

Statistical models and computational tools for predicting complex traits and diseases

(정원일 교수님) 논문 리뷰

Use of a few published SNPs

β_j - SNP j의 효과 크기 (유전적 기여도 $\neq 0$)

x_{ij} - 개인 i의 SNP j에 대한 유전형

y_i - 개인 i의 표현형

$\hat{y}_i = \sum \beta_j x_{ij} = \text{PRS (다윈과 유전 점수)}$

표현형 지표

연속형 특성 - Predicted $R^2 \approx$ 실제 표현형 값과 예측 표현형 차이 상관관계의 제곱

이진 특성 - AUC (ROC curve), 약 R^2 , 정확도 R^2

예를 들면 GWAS-선택된 SNP가 설명한 유전 (R^2_{SNPs})에 의해 제한

→ 특정 SNP의 설명 정도를 설명할 수 있는 표현형 분산의 최대 비율

SNP의 예측 R^2 는 표현형의 유전적 강도 따라 달라짐

무한한 유전 강도 - 모든 SNP가 완벽하게 유전적 강도 예측 R^2 에 상응함

비한적 유전 강도 - 일부 SNP만이 유전적 강도에 큰 영향을 가지며, 대부분의 SNP는 효과가 0

상대적 유전 강도 예측 R^2

Use of all SNPs from GWAS studies

PRS (다윈과 유전 점수)

$\hat{y}_i = \sum \beta_j x_{ij}$ - GWAS-선택된 유전적 강도에 모든 SNP 사용

$E[R^2] \approx \frac{h^2}{1 + h^2 N}$

h^2 - 설명하는 유전적 강도

M - SNP의 총 개수

N - 개인의 수

↓

유전적 강도와 SNP의 설명 정도를 설명할 수 있는 표현형 분산의 최대 비율

모든 SNP를 활용한 PRS 계산의 주요 고려사항

1. 표현형의 비유전적 유전-강

p-value 사용

- 특정 임계값 P_t 이하의 p-value만 가진 SNP만 고려

- 최적의 P_t 임계값은 특정 생물학적 예측 정확도 가장 높을 때 선택

- 유전적 강도 검증 필요할 때 data

k-fold cross-validation을 통해 최적의 임계값을 선택하는 방법

2. LD 구조 고려

• LD 제거: 유전적 상관관계(r^2)에 기반하여 연관된 SNP 상 중 하나를 제외

• LD 클러스터링: 연관된 SNP 상 중 표현형에 대해 더 높은 p-value를 가진 SNP를 제거

but 이 방법 모두 최대 예측 정확도를 달성하지 못함

LD - Based prediction

LD 구조를 LD 자기 or LD 클러스터링 같은 방식으로 해결해왔지만 최근 선택 동계로 필요한 LD pred가 개발되었다.

선택 동계 추정.

$$P(\theta | \text{data}) = \frac{P(\text{data} | \theta) P(\theta)}{P(\text{data})}$$

LD - pred.

$$\hat{y}_i = \frac{\sum_j E(\beta_j | \hat{\beta}_j) x_{ij}}{L_{\text{SNP}} \text{의 선택 동계 효과 크기}}$$

SNP가 LD가 없는 특정한 경우.

① Under a Gaussian infinitesimal prior

$$\beta_j \sim N(0, \frac{h^2}{M})$$

$$E(\beta_j | \hat{\beta}_j) = \frac{h^2}{h^2 + M_e/N} \hat{\beta}_j$$

$$\hat{\beta}_j \sim \beta_j + e_j \sim e_j \sim N(0, \frac{1}{N})$$

SNP의 β_j 의 추정을 해.

② under a Gaussian non-infinitesimal prior

$\beta_j \sim N(0, \frac{h^2}{M_p})$ 의 확률 $P, \beta_j = 0$ 의 확률 $1-P$

$$E(\beta_j | \hat{\beta}_j) = \frac{h^2}{h^2 + M_e/N} \hat{\beta}_j$$

$\hat{\beta}_j$ - SNP가 causal인 확률.

$\hat{\beta}_j$ 로 비관망까지 추론 가능.

SNP가 LD가 없는 경우.

① Gaussian infinitesimal prior.

$$E(\beta_j | \hat{\beta}_j) = \frac{D + I}{L_{\text{LD}}} \hat{\beta}_j$$

② Gaussian non-infinitesimal prior

mcmc 기법 사용.

$$\beta_j \sim N(D\mu, D/N)$$

$$f(\beta_j | \hat{\beta}_j) = f(\beta_j) e^{-\frac{1}{2}(\hat{\beta}_j - D\mu)^T D^{-1}(\hat{\beta}_j - D\mu)}$$

BLUP - Based prediction.

(Best Linear Unbiased prediction)
모든 SNP의 효과 크기를 동시에 추정.

GBLUP

$$Y = X\beta + G + e$$

(Nx1) (Nx1) (Nx1) (Nx1) (Nx1)

G - 모든 개체에 대한 유전적 효과 $\sim N(0, \sigma_g^2 A)$

e - 유전적 효과 $\sim N(0, \sigma_e^2 I)$

$$\hat{g} = E(g|y) = \sigma_g^2 A (\sigma_g^2 A + \sigma_e^2 I)^{-1} (y - X\beta)$$

GBLUP $\xrightarrow{\text{유전적 효과}} \text{고지 효과 BLUP 모형}$

$$y = X\beta + W u + e$$

(Nx1) (Nx1) (Nx1)

W - 유전적 효과 관련 행렬.

u : 유전적 SNP 효과 벡터 $u \sim N(0, \sigma_u^2 I)$

$$\hat{u} = E(u|y) = W^T A^{-1} \hat{g} / M$$

A - 유전적 관계 행렬.

GBLUP은 개체 간의 유전적 관계를 고려하여 유전 효과를 추정하므로, 모든 개체에게 균등하게 유전적 효과를 예측할 수 있다.

SBLUP

GBLUP은 개체 간의 유전적 관계를 고려하여 유전 효과를 추정하는 방법이지만, 유전적 효과가 없는 개체를 포함할 경우 문제가 발생한다.

모든 개체 SBLUP.

LD pred 모델과 유사한 In infinitesimal model 관련 정리.

$$\hat{g} = (W^T W + \lambda I)^{-1} W^T y$$

W - 유전적 효과 관련 행렬.

λ - 정규화 파라미터

$$E(W^T W) = V^T V = B$$

V : 유전적 효과 관련 행렬

$$E[W^T y] \approx \text{diag}(B) \beta$$

$$\hat{u} = (B + \lambda I)^{-1} \text{diag}(B) \beta$$

$$h_g^2 = \frac{M \sigma_g^2}{\sigma_g^2 + \sigma_e^2} \quad \text{유전적 효과 관련 행렬}$$

$$\lambda = M \times \left(\frac{1}{h_g^2} - 1 \right)$$

$$\hat{g}_{\text{new}} = W_{\text{new}} \hat{u}$$

W_{new} : 새로운 데이터셋의 유전적 효과 관련 행렬

BMR- Based prediction

BMR- 표준 실험 결과 모형을 확장하여 SNP
효과에 대한 대체 사전 확률
포함함으로써 예측 정확도를 향상시킨다.

Bayes R - 가정: y 가 다수의 SNP 효과(β)와
간섭효과(e)의 선형 결합으로
표현되며 다중 실험 결과 모형을
가정.

$$y = X\beta + e$$

Bayes R은 각 SNP 효과가 서로 다른 분산을
가진 여러 개의 정규분포에서 유래한다고 가정

$$P(\beta_j | \pi, \sigma^2_{\beta_j}) = \sum_{\pi} \pi \cdot N(\beta_j | 0, \sigma^2_{\beta_j})$$

$$\sim N(\beta_j | 0, \sigma^2_{\beta_j})$$

$$\text{posterior } \beta : P(\beta | \pi, \sigma^2_{\beta_j}, \sigma^2_e)$$

$$\propto P(\beta | \pi, \sigma^2_{\beta_j}) P(\pi) P(\sigma^2_e) P(\sigma^2_e)$$

SBayesR .

GWAS 효과 통계를 활용하여 다중 실험
결과 개수 (β)의 추장치를 얻는 방법.

$$y = X\beta + e$$

SBayes R은 GWAS 효과 통계만을
사용하여 SNP 효과를 추장하고,
이를 기반으로 PRS를 계산.

→ 개별 수치의 유전적 영향 및

표현형 데이터가 부족한 상황에서도
유용.

Penalized Regression - Based prediction

sup 기준 - 모든 sup 기준 양쪽의 기준
but 실제로 복잡한 동등성 기법은 존재하지
않고 몇 가지의 sup와 관련이 있음



Penalized regression을 사용하여 다양한
기준 기준을 포함한 PLS 기준

Lasso and elastic net.

The elastic net regression.

$$f(\beta) = (Y - X\beta)^T (Y - X\beta) + \lambda \alpha \|\beta\|_1 + (1 - \alpha) \|\beta\|_2^2$$

$\|\beta\|_1$ 은 L1 기준

$\|\beta\|_2$ 은 L2 기준

$\alpha = 1 \rightarrow$ 라소화

$\alpha = 0 \rightarrow$ 리귀처

~~Lasso~~

Multi-trait Approaches

다중 형질 동등성 다중 기준과 예측의
정확도는 더욱 향상시킬 수 있다.