A Unified Model based MDR framework for detecting gene-gene interactions

Taesung Park[1]

Joint work with

Wenbao Yu[2], **Yongkang Kim**[2], and **Seungyeoun Lee**[3]

[1] Seoul National University, Korea

[2] U of Penn, USA

[3] Sejong University, Korea

Outline

- Introduction
- Method
- Simulation
- Application

MDR Overview

- Method for detecting and characterizing interactions in common complex multifactorial disease (Ritchie et al., 2001)
- Applicable even when sample size is small or dataset contains alleles in LD
- Indicate which alleles or genotypes increase susceptibility (High, Low)

MDR Overview

- For simplicity, assume two SNPs
- $n_{i,j}^{\text{case}}$: frequency of case in (S1=i, S2=j)
- $n_{i,j}^{\text{ctl}}$: frequency of control in (S1=i, S2=j)

- High risk group

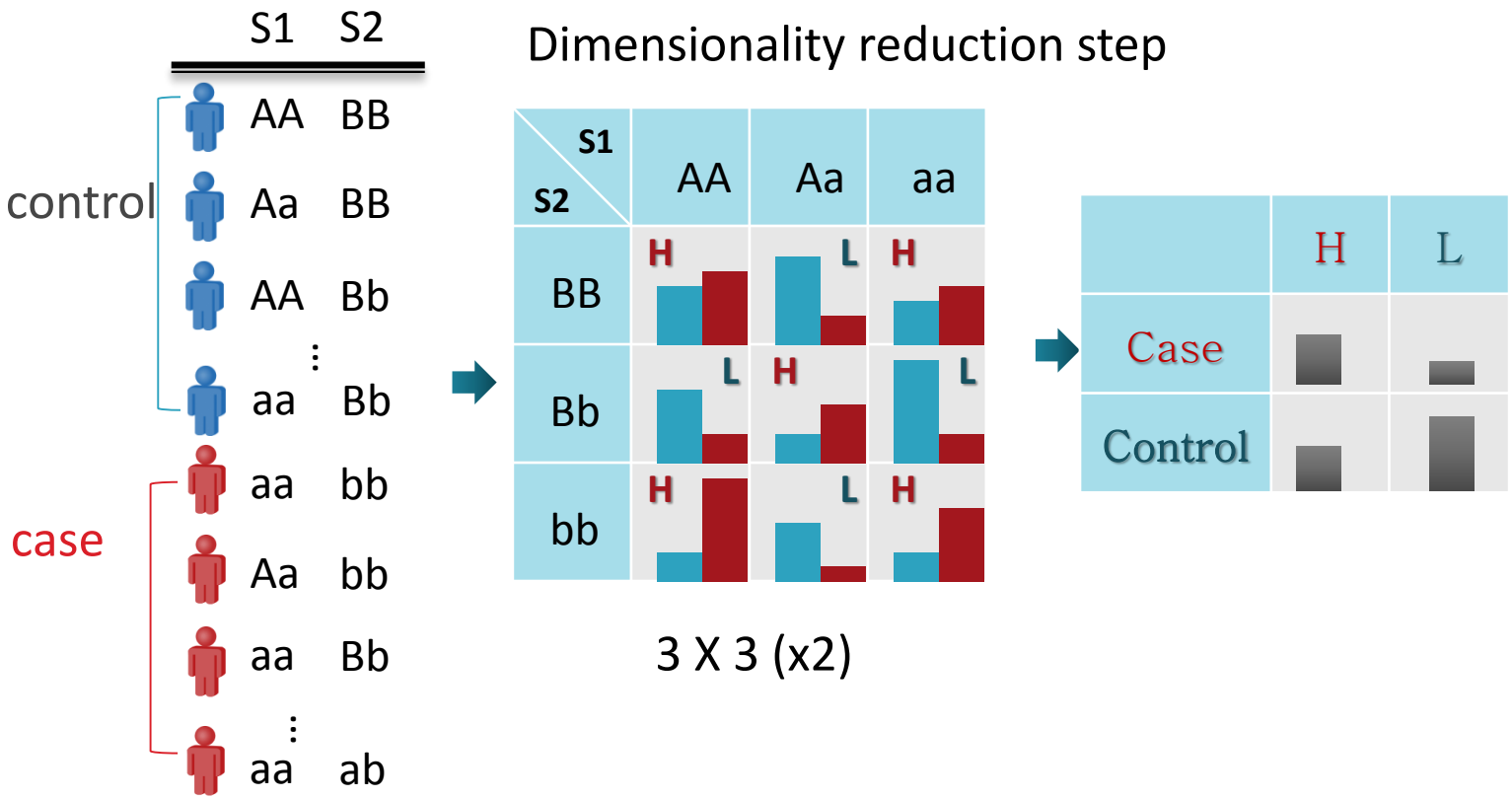
$$\Leftrightarrow \frac{n_{i,j}^{\text{case}}}{n_{i,j}^{\text{ctl}}} \geq \frac{n^{\text{case}}}{n^{\text{ctl}}}$$

- Low risk group

$$\Leftrightarrow \frac{n_{i,j}^{\text{case}}}{n_{i,j}^{\text{ctl}}} < \frac{n^{\text{case}}}{n^{\text{ctl}}}$$

MDR overview

Two-way interactions



MDR Overview

- 10-fold cross validation
- Accuracy
  - Ratio of correct classification to the total number of instances classified
- Balanced accuracy (BA)

$$BA = \left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP} \right) / 2$$

- Cross-validation Consistency (CVC)
  - Number of times that a SNP combination is identified as the best combination across the 10 CV datasets

	H	L
Case	TP	FN
Control	FP	TN

Extensions for MDR

- **Generalization via statistical modeling**
  - Generalized MDR(GMDR) (Lou *et al.* 2007), MB-MDR (Cattaer *et al.* 2011)
  - Odds Ratio MDR (Chung *et al.* 2007), Log-linear MDR (Lee *et al.* 2008)
  - New Measures for MDR (Namkung *et al.* 2010), Ordinal MDR (Kim *et al.* 2013)
  - Gene-based MDR (Oh et al., 2013), Entropy MDR (Kwon, et al, 2014)
- **Family-based data**
  - FAM-MDR (Cattaert *et al.* 2010), PGMDR (Chen *et al.* 2011) and MDR-PDT (Edwards *et al.* 2010)

Extensions of MDR

- **Survival data**
  - Surv-MDR (Gui *et al.* 2011) and Cox-MDR (Lee *et al.* 2012)
- **Quantitative traits**
  - Quantitative MDR (Gui *et al.* 2013)
- **Multi-phenotypes**
  - Multivariate generalized MDR (Choi *et al.* 2013),
  - Multivariate quantitative MDR (Yu *et al.2015*)

Extensions of MDR by Our Group

Odds ratio based multifactor-dimensionality reduction method for detecting gene–gene interactions
Yujin Chung[1], Seung Yeoun Lee[2], Robert C. Elston[3] and Taesung Park[1,*]
Bioinformatics

Log-linear model-based multifactor dimensionality reduction method to detect gene–gene interactions
Seung Yeoun Lee[1], Yujin Chung[2], Robert C. Elston[3], Youngchul Kim[4] and Taesung Park[4,*]
Bioinformatics

New evaluation measures for multifactor dimensionality reductio classifiers in gene–gene interaction analysis
Junghyun Namkung[1,†], Kyunga Kim[2,†], Sungon Yi[2], Wonil Chung[2], Min-Seok Kwon[1] and Taesung Park[1,2,*]
Bioinformatics

Identification of Gene-Gene Interactions in the Presence of Miss Data Using the Multifactor Dimensionality Reduction Methoc
Junghyun Namkung,[1,2] Robert C. Elston,[3] Jun-Mo Yang,[2] and Taesung Park[1,4*]
Genetic Epi

A novel method to identify high order gene-gene interactions in genome-wide association studies: Gene-based MDR
Sohee Oh[1], Jaehoon Lee[1], Min-Seok Kwon[2], Bruce Weir[3],
BMC Bioinformatics

Gene–gene interaction analysis for the survival phenotype based on the Cox model
Seungyeoun Lee[1,*], Min-Seok Kwon[2], Jung Mi Oh[3] and Taesung Park[2,4,*]
Bioinformatics

Identification of multiple gene-gene interactions for ordinal phenotypes
BMC Medical Genomics

Multivariate generalized multifactor dimensionality reduction to detect gene-gene interactions
Jiin Choi[1], Taesung Park[1§]
BMC Systems Biology

IGENT: Efficient Entropy based Algorithm for Genome-wide Gene-Gene Interaction Analysis
Min-Seok Kwon[1], Mira Park[2] and Taesung Park[1, 3§]
BMC Medical Genomics

Software by Our Lab

OR-MDR	Odds ratio based multifactor-dimensioinality reduction method	**R** package
GWAS-MDR	A program for genome-wide association analysis based on multifactor dimensionality reduction	**CPU** based clusters
GWAS-GMDR	A generalized GWAS-Multi that permits adjustment for covariates.	
Ordinal MDR	Multi method for ordinal phenotypes in Gene-Gene interaction analysis	
GPU-G/MDRi	Ultra-high performance G/Multi program based on GPU (graphic processing unit)	**GPU** based system
CuGWAM	A program for visualizing gene-gene interaction in genetic association analysis	

Some drawbacks of MDR based approaches

- It is difficult to measure the significance of a multi-locus model
- Computational burden because of permutation is required for each multi-locus model
- MDR can not distinguish marginal effects from the pure interaction effects

Outline

- Introduction
- Method
- Simulation
- Application

Unified Model-based MDR (UM-MDR) approach

Two-step approach

1. Classification step:

   For each given v-order SNP combination, classify each genotype combination cell into H/L, and let S stands for a cell's H/L status

2. Modeling step:

$$g(\boldsymbol{\mu}) = \alpha_0 + \beta S + \boldsymbol{\gamma}^{\boldsymbol{T}}\boldsymbol{X}$$

   where $\boldsymbol{Y}$ and $\boldsymbol{X}$ stand for the trait and covariates, respectively, and $\boldsymbol{\mu}$ is the mean vector of $\boldsymbol{Y}$ and $g(\cdot)$ is the link function

About classification step

- The classification rule used in this step is flexible
- For a quantitative trait, similar to QMDR, we assign S=H to a cell when the mean value of Y in the cell is greater than the global mean of Y
- For a case/control trait, similar to MDR, we assign S=H to the cell when the ratio of the # cases to the # controls in the cell is greater than the global ratio
- The classification rule for GMDR can also be used for both case/control trait or quantitative trait

Adjustment of marginal effects

- Motivation: There may be a locus, say locus A, that has a strong marginal effect, and some multi-locus models including locus A may be significant just because of locus A

- For a given multi-locus model, for example, SNP1 and SNP2

$$g(\boldsymbol{\mu}) = \alpha_0 + \beta S + \boldsymbol{\gamma}^{\boldsymbol{T}}\boldsymbol{X} + \alpha_1 SNP1 + \alpha_2 SNP2$$

- Penalized regression can be used for handling a multicolinearity problem.
- Similarly, if we want detect pure high-order interactions, we can put lower-order interaction terms in the right side of the model

Advantage

1. Perform significant tests for MDR approaches
2. Cross-validation and permutation in the traditional MDR approaches are not necessary and therefore, the computation cost is significantly reduced
3. The statistical significance of a gene-gene interaction is easily obtained, with adjustment of the covariate effects
4. Test high-order interaction models easily
5. Many existing classification methods can be used to define H/L in the first step

p-value calculation

- We use the Wald type statistic for measure the significance of a multi-loci model.
- What is the null distribution? It may not be a chi-squared distribution.

Why need correction of the p-value (type I error inflation)?

- For a very simple case

$$Y_i = \alpha + \beta S_i + \varepsilon_i$$

- The LSE(MLE) of β is

$$\hat{\beta} = \frac{N}{N_L}(\bar{Y} - \hat{Y}_H),$$

  - $\bar{Y}$ and $\hat{Y}_H$ are the global mean and the mean of H group, respectively
  - N and N_L are the total sample size and sample size of the L group, respectively

- Therefore, to test $\beta = 0$, we actually test $E(Y) = E(Y_H)$
- Note that for QMDR, we classify cell into H if its mean is larger than the global mean, it seems that we have automatically set $\hat{Y}_H > \bar{Y}$ in the first step

Proposed method for correcting the p-value

- Assume the null distribution is a non-central chi-squared distribution
- Estimate the non-central parameter by a few permutation (5-10 times for example)
- Re-calculate the p-value based on the non-central chi-squared distribution

Outline

- Introduction
- Method
- Simulation
- Application

Motivation of simulation

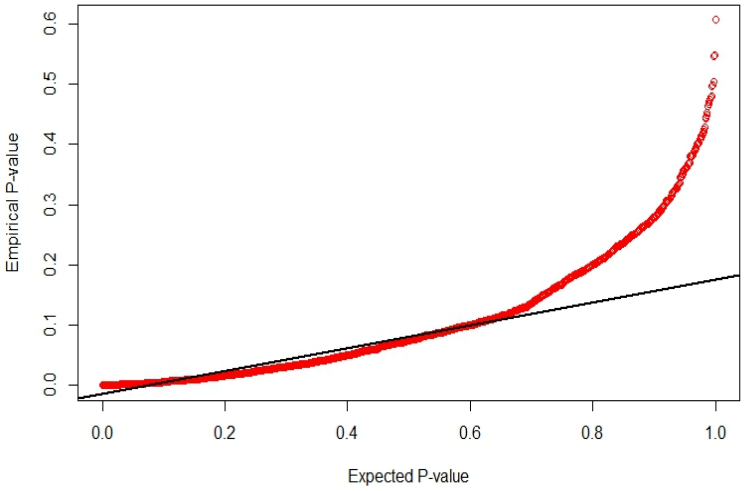
- The motivation of these simulation studies includes:
  - To check whether the proposed approach can control type I error rate
  - To check whether the proposed approach can identify the causal interaction with/without marginal effects
  - To check whether the proposed approach can detect high order interactions

Type I error study

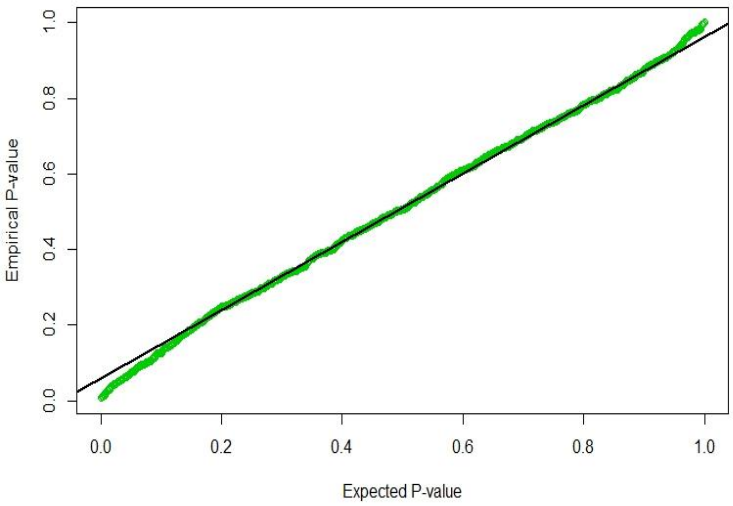
MAF	Binary trait		Quantitative trait	
	Uncorrected	Corrected	Uncorrected	Corrected
0.05	0.13	0.04	0.18	0.04
0.10	0.20	0.03	0.22	0.05
0.20	0.29	0.04	0.48	0.05
0.30	0.48	0.04	0.50	0.01
0.40	0.60	0.03	0.60	0.02

1. N=1000, nominal size = 0.05, and traits randomly generated; No LD, 2 SNPs
2. Uncorrected and Corrected correspond to the results of using uncorrected and corrected p-values

QQ-plot (under the null)



Uncorrected



Corrected

Simulation I—no marginal effects (binary trait)

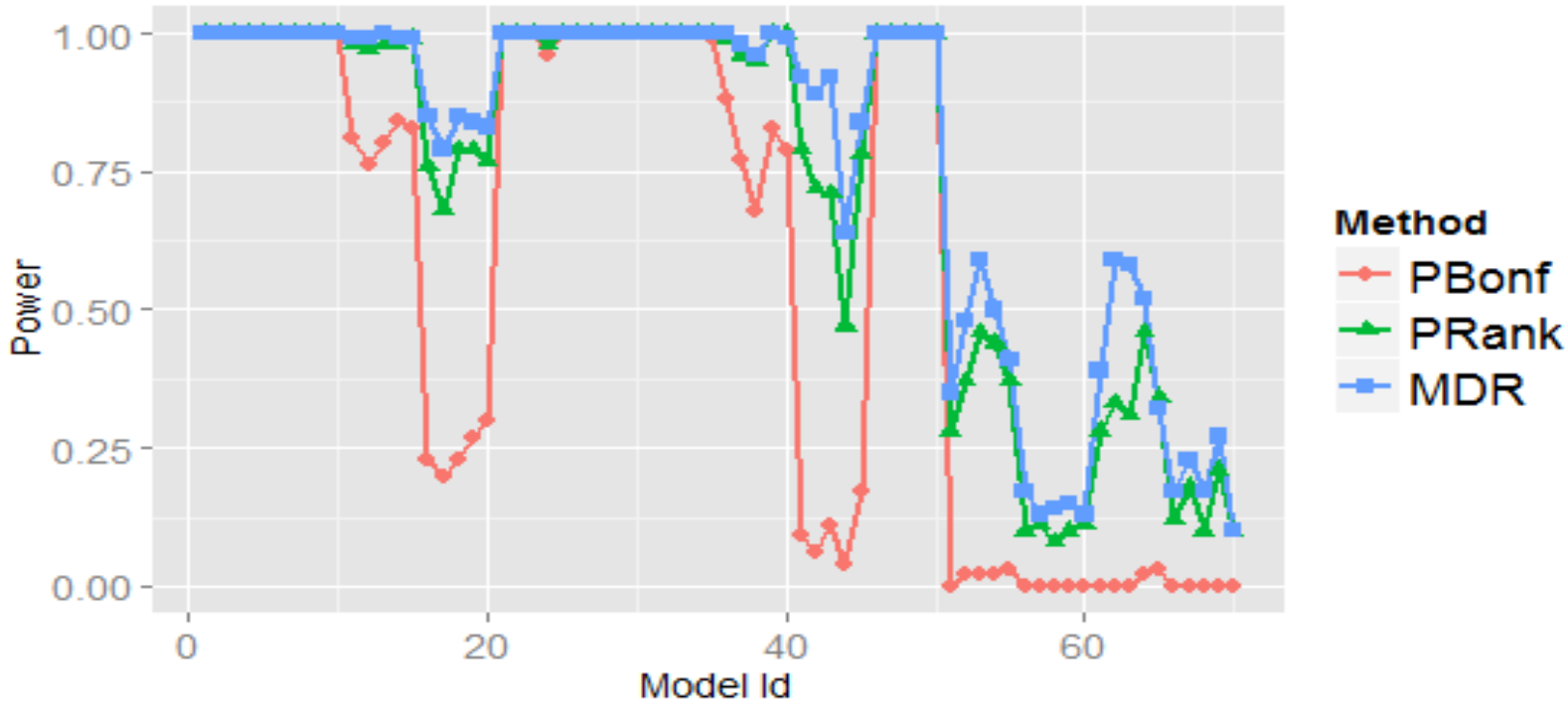
- Setting:

1. Sample size N=1000
2. 20 SNPs, S1-S2, the first two are causal interaction, no LD; binary trait
3. Study the performance on 70 penetrance models (*Velez et al. 2007*)
4. 100 data sets for each model

Power definition

1. The power of UM-MDR defined as the rate of the corrected p-value (after Bonferroni correction) of the causal model being smaller than a nominal size, say 0.05, ---denote as ***PBonf***
2. Such definition of power is different from the original definition for MDR's, which is the detection rate of the causal model being the best model.
3. To compare power fairly, we define the power of UM-MDR as the rate of the causal model being ranked 1st by the corrected p-value --- note as ***PRank***

Power comparison



UM-MDR with PRank achieves similar powers for most 70 models as MDR for binary trait

Simulation II—no marginal effects (quantitative trait)

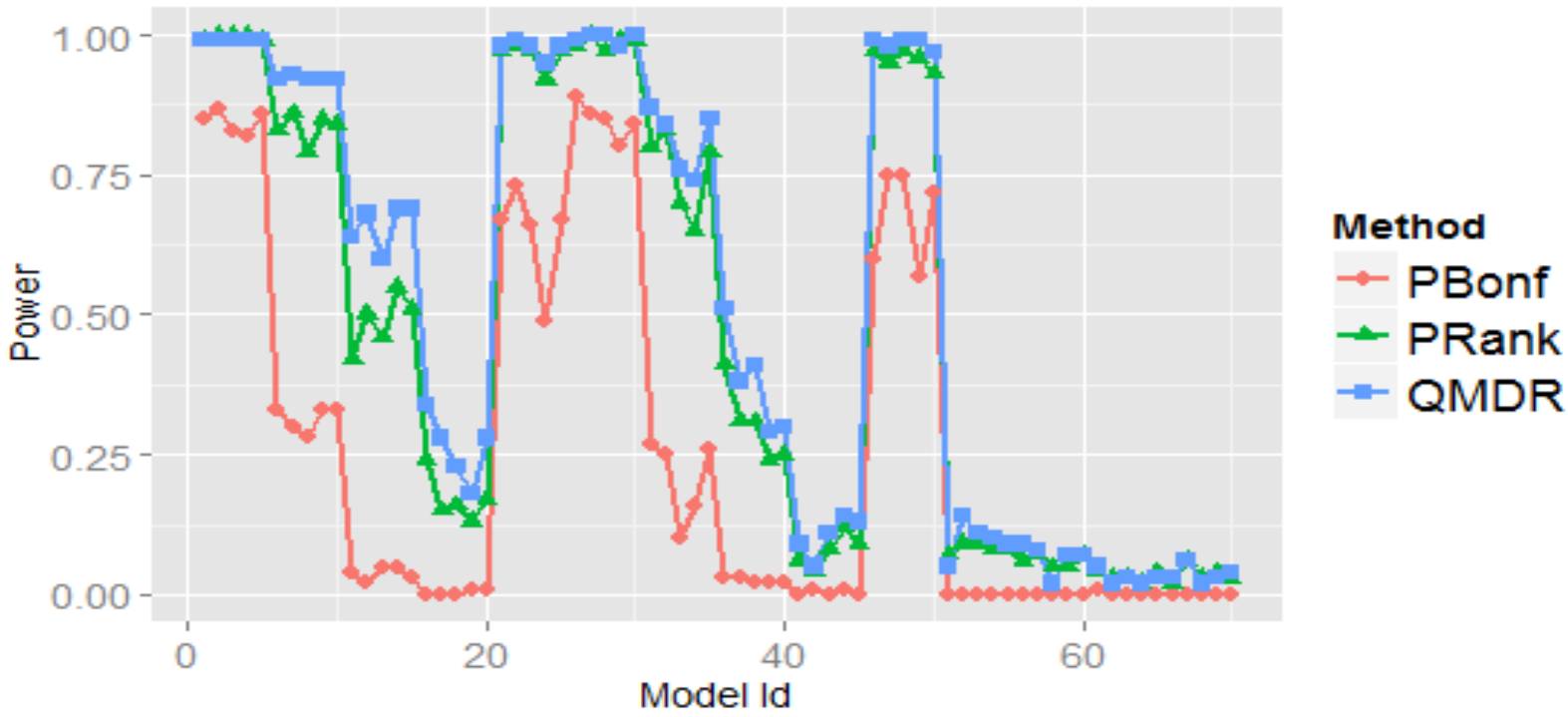
- Setting:

1. The same setting as simulation I, except
2. a single quantitative trait,

$$y|S1 = i, S2 = j \sim N(\mu f_{ij}, 1)$$

3. $f_{ij} = P(high\ risk|S1 = i, S2 = j)$ is the given penetrance function (the 70 models from *Velez et al. 2007*) and μ is the effect size (default is 1)

Power comparison

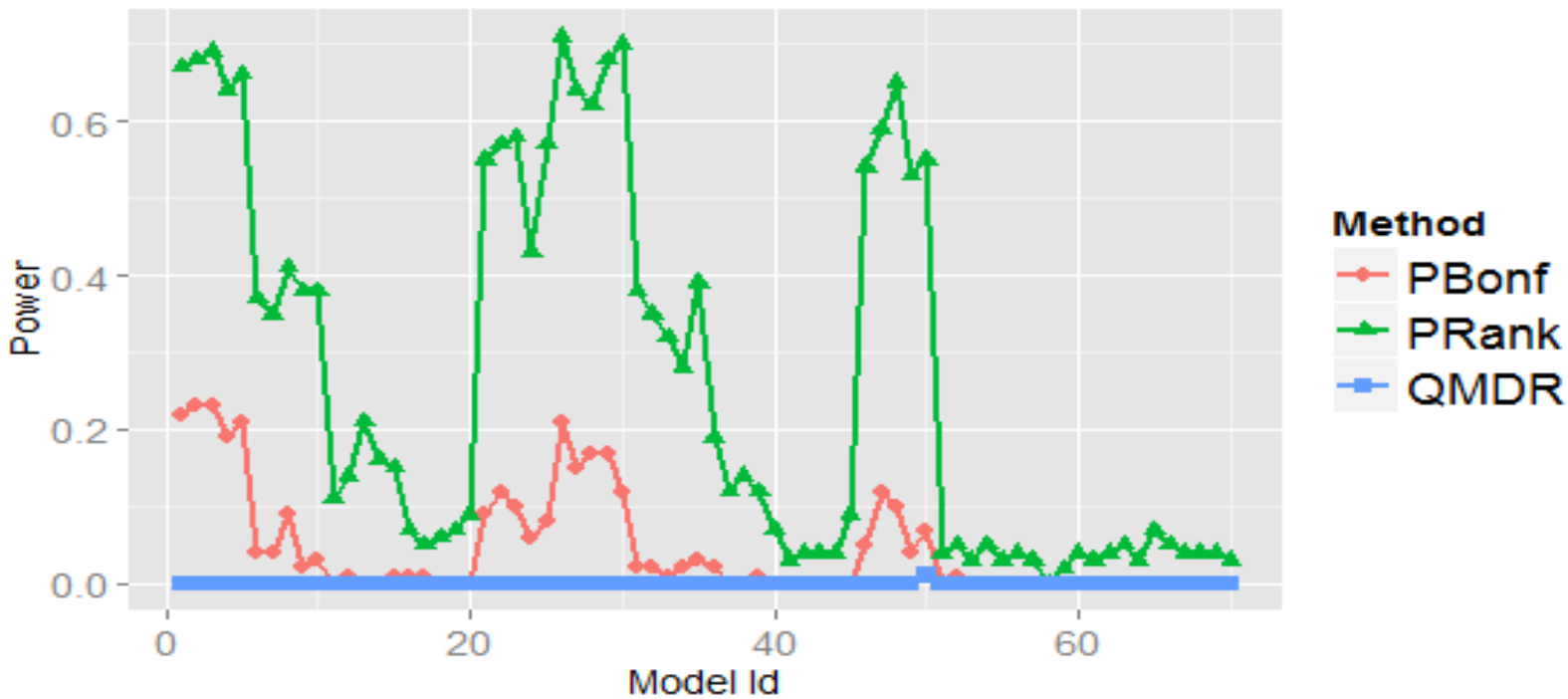


Similar pattern has been achieved as in binary trait

Simulation III—with marginal effects (quantitative trait)

- The same setting as simulation II, except adding marginal effect for S3
- $y|S1 = i, S2 = j \sim N(\mu f_{ij}, 1) \ + N(\alpha\, S3, 1)$
- This simulation is to check whether our approach can avoid detecting the two-locus models (S3, other)
- $\alpha = 1.0$
- Adjust a marginal effect for the proposed approaches

Power comparison

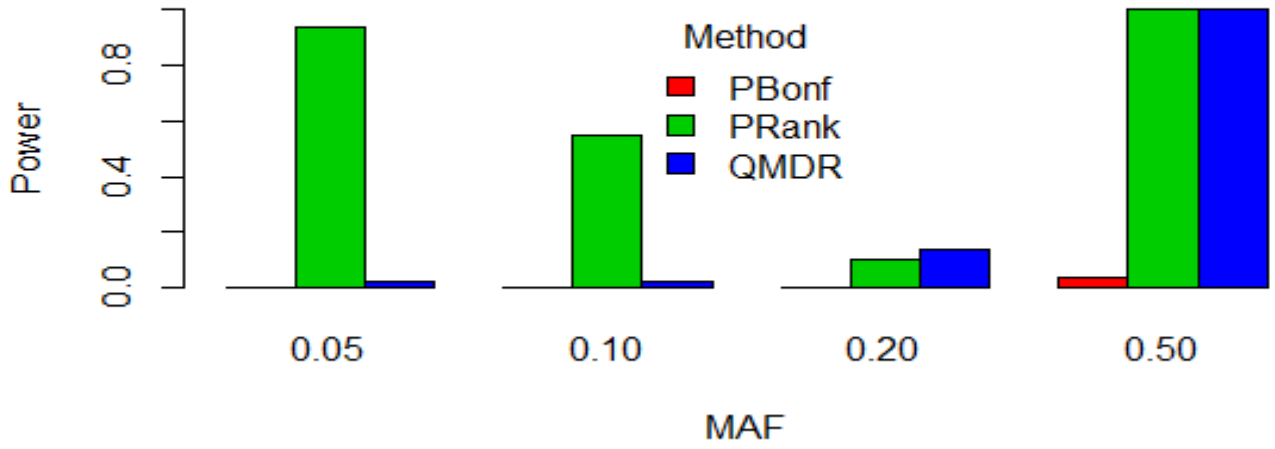


QMDR has no power due to the marginal effect of S3

Simulation IV: three-order interaction (binary trait)

- A model defined in Zhang & Liu (2007)
- For a certain genotype combination, there is an increased disease risk
- The interaction effects are decided in such a way that the marginal effect of each locus equals a specified value
- We set the marginal effect (the odd ratio minus 1), for instance, to be 0.2 for four different MAF values

Power Comparison



UM-MDR with PRank achieves the highest power across all different MAFs

Summary for simulation

1. When there are only pure interaction effects, our approach *PRank* has similar power as MDR (QMDR)

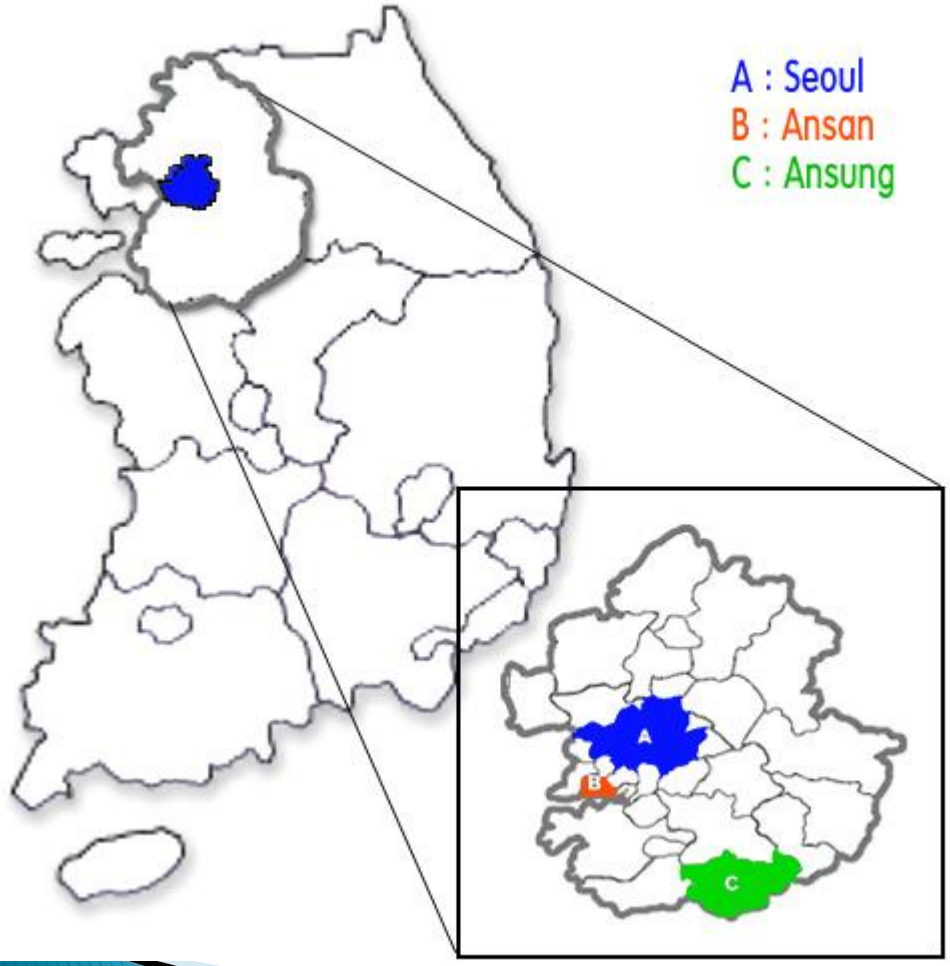
2. When there are marginal effects of SNPs, *PRank* outperforms MDR(QMDR)

3. When there is high-order causal effects, *PRank* outperforms MDR(QMDR)

Outline

- Introduction
- Method
- Simulation
- Application

Korea Association Resource (KARE) Data: Characteristics



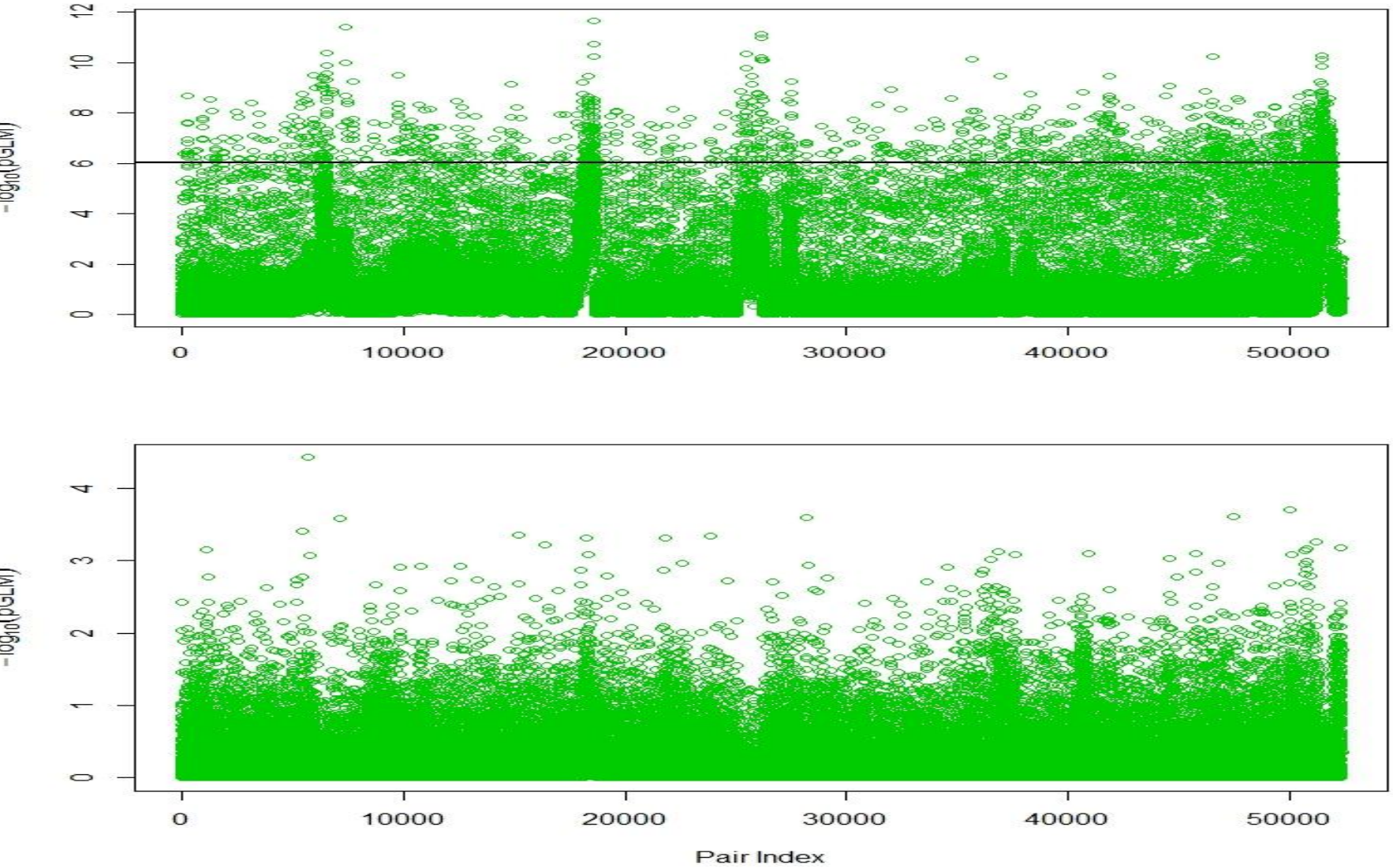
	Baseline study	
	Ansung	Ansan
Participants	5,018	5,020
Sex (women/men)	2,778/ 2,240	2,497/ 2,523
Age (mean)	55.5	49.1
40th (%)	31.2	62.8
50th (%)	29.1	23.0
60> (%)	39.6	14.3

Courtesy of KNIH

Analysis I: Application to KARE data for Lipid traits

- Korean Association Resource (KARE) project (Cho *et al.* 2009)
- The multivariate quantitative phenotypes for metabolic traits
  - **HDL**: high density lipoprotein cholesterol, **TG**: triglyceride, **LDL**: low density lipoprotein cholesterol
  - cor(HDL, TG) = −0.38, cor(HDL, LDL) = 0.10, cor(TG, LDL) = −0.06
- 8581 unrelated individuals, 344,596 SNPs; Recruitment area, Gender and Age are covariates
- 324 SNPs selected from preliminary analysis
  - Fitting linear regression of each phenotype to the covariates and a single SNP
  - The SNPs with p-values less than 0.0001 were kept for next step
- The trait HDL is used as the phenotype for GGI

Plot of p-value



Top detected model for analysis I (without marginal adjustment)

Top	SNP1	SNP2	P-value
1	rs271	rs10495536	2.25×10^{-12}
2	rs4970834	rs4713525	3.80×10^{-12}
3	rs486394	rs11782155	7.76×10^{-12}
4	rs486394	rs9288811	1.01×10^{-11}
5	rs271	rs2645371	1.86×10^{-11}

Top pairs (with marginal adjustment)

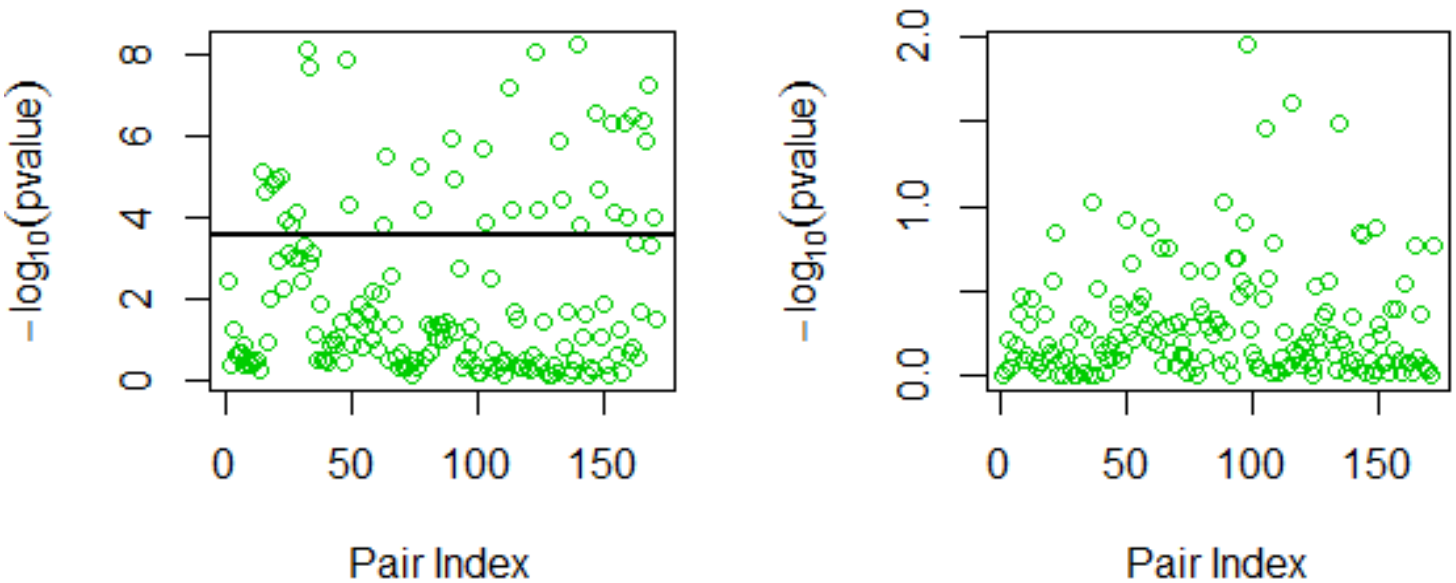
Top	SNP1	SNP2	P-value
1	rs765547	rs495348	3.77×10^{-5}
2	rs7079742	rs4512110	2.01×10^{-4}
3	rs7514421	rs17041893	2.48×10^{-4}
4	rs17120157	rs3909541	2.50×10^{-4}
5	rs12229654	rs1077410	2.64×10^{-4}

For QMDR: (rs11066280, rs12994068) is identified with CVC=3, which is quite different from the top pairs for our method

Analysis II—using candidate SNPs from literature

- The same KARE data, but using 19 candidate SNPs from the literature (*Willer et al. 2008*)
- Use the trait HDL as phenotype

Plot of P-value



Negative log(P-value) plots for all two-order multi-locus Models
Left --- without marginal adjustment
Right --- with marginal adjustment

Top detected model for analysis II (without marginal adjustment)

Top	SNP1	SNP2	P-value
1	rs10402271	rs780049	5.62×10^{-9}
2	rs12596776	rs780049	7.19×10^{-9}
3	rs4149268	rs780049	8.44×10^{-9}
4	rs1566439	rs780049	1.18×10^{-8}
5	rs12596776	rs17321515	1.82×10^{-8}

Top detected models (with marginal adjustment)

Top	SNP1	SNP2	P-value
1	rs2144300	rs10402271	0.011
2	rs2156552	rs17145738	0.025
3	rs2338104	rs17145738	0.032
4	rs2144300	rs1748195	0.034
5	rs693	rs6586891	0.094

Without multiple test correction

Some comments on real example results

- The top four models were significant at the 5% significance level
- These models may have a higher chance of being true epitasis, since we have already adjusted the marginal effects.
- QMDR, on the other hand, identifies **(rs12596776, rs17321515)** as the best two-order model with CVC=6, which may due to the marginal effect at large, because its p-value is 0.34 estimated by our study

Summary of real application

- With a marginal effect adjustment, many multi-locus models are not detected, which indicates most models identified without marginal effect adjustment may not be true epitasis

- The model identified by MDR approach may not be significant by our approach

Conclusion

- Proposed a unified model-based MDR approach, including a classification step and a subsequent modeling step
- The proposed approach is flexible in the sense that the various classification rules can be applied and different types of trait/traits can be used
- The modeling approach can provide significance of any multi-locus model, including traditional MDR approaches, while avoiding a large number of permutation
- Provide an easy way of measuring the significance of high-order interaction model

Further works

- More extensive simulation studies are needed, including higher order interaction and the multivariate traits analysis
- Compare the findings with existing MDR based approaches for real applications

Acknowledgement

Thank you for listening

What is different from MB-MDR

1. Classification step is more flexible; no intermediate group; not necessary to do multiple tests

2. Penalized regression can be used in the modeling step to account for pure marginal effects

3. Get p-value differently---not a huge number of permutation

Ridge regression

- To a standard linear regression model:

$$\boldsymbol{Y} = \boldsymbol{Z\beta} + \boldsymbol{\epsilon}$$

  - Y and Z stand for the response vector and design matrix, respectively, and $\epsilon_i \sim N(0,\ \sigma^2)$ iid

- The ordinary least squares estimator for β is $(Z'Z)^{-1}Z'Y$

- The ridge regression estimator is

$$\widehat{\boldsymbol{\beta}}^{\lambda} = (\boldsymbol{Z'Z} + \lambda \boldsymbol{I})^{-1}\boldsymbol{Z'Y}$$

  - where λ is a positive tuning parameter, and $\boldsymbol{I}$ is the identity matrix

- $var\left(\widehat{\boldsymbol{\beta}}^{\lambda}\right) = \sigma^2(\boldsymbol{Z'Z} + \lambda \boldsymbol{I})^{-1}\boldsymbol{Z'Z}(\boldsymbol{Z'Z} + \lambda \boldsymbol{I})^{-1}$

Logistic ridge regression

- The logistic regression model is:

$$\log\left(\frac{P(Y=1|\boldsymbol{X})}{P(Y=0|\boldsymbol{X})}\right) = \boldsymbol{X\beta}$$

  - For logistic regression model with ridge penalty, the estimator can be found by (*Vago and Kenndy 2006*)
  - $$\widehat{\boldsymbol{\beta}}^{\lambda} = \text{argmin } -\text{Log}\big(\text{L}(\boldsymbol{\beta}\ )\big) + \lambda\|\boldsymbol{\beta}\|_2$$
  - $\text{L}(\boldsymbol{\beta})$ is the likelihood function and Newton-Raphson algorithm is used to find $\hat{\beta}^{\lambda}$

- $$var\left(\widehat{\boldsymbol{\beta}}^{\lambda}\right) = (\boldsymbol{Z'WZ} + 2\lambda\boldsymbol{I})^{-1}\boldsymbol{Z'WZ}(\boldsymbol{Z'WZ} + 2\lambda\boldsymbol{I})^{-1}$$

  where $\boldsymbol{W} = diag[\hat{p}_i(1-\hat{p}_i)]$ and $\hat{p}_i = e^{X_i\widehat{\boldsymbol{\beta}}^{\lambda}}/(1+e^{X_i\widehat{\boldsymbol{\beta}}^{\lambda}})$

Choice of λ

- The λ was chosen to minimize the deviance by 10-fold cross validation
  - Deviance is defined as the mean squared error for ridge regression
  - Deviance is defined as $-2\log(\frac{L}{L_s})$, with L and L_s be the likelihood of fitted and saturated model, respectively.
  - When the satuarated model is not available, use $-2\log(L)$ instead

Acknowledgement

Lab members @SNU

Post D.

Ph.D. degree

Master degree

창의연구단
National Creative Research Initiatives