0.25	0.1	0.05	0.03
IHT	0.499	0.25	0.096	0.062	0.057
lasso	0.448	0.199	0.045	0.01	0.008

Logistic (binary trait)

True beta	0.5	0.25	0.1	0.05	0.03
IHT	0.504	0.256	0.128	0.114	NA
lasso	0.379	0.142	0.02	0.01	NA

Both simulations have 5k samples, 10k SNPs, run 100 replicates.

Conclusion: IHT is better at estimating parameters

Case study: IHT vs Lasso vs Marginal testing

	Normal	Logistic	Poisson	Neg Binomial
IHT TP	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



= true positives (higher means good)



= false positives (lower means good)

Simulation with 10 causal SNPs.

Our contribution!

Conclusion: IHT is better than lasso and marginal testing.

What are our contributions?

- Extend IHT to generalized linear models (previous slide)
- Extend IHT to double sparsity: limited groups and limited number of SNPs per group
- Enable prior weighting in IHT
- **Efficient IHT implementation**
 - Handles ~200k samples and ~500k SNPs (PLINK)
 - ~50k samples and 100k dosages (VCF, general matrices)

Software showcase!

- Code open source: <https://github.com/OpenMendel/MendelIHT.jl>
 - Tested on Mac, Linux, Windows
- Paper: <https://www.biorxiv.org/content/10.1101/697755v2>
- Code for reproducing all figures in our paper:
<https://github.com/OpenMendel/MendelIHT.jl/tree/master/figures>
- Documentation: <https://openmendel.github.io/MendelIHT.jl/latest/>