# Prediction intervals for polygenic scores

Fred Boehm

September 23, 2022

# Outline

1. Polygenic scores
2. DBSLMM
3. Prediction intervals for PGS
4. Jackknife+ for prediction intervals
5. Crossvalidation+ for prediction intervals

# Polygenic scores

# Polygenic scores

- Polygenic scores aim to summarize genetic contributions to complex traits:

$$\sum_{SNPs} (\text{SNP genotype}) * (\text{SNP effect})$$

- Our goal is to develop a strategy for constructing prediction intervals for PGS (quantitative or binary traits)

# Popular PGS methods with GWAS summary statistics

1. DBSLMM [YZ20]
2. ldpred2
3. lassosum2
4. SBLUP
5. C + T

# Prediction intervals for PGS

# Existing approaches to prediction intervals in PRS

1. Mondrian cross-conformal prediction intervals for PRS [Sun+21]
    - One approach to conformal prediction
2. Bayesian credible intervals for PRS [Din+21]
    - ldpred2 used to obtain posterior samples
    - Observe large variances in PRS

# Importance of prediction intervals for PGS

- Clinical utility of an interval estimate in addition to point estimate
- Large variability in PRS with ldpred2
  - Method applies only to ldpred2

# DBSLMM Model & Methods

# DBSLMM

- All SNPs have nonzero effects on the trait
- Each SNP effect arises from one of two normal distributions
  - Large variance or small variance
- Treat the large variance SNP effects as fixed effects & small variance SNP effects as random effects (omnigenic hypothesis)

# DBSLMM Model

$$y = X\beta + \epsilon$$

- trait $y$: n vector
- $X$: n by m matrix of standardized SNP genotypes
- $\beta$: m vector of SNP effects
- $\epsilon$: n vector of random errors with precision $\tau$

# DBSLMM Model

$$y = X\beta + \epsilon$$

$$\beta_j \sim \pi N(0, \sigma_l^2 \tau^{-1}) + (1-\pi) N(0, \sigma_s^2 \tau^{-1})$$

- $\pi$ proportion of SNPs in the large variance component

# DBSLMM Model Fitting

$$y = X\beta + \epsilon$$

- BSLMM: MCMC for model fitting is slow with large memory requirements
- Large effect SNPs should be easy to identify from GWAS analysis
- Small effect SNPs can't be inferred accurately
- But polygenic effects may be inferred with accuracy

# DBSLMM Model

$$y = X_l\beta_l + X_s\beta_s + \epsilon$$

- $X_l$ : n by $m_l$ SNP genotypes matrix for large effect SNPs
- $\beta_l$ : $m_l$ effects vector for large effect SNPs
- $X_s$ : n by $m_s$ SNP genotypes matrix for small effect SNPs
- $\beta_s$ : $m_s$ effects vector for small effect SNPs

$$\beta_{lj} \sim N(0, \sigma_l^2\tau^{-1})$$

$$\beta_{sj} \sim N(0, \sigma_s^2\tau^{-1})$$

- Set $\sigma_l^2 \to \infty$ & treat $\beta_l$ as fixed effects

# DBSLMM Model: Parameter estimation

- Clumping and Thresholding (C + T) procedure in PLINK to select large effect SNPs
  - One chromosome at a time
  - p-value threshold: $10^{-6}$
  - region size: 1 MB
  - LD threshold: $r^2 = 0.1$
- Combine large effect SNPs across genome to get $m_l$ SNPs

# DBSLMM Model: Parameter estimation

$$\hat{\beta}_l = (X_l^T H^{-1} X_l)^{-1} X_l^T H^{-1} y$$

$$\hat{\beta}_s = \hat{\sigma}_s^2 X_s^T H^{-1} (y - X_l \hat{\beta}_l)$$

$$Var(y) = H = \hat{\sigma}_s^2 X_s X_s^T + I_n$$

# DBSLMM Model: Parameter estimation

- Set $\hat{\sigma}_s^2$ to predetermined value instead of estimating it
  - LD score regression to get SNP heritability, $\hat{h}^2$
  - Set $\hat{\sigma}_s^2 = \frac{\hat{h}^2}{m}$

# DBSLMM Model: Parameter estimation

- Use Woodbury matrix identity to calculate $H^{-1}$

$$H^{-1} = I_n - X_s(\sigma_s^{-2} I_{m_s} + X_s^T X_s)^{-1} X_s^T$$

# Prediction intervals for PGS

# Jackknife-plus for prediction intervals [Bar+20]

- Uses ideas and results from conformal prediction theory
- Comes with probabilistic coverage guarantees
- Assumes exchangeability of observations
- Uses leave-one-out residuals

# Jackknife-plus for prediction intervals [Bar+20]

- JK+ constructs prediction interval for $Y_{n+1}$ as a function of $n$ training points $(X_i, Y_i)$ & $X_{n+1}$
- Naively, we might want to use residuals from the training data to construct the interval:

$$(X_{n+1}) \pm (\text{the } (1-\alpha) \text{ quantile of the } n \text{ absolute residuals})$$

- Residuals are $|Y_1 - \hat{\mu}(X_1)|, \ldots, |Y_n - \hat{\mu}(X_n)|$
- Due to overfitting, $n$ training residuals tend to be smaller than that of the $(n+1)^{th}$ point

# Jackknife-plus for prediction intervals [Bar+20]

- JK computes leave-one-out residuals:
  - $R_i = |Y_i - \hat{\mu}_{-i}(X_i)|$
- And computes the regression function $\hat{\mu}$ with all $n$ training points
- And outputs the interval:
  - $\hat{\mu}(X_{n+1}) \pm (\text{the } (1-\alpha) \text{ quantile of } R_1, \ldots, R_n)$
- [Bar+20] point out that JK has no universal theoretical guarantees & may lose predictive coverage in some settings

# Jackknife-plus for prediction intervals [Bar+20]

- JK+ is a modification of JK
  - Replace $\hat{\mu}$ with $\hat{\mu}_{-i}$

# Jackknife-plus for prediction intervals [Bar+20]

- Notation
  - $\hat{q}^{+}_{n,\alpha}\{v_i\} = \text{the } \lceil (1-\alpha)(n+1) \rceil \text{-th smallest value of } v_1, \ldots v_n$
  - $\hat{q}^{-}_{n,\alpha}\{v_i\} = \text{the } \lfloor \alpha(n+1) \rfloor \text{-th smallest value of } v_1, \ldots v_n$

# Jackknife-plus for prediction intervals [Bar+20]

- $\hat{C}^{\text{naive}}_{n,\alpha}(X_{n+1}) = \hat{\mu}(X_{n+1}) \pm \hat{q}^{+}_{n,\alpha}\{|Y_1 - \hat{\mu}(X_1)|, \ldots, |Y_n - \hat{\mu}(X_n)|\}$
- $\hat{C}^{\text{JK}}_{n,\alpha}(X_{n+1}) = \left(\hat{q}^{-}_{n,\alpha}\{\hat{\mu}(X_{n+1}) - R^{LOO}_i\}, \hat{q}^{+}_{n,\alpha}\{\hat{\mu}(X_{n+1}) + R^{LOO}_i\}\right)$
- $\hat{C}^{\text{JK+}}_{n,\alpha}(X_{n+1}) = \left(\hat{q}^{-}_{n,\alpha}\{\hat{\mu}_{-i}(X_{n+1}) - R^{LOO}_i\}, \hat{q}^{+}_{n,\alpha}\{\hat{\mu}_{-i}(X_{n+1}) + R^{LOO}_i\}\right)$

# Jackknife-plus for prediction intervals [Bar+20]

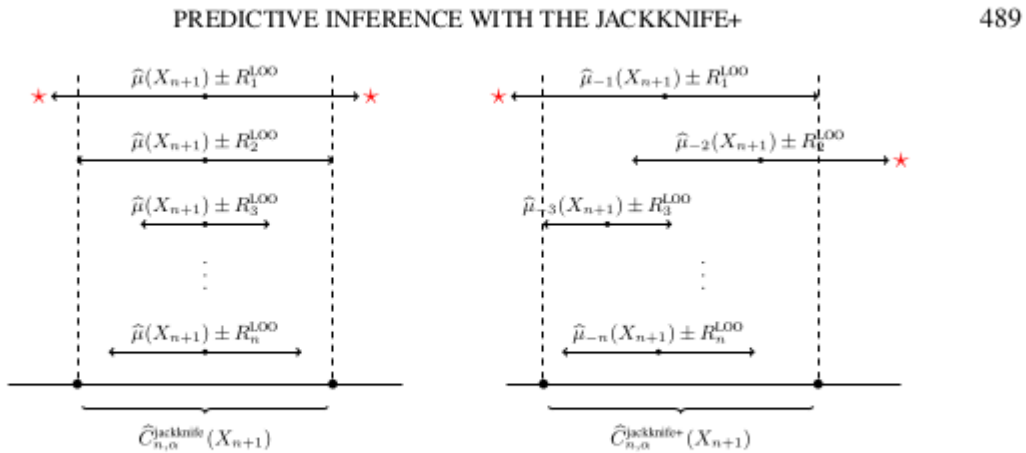



FIG. 1. *Illustration of the usual jackknife and the new jackknife+. The resulting prediction intervals are chosen so that, on either side, the boundary is exceeded by a sufficiently small proportion of the two sided arrows—above, these are marked with a star.*

# Cross-validation+ for $K$-fold crossvalidation

- Split training set into $K$ disjoint sets of equal size, $m = \frac{n}{K}$
- $\hat{\mu}_{-S_k} = \mathcal{A}\left((X_i, Y_i) : i \in \{1, \ldots, n\} \setminus S_k\right)$
- $R_i^{CV} = |Y_i - \hat{\mu}_{-S_{k(i)}}(X_i)|$ with $i \in S_{k(i)}$
- $\hat{C}^{CV+}_{n,K,\alpha}(X_{n+1}) = \left(\hat{q}^{-}_{n,\alpha}\{\hat{\mu}_{-S_{k(i)}}(X_{n+1}) - R_i^{CV}\}, \hat{q}^{+}_{n,\alpha}\{\hat{\mu}_{-S_{k(i)}}(X_{n+1}) + R_i^{CV}\}\right)$
- CV+ requires $K$ model fits instead of $n$ for JK+
  - CV+ intervals may be wider due to smaller sample size

# Simulations

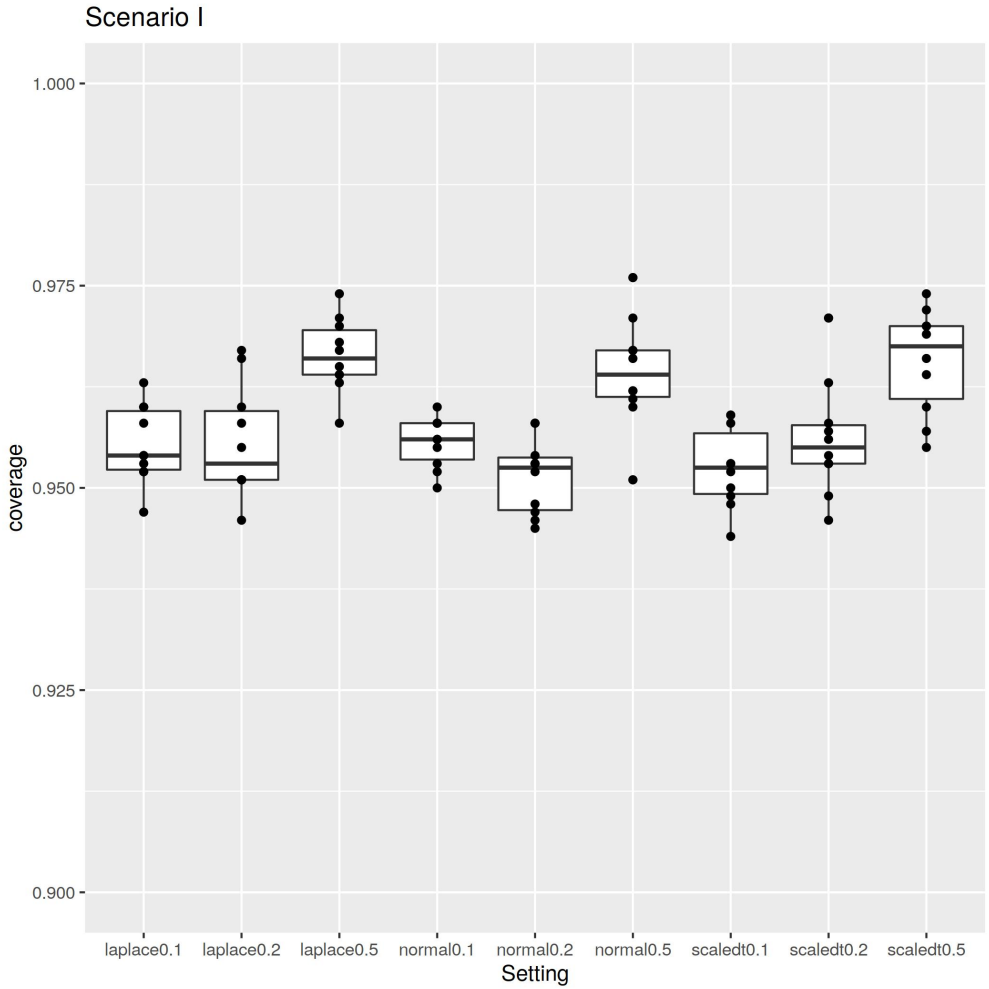
# Simulations study design

- 14,500 subjects randomly chosen from 337129 UKB subjects
  - 12500 randomly assigned to training set
  - 1000 randomly assigned to validation set
  - 500 validation subjects also randomly assigned to reference set
  - 1000 remaining subjects assigned to "verification" set
- 5-fold cross-validation used with the 12,500 training set
- Chose all ~95,000 Chr1 SNPs for simulations

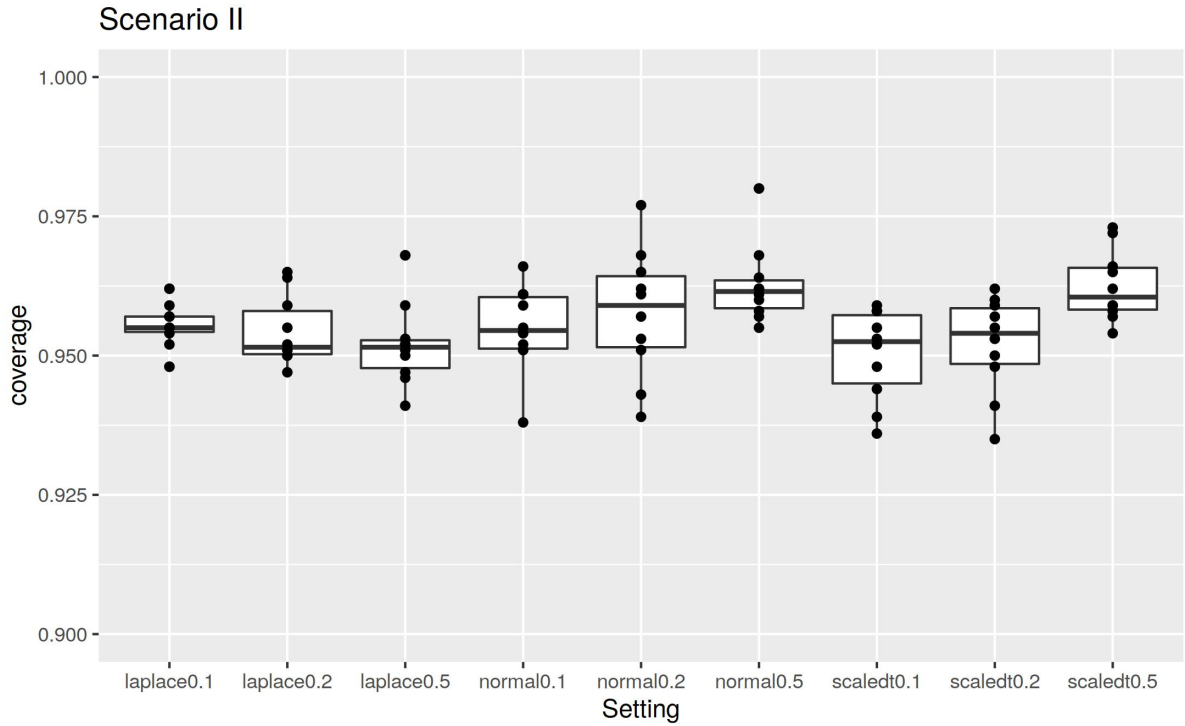
# Simulations study design

- Quantitative traits simulated - with GCTA - according to four scenarios:
  - Scenario I: Polygenic (all SNPs are causal)
  - Scenario II: Sparse (0.1% of SNPs are causal)
  - Scenario III: Hybrid (all SNPs are causal, and 0.1% of SNPs have large effects, PGE = 0.2)
  - Scenario IV: Hybrid (all SNPs are causal, and 0.1% of SNPs have large effects, PGE = 0.5)
  - 3 distributions: Laplace, normal, scaled t
  - 3 heritabilities: 0.1, 0.2, 0.5
  - 10 replicates per setting

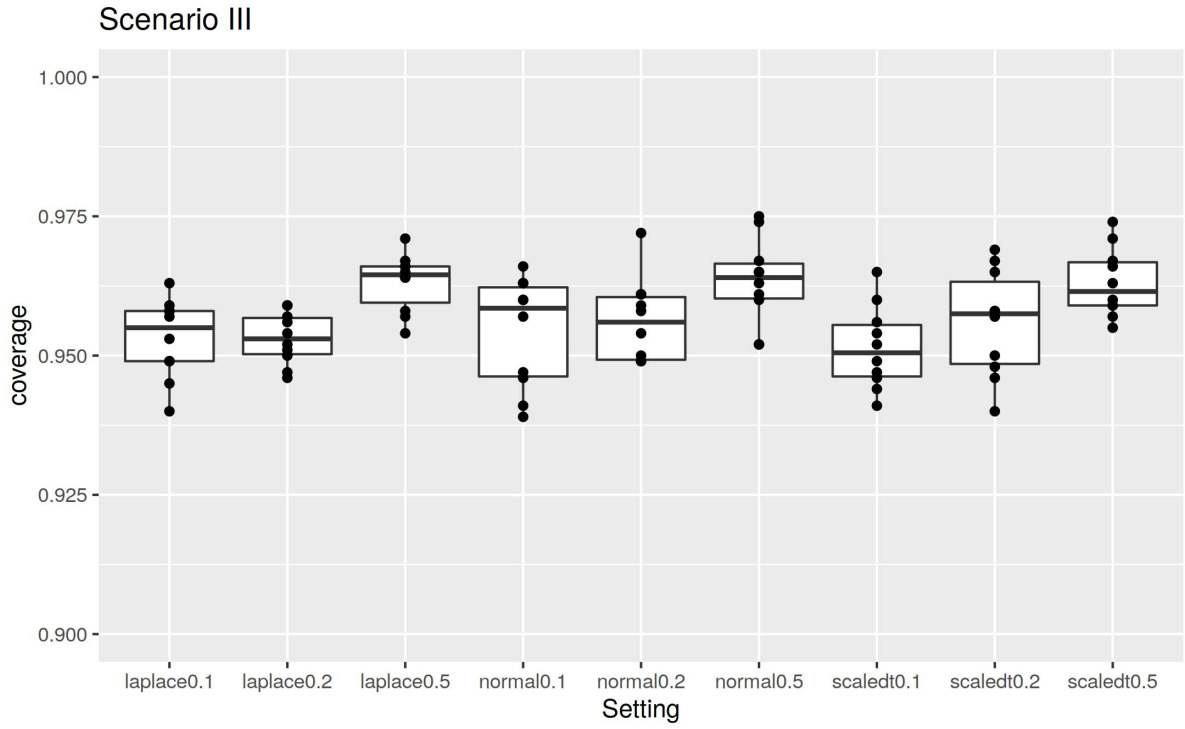
# Simulations results



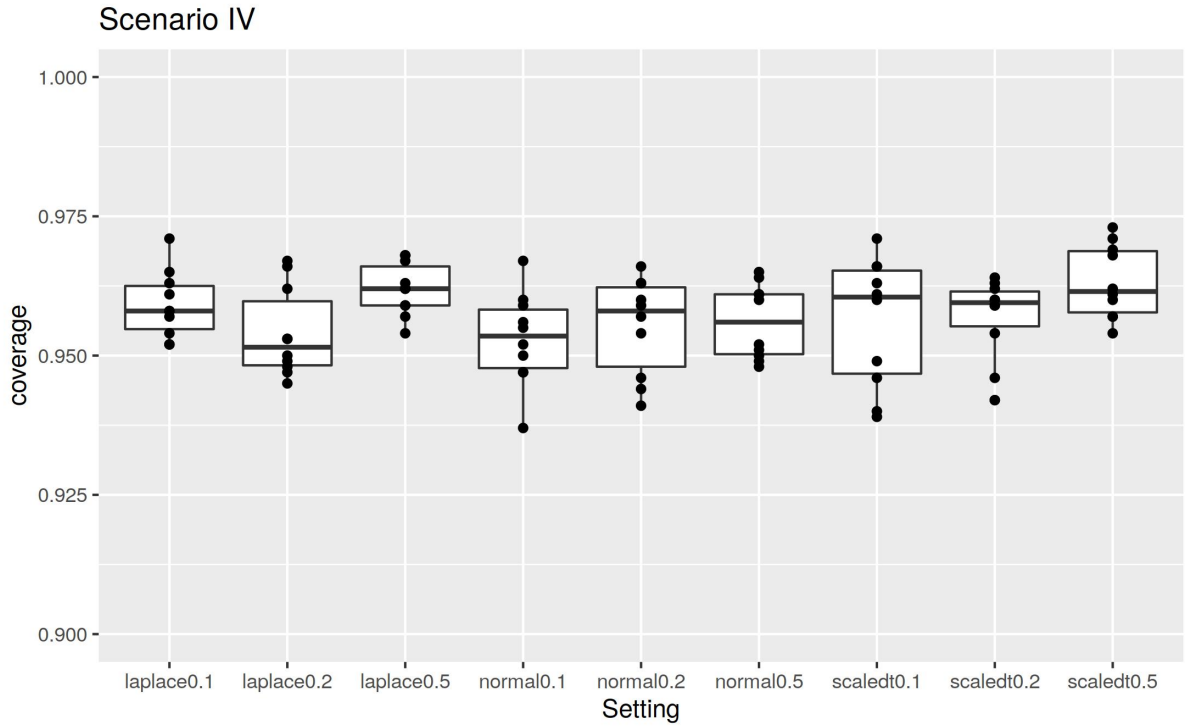
# Simulations results



# Simulations results



# Simulations results



# UKB quantitative traits

- Sheng Yang processed & analyzed 25 quantitative traits from UKB
- I've constructed CV+ prediction intervals for the 25 traits
- Coverage ranges from 0.943 to 0.965 for nominally ~95% intervals

# Next Steps

- Troubleshoot simulations
  - Is there a bug in my CV+ R code?
- Analyze binary traits from UKB
  - Sheng Yang already created the needed summary statistics files (gemma outputs)

# Thank you!

# References


Barber, R. F., E. J. Candes, A. Ramdas, et al. (2020). "Predictive inference with the jackknife+". In: *arXiv:1905.02928 [stat]*. arXiv: 1905.02928. URL: http://arxiv.org/abs/1905.02928 (visited on Sep. 12, 2021).

Ding, Y., K. Hou, K. S. Burch, et al. (2021). "Large uncertainty in individual polygenic risk score estimation impacts PRS-based risk stratification". In: *Nature Genetics*. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-021-00961-5. URL: https://www.nature.com/articles/s41588-021-00961-5 (visited on Dec. 25, 2021).

Sun, J., Y. Wang, L. Folkersen, et al. (2021). "Translating polygenic risk scores for clinical use by estimating the confidence bounds of risk prediction". In: *Nature Communications* 12.1, p. 5276. ISSN: 2041-1723. DOI: 10.1038/s41467-021-25014-7. URL: https://www.nature.com/articles/s41467-021-25014-7 (visited on Oct. 10, 2021).

Yang, S. and X. Zhou (2020). "Accurate and Scalable Construction of Polygenic Scores in Large Biobank Data Sets". In: *The American Journal of Human Genetics* 106.5, pp. 679-693. ISSN: 00029297. DOI: 10.1016/j.ajhg.2020.03.013. URL: https://linkinghub.elsevier.com/retrieve/pii/S0002929720301099 (visited on Jun. 01, 2021).