A Fast and Scalable Framework for Large-scale and Ultrahigh-dimensional Sparse Regression with Application to the UK Biobank

https://www.biorxiv.org/content/10.1101/630079v2.full.pdf

Jodie Lord
Journal Club | October 2020

PREMISE FOR PAPER

- Biobanks = huge data sources. Massive potential.
  BUT… **ultra-high dimensional**

- Shape of data drives modelling decision…

- "BIG"

Var_1	Var_2	Var_3	Var_4	Var_5	Var_6	…	Var_50000

Feature selection
e.g. Lasso

- "BIG"

Machine Learning

Samp_1
Samp_2
Samp_3
Samp_4
Samp_5
Samp_6
…
Samp_5000000

PREMISE FOR PAPER

- Biobanks = huge data sources. Massive potential.
  BUT… **ultra-high dimensional**

- Shape of data drives modelling decision…

- "BIG"

- "BIG"

Var_1	Var_2	Var_3	Var_4	Var_5	Var_6	…	Var_50000

Feature selection
e.g. Lasso

Machine Learning

Samp_1
Samp_2
Samp_3
Samp_4
Samp_5
Samp_6
…
Samp_5000000

REGULARIZATION

- Constraints applied to shrink regression coefficients

  ↑ Model stability ↓ Overfitting

LASSO

- L1 penalty
- Shrinks **to** 0
- Variable selection
- Sum of coefficients = penalty

$$\lambda \sum_{j=1}^{m} |\hat{\beta}_j|$$

Tuning parameter

RIDGE

- L2 penalty
- Shrinks **towards** 0
- Multicollinearity → grouping
- SS coef = forms penalty

$$\lambda \sum_{j=1}^{m} \hat{\beta}_j^2$$

ELASTIC NET

- Alpha parameter
- a=0: Ridge
- a=1: Lasso
- 0<a<1: Combination

$$\lambda \left(\frac{1-\alpha}{} \sum_{j=1}^{m} \hat{\beta}_j^2 + \frac{\alpha}{} \sum_{j=1}^{m} |\hat{\beta}_j| \right)$$

Tuning Ridge Lasso

REGULARIZATION

- Constraints applied to shrink regression coefficients

  ↑ Model stability ↓ Overfitting

LASSO

- L1 penalty
- Shrinks **to** 0
- Variable selection
- Sum of coefficients = penalty

$$\lambda \sum_{j=1}^{m} |\hat{\beta}_j|$$

Tuning parameter

RIDGE

- L2 penalty
- Shrinks **towards** 0
- Multicollinearity → grouping
- SS coef = forms penalty

$$\lambda \sum_{j=1}^{m} \hat{\beta}_j^2$$

ELASTIC NET

- Alpha parameter
- a=0: Ridge
- a=1: Lasso
- 0<a<1: Combination

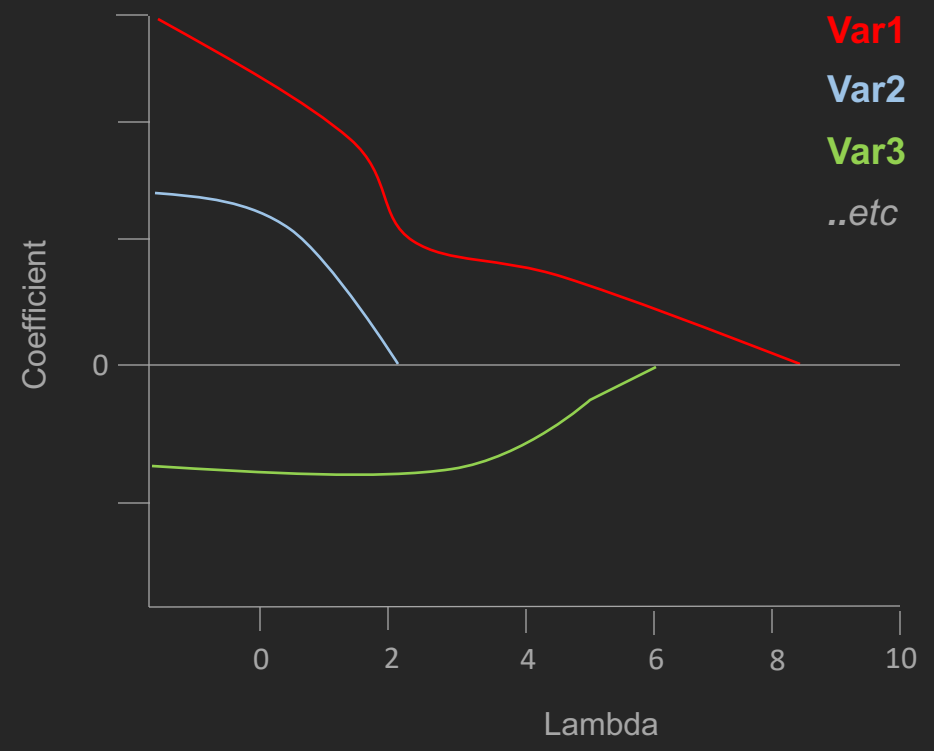
$$\lambda \left(\underline{1-\alpha} \sum_{j=1}^{m} \hat{\beta}_j^2 + \underline{\alpha} \sum_{j=1}^{m} |\hat{\beta}_j| \right)$$

Tuning Ridge Lasso

GLMNET

- Existing package in R, developed by Hastie and co.
- Computes Lasso coefficient paths.
  (optimal lambda unknown → compute over a grid of lambda values)
- Cross validation to select optimal lambda.
- So this is not a new idea…

  BUT… scaling remains an issue



RELEVANCE TO GENOMIC PREDICTION

- Genetic variance currently explained by GWAS

RELEVANCE TO GENOMIC PREDICTION

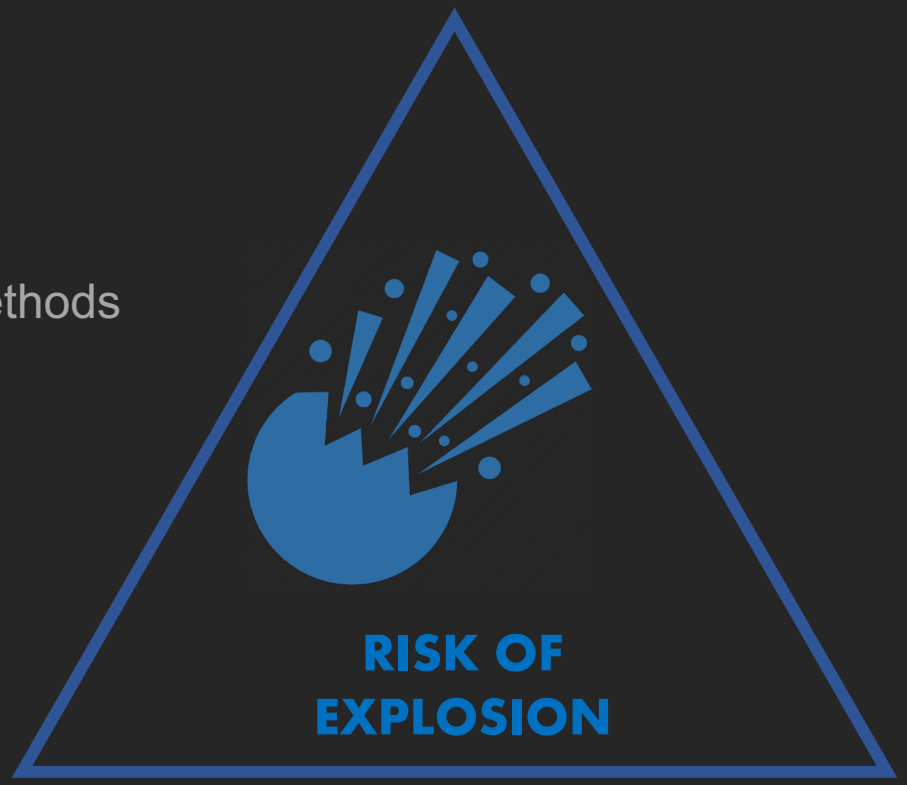
biobank uk
Improving the health of future generations

- 500,000 samples → more predictive power
- Limitations of univariate approaches

RELEVANCE TO GENOMIC PREDICTION

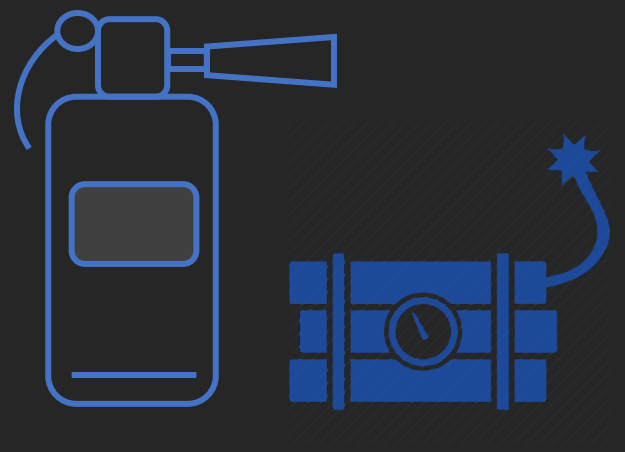
biobank uk
Improving the health of future generations

- Full-scale multivariable methods



SNPNET

- Extension of GLMNET
- **BA**tch **S**creening **I**terative **L**asso → BASIL
- R package → easy acquisition
- Full-scale multivariable methods



METHODS

BASIL

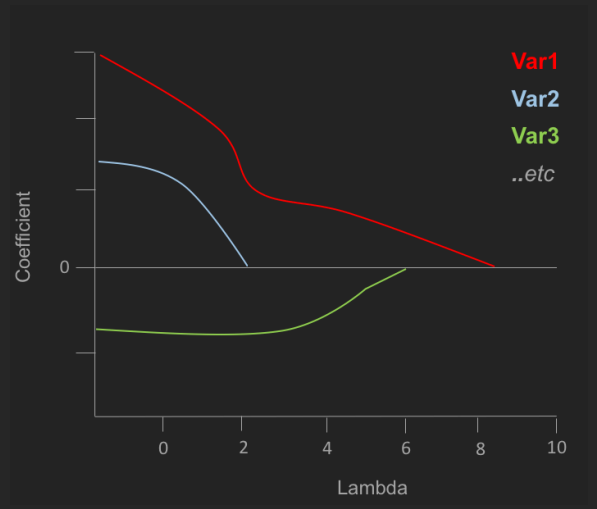
Screen → solve → check

- Screen: Strong set
  Inner product at $\lambda l > \lambda l_{+1}$-($\lambda l$- λl_{+1})
- Solve: Computed strong set
  Compute lambda solution only on these subset
- Check: Entire variable matrix
  Evaluate against KTT condition (🤷‍♀️)

Not new – all done in GLMNET

- **BATCH** screening of
  multiple lambdas → single whole data check

Specific to SNPNET



EXTRACT FROM PAPER

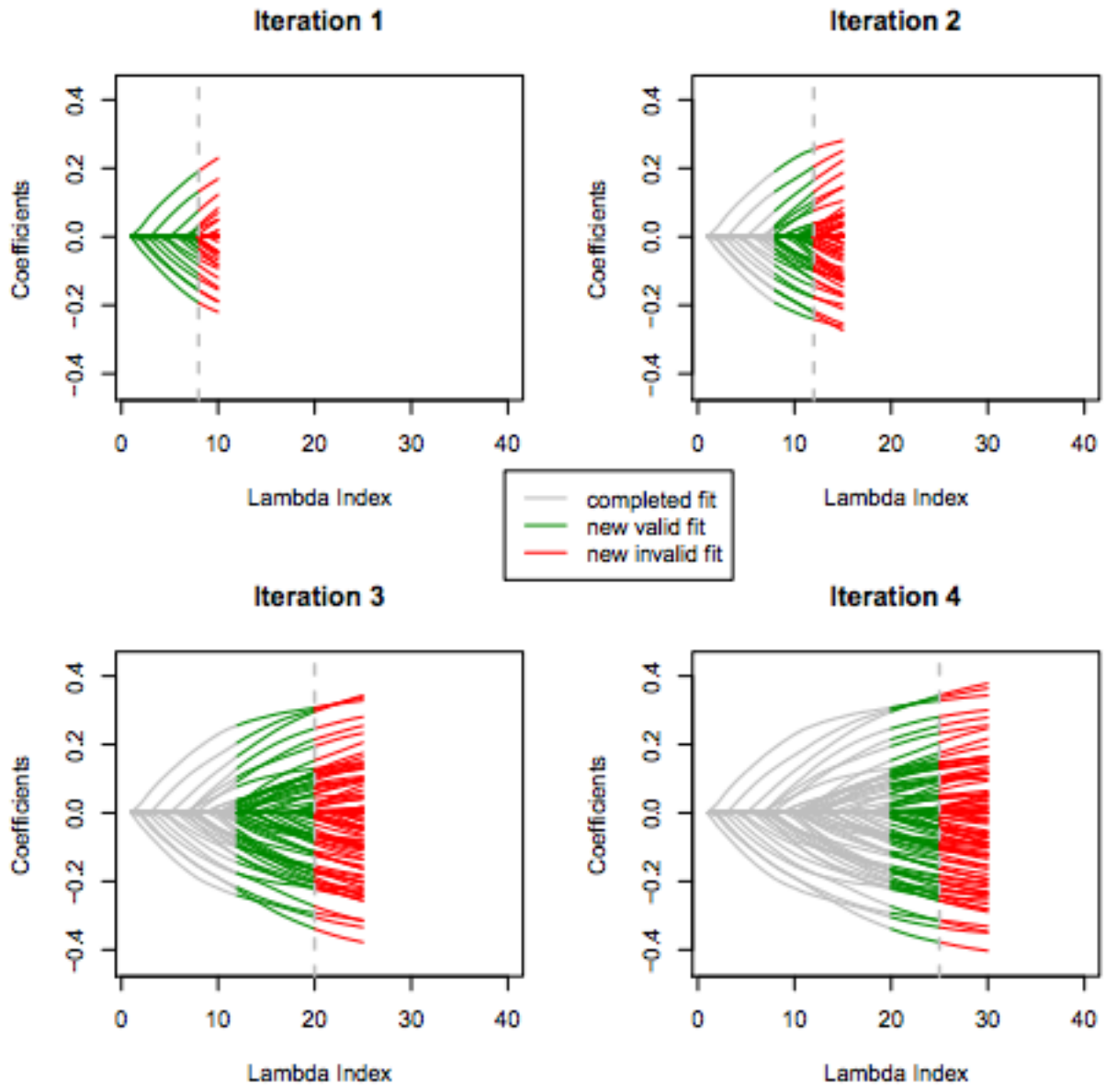


Figure 1: The lasso coefficient profile that shows the progression of the BASIL algorithm. The previously finished part of the path is colored grey, the newly completed and verified is in green, and the part that is newly computed but failed the verification is colored red.

BASIL

- Linear
- Logistic
- Survival
- Elastic net

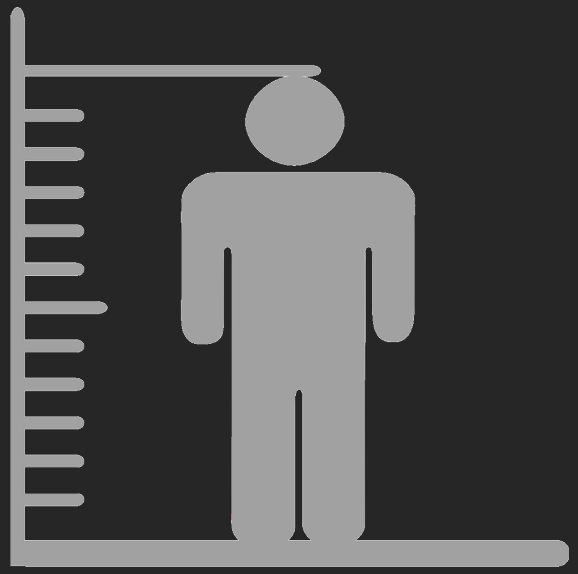
TESTING ON UKBB DATA

- 337,199 unrelated individuals (of the 500,000 in UKBB)
- Training | Validation | Test
- 805,426 measured variants
- Covariates = age, sex, principal components (40? 10?)

TESTING ON UKBB DATA

- Continuous phenotypes → R2

1.



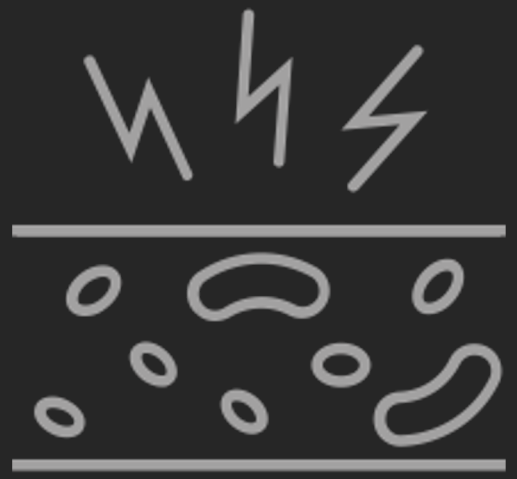
2.

TESTING ON UKBB DATA

- Binary phenotypes → AUC

1.

2.



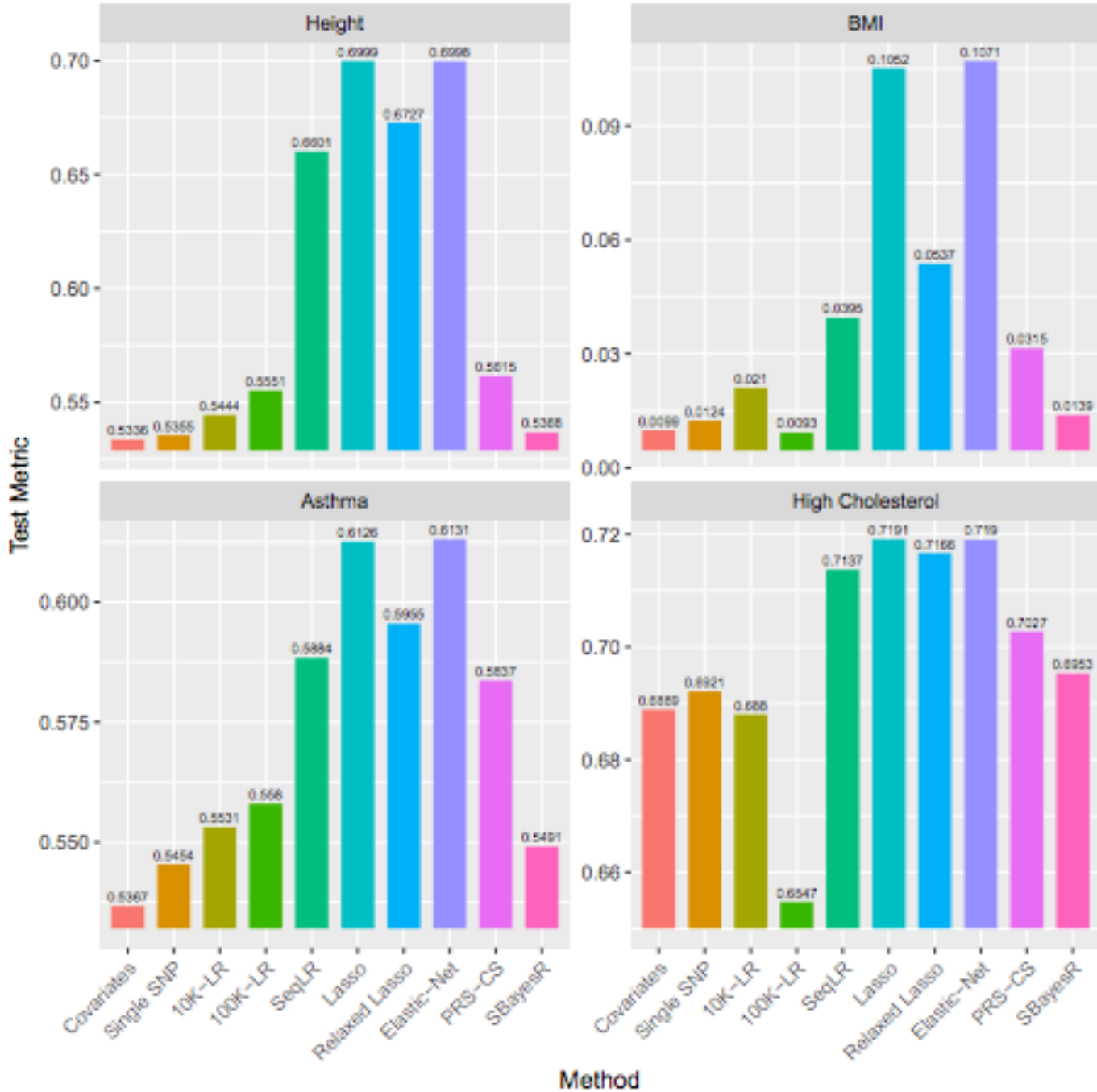
COMPARATIVE MODELS

Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10
Baseline	Covariate only	Single SNP	Relaxed threshold 10k	Relaxed threshold 100k	Sequential	PRS-CS	SBayesR	Elastic net	Relaxed LASSO
Sex Age	Sex Age 10 PCs	Strongest univariate +model 2	Single var from linear combinations using univariate coefficients +model 2		Multiple regression - sequential increase SNP N	Default settings 🤷‍♀️		α = 0.1 α = 0.5 α = 0.9	OLS variables selected for the LASSO

Most similar to PRSice

EXTRACT FROM PAPER

Results on test-sample set



R2

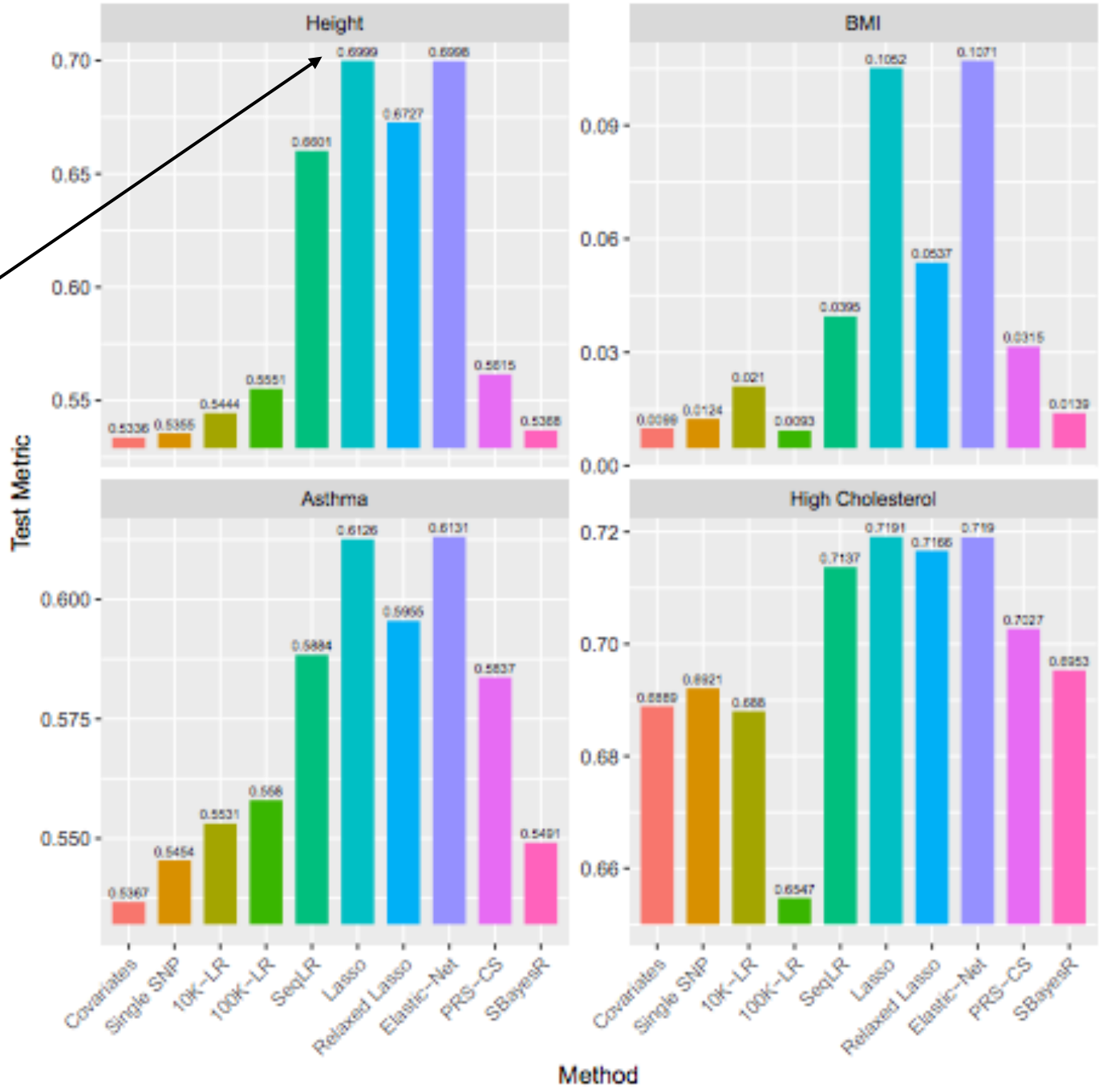
AUC

RESULTS

EXTRACT FROM PAPER

Results on test-sample set

R2=0.6999…



R2

AUC

STRENGTHS

- Thorough methodological overview
- Builds on existing method → package readily available
- Examples using real, accessible, relevant data
- Comparisons with potential alternatives
- Continuous and binary outcomes demonstrated

POTENTIAL LIMITATIONS

- Not fully clear whether individual level data required
- Examples based on split datasets…
  Cross validation?
- Model fit unusually large? Covariate effect?
- Linkage disequilibrium?
- Cherry picking traits?

DISCUSSION

YOUR THOUGHTS?