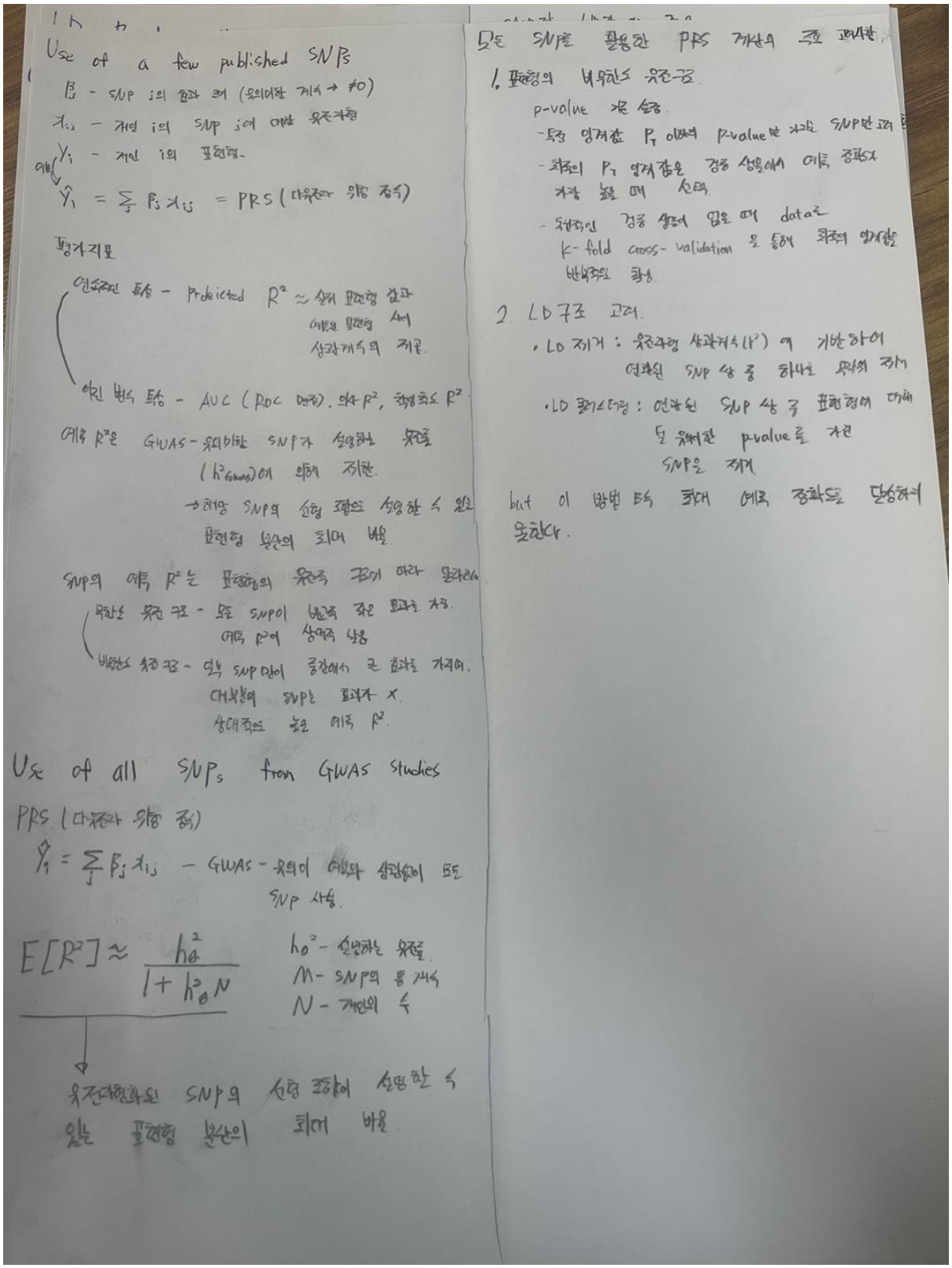
Statistical models and computational tools for predicting complex traits and diseases

(정원일 교수님) 논문 리뷰



LD - Based prediction

LD 구조를 LD 제거 or LD 클럼핑 같은 방식으로 해결해왔지만 최근 요약 통계를 활용한 LD pred가 개발되었다.

사후 평균 측정.

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

LD - Pred.

$$\hat{Y}_i = \sum_j E(\beta_j | \hat{\beta}_j) x_{ij}$$

└ SNP j의 사후 평균 효과 크기

SNP 간 LD가 없는 독립한 경우.

① Under a Gaussian infinitesimal prior

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

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

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

SNP j의 $\hat{\beta}_j$의 축소 추정으로 해석.

② under a Gaussian non-infinitesimal prior

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

$$E(\beta_j | \hat{\beta}_j) = \frac{h_g^2}{h_g^2 + M_p/N} \bar{p}_j \hat{\beta}_j$$

$\bar{p}_j$ - SNP j가 causal일 확률.

$\hat{\beta}_j$를 비선형적으로 축소한 형태.

SNP 간 LD가 있는 경우.

① Gaussian infinitesimal prior.

$$E(\beta_j | \hat{\beta}_j) = [D + I]^{-1} \hat{\beta}_j$$

└ LD 행렬.

② Gaussian non-infinitesimal prior

mcmc 기법 사용.

$$\hat{\beta}_j \sim N(D\beta, D/N)$$

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

BLUP - Based prediction.

(Best Linear Unbiased prediction)

모든 SNP의 효과 크기를 동시에 추정

GBLUP

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

(N×1) (N×c) (c×1) (N×1) (N×1)

- g - 모든 개체에 대한 무작위 유전 벡터 $\sim N(0, \sigma_g^2 A)$
- e - 무작위 오차 벡터 $\sim N(0, \sigma_e^2 I)$
- $\hat{g} = E(g \mid y) = \sigma_g^2 A(\sigma_g^2 A + \sigma_e^2 I)^{-1}(y - X\beta)$

GBLUP —(변환 가능)→ 리지 회귀 BLUP 모델.

$$y = X\beta + Wu + e$$

(N×m) (m×1)

W - 표준화된 유전자형 행렬.

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

$$\hat{u} = E[u \mid g] = W^T A^{-1} \hat{g} / M$$

A - 유전 관계 행렬

GBLUP는 개체 간의 유전적 관계를 고려하여 유전 효과를 추정하므로, 보다 통계 기법보다 더 정확한 예측과 생성할 수 있다.

SBLUP

GBLUP는 개별 수준의 유전자형과 표현형 데이터가 필요. 이 데이터가 없는 경우 근사화 하는 모델이 SBLUP.

LD pred 모델과 비슷하게 infinitesimal model 만을 고려.

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

W - 표준화 된 유전자형 행렬.

λ - 정규화 파라미터

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

V: 참조 패널의 표준화된 유전자형 행렬.

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

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

$$h_g^2 = \frac{M\sigma_g^2}{\sigma_p^2}$$

— 유전적 분산

— 표현형 분산.

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

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

W_{new}: 대상 데이터셋의 표준화된 유전자형

BMR- Based prediction

BMR- 공분 선형 혼합 모델을 확장하여 SNP 효과에 대한 대체 사전 확률을 모형 함수로서 예측 정확도를 향상시킨다.

Bayes R - 가정: Y가 다수의 SNP 효과(β)와 관찰오차(e)의 선형 결합으로 표현되는 다중 선형 회귀 모형 가정.

$$Y = X\beta + e$$

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

$$P(\beta_j \mid \pi, \sigma^2_\beta) = \sum_{c=1}^{C} \pi_c \, N(\beta_j \mid 0, \sigma^2_{\beta c})$$

$$\sim N(\beta_j \mid 0, \sigma^2_{\beta c})$$

posterior β : $P(\beta \mid \pi, \sigma^2_\beta, \sigma^2_e)$

$$\propto P(\beta \mid \pi, \sigma^2_\beta) P(\pi) \; P(\sigma^2_\beta) P(\sigma^2_e)$$

SBayesR.

GWAS 요약 통계를 활용하여 다중 선형 회귀 계수 (β)의 측정치를 얻는 방법.

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

SBayesR은 GWAS 요약 통계만을 사용하여 SNP 효과를 측정하고, 이를 기반으로 PRS를 계산.

→ 개별 수준의 유전자형 및 표현형 데이터가 부족한 상황에서도 유용.

Penalized Regression-Based prediction

GBLUP 계통 - 표현도 유전 구조를 반영하지 못함
but 실제로 복잡한 특성이나 질병은 유전체에서
수 많은 존재에 의해 SNP와 관련이 있음.

↓

Penalized regression을 사용하여 비유전적
유전 구조를 수용하는 모델로 구축.

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 규제
$||\beta||_2$ 은 L2 규제
$\alpha=1$ → 라쏘회귀
$\alpha=0$ → 리지회귀

~~Lasso~~

Multi-trait Approaches

다중 형질을 동시에 다룸 다중 유전자 예측의
정확도를 더욱 향상시킬 수 있다.