P16.018.A

A three-level modeling for identifying important predictor variables in genome-wide association studies suffering from p >> n

Jakub Liu, Dawid Słomian, Paula Dobosz, Joanna Szyda

Jliu@ump.edu.pl

Biostatistics Group, Department of Genetics, Wroclaw University of Environmental and Life Sciences, Wroclaw, Poland

National Institute of Animal Breeding, Cracow, Poland

University Cancer Diagnostic Center, Poznan University of Medical Sciences, Poznan, Poland

Conclusions

- **Feature preselection based on submodels** → effective way to circumvent p >> n problem
- **Mixed linear model** → feasible for high dimentional data → shrinkage by N(0,V)
- **Multidimentional data from WGS** → conventional significance replaced by estimate clustering

Motivation

- **modern datasets**
  - # predictors (p) >> # observations (n)
- **problems:**
  - mathematical untrainable models
  - low explainability
  - overly high complexity
  - high computational demands

Feature selection required **prior to** actual modeling and effect estimation.

Aim

implement the feature selection approach for preselection of SNPs in multidimentional data

all predictors → feature selection → only important predictors → further analysis possible

Material

- **1,222 individuals with COVID-19 infection classification → resistant / non-resistant**
- **41,836,187 SNPs from WGS**

Methods

1. preprocessing

- VCF filterd on MAF
- 43 469 928 SNPs before filtering
- 6 949 073 SNPs after filtering (**X**)

vcftools, PLINK, Python

2. M LogReg models

- Each model → subset of SNPs 123
- SNP estimates
- Model performance → deviance
- Model performance matrix
- 62 045 (**M**)

Python

submodel	deviance	SNP1	…	SNP_**X**
1	MP_1	FP1		0.0
…	…	…	…	…
M	MP_**M**	0.0		FP_X

3. final model

- Mixed linear model
- **y = μ + Zb + e**

y model performance MP 1×**M**
μ mean
Z feature performance FP **M**×**X**
b SNP effects 1×**X**

$\boldsymbol{b} \sim N\left(0, \boldsymbol{I}\sigma_b^2\right)$

e residual $\boldsymbol{e} \sim N(0, \boldsymbol{I}\sigma_e^2)$

- Estimation → PCG
- iterations: 75, time: 2.5 days

MIX99 in Fortran

4. important SNPs

- 1D K-means clustering of $\widehat{\boldsymbol{b}}$
- 2 clusters

→ **important (1)**
1,313 SNPs **center -0.03**
→ non-important (0)
6,947,759 SNPs center $6.7 \cdot 10^{-6}$