mtPGS: Multi-Trait Assisted Polygenic Scores

Leveraging Correlated Traits for Accurate Genetic Prediction

Chang Xu, Santhi K. Ganesh, Xiang Zhou

University of Michigan School of Public Health

June 10, 2025

Overview

1. **Motivation & Problem Statement**
2. **Technical Challenges in Current Methods**
3. **mtPGS Methodology & Statistical Framework**
4. **Algorithmic Implementation & Computational Complexity**
5. **Experimental Validation & Results**
6. **Discussion & Future Directions**

Motivation & Problem Statement

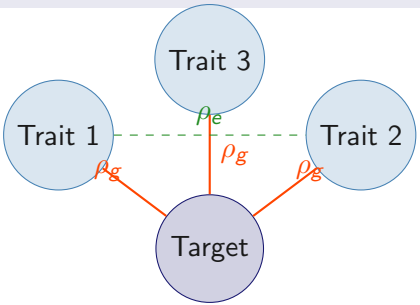
Limitations of Current Polygenic Score Construction

Current Challenges:

- SNP effect size estimation suffers from noise
- Univariate methods ignore trait correlations
- Existing multivariate methods have limitations:
  - Inflexible modeling assumptions
  - Ignore environmental correlations
  - Computationally intensive (MCMC)

Key Insight:

- Genetic correlations among traits contain information
- Environmental correlations due to sample overlap
- Need a flexible, scalable framework



PGS Formula:

$$\text{PGS}_i = \sum_{j=1}^{p} x_{ij}\hat{\beta}_j$$

where x_{ij} is genotype, $\hat{\beta}_j$ is effect size

Technical Challenges in Multivariate PGS Methods

Limitations of Existing Approaches

Method	Input	Architecture Model	Key Limitations
wMT-SBLUP	Summary	Multivariate Normal	Simple covariance structure
MTAG	Summary	Linear Combination	No environmental correlation
XPXP	Summary	Cross-population	Limited to population differences
BVR	Individual	Bivariate Ridge	Requires individual data
MTGBLUP	Individual	Multivariate BLUP	Not scalable to biobanks

Fundamental Issues:

1. **Inflexible Effect Size Distribution:** Single multivariate normal cannot capture complex genetic architectures
2. **Environmental Correlation Ignored:** Sample overlap creates dependencies not modeled
3. **Computational Scalability:** MCMC methods don't scale to biobank data ($n > 500K$)
4. **Over-shrinkage Problem:** Large effects get inappropriately shrunk toward zero

mtPGS Statistical Framework

Flexible Bivariate Modeling with Environmental Correlations

Regression Model Setup: For target trait y_0 and relevant trait y_m:

$$y_0^* = X_0^* \beta_0 + \epsilon_0^* \quad \text{(non-overlapping individuals)} \tag{1}$$

$$y_m^* = X_m^* \beta_m + \epsilon_m^* \quad \text{(non-overlapping individuals)} \tag{2}$$

$$[\tilde{y}_0, \tilde{y}_m] = \tilde{X}[\beta_0, \beta_m] + [\tilde{\epsilon}_0, \tilde{\epsilon}_m] \quad \text{(overlapping)} \tag{3}$$

Flexible SNP Effect Prior - Mixture of Bivariate Normals:

$$\begin{bmatrix} \beta_{0j} \\ \beta_{mj} \end{bmatrix} \sim \sum_{k=1}^{4} \pi_k \mathcal{BN}(\mathbf{0}, \mathbf{\Sigma}_k)$$

Four Component Mixture:

- π_{11}: Large effects on both traits
- π_{10}: Large on target, small on relevant
- π_{01}: Small on target, large on relevant
- π_{00}: Small effects on both traits

Covariance Structure:

$$\mathbf{\Sigma}_k = \begin{bmatrix} \sigma^2_{0l/s} & \rho_g \sigma_{0l/s} \sigma_{ml/s} \\ \rho_g \sigma_{0l/s} \sigma_{ml/s} & \sigma^2_{ml/s} \end{bmatrix}$$

l/s indicates large/small effect

Environmental Correlation Modeling

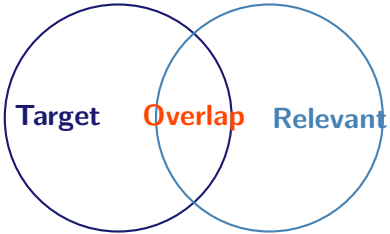
Explicit Treatment of Sample Overlap Dependencies

Environmental Effect Model:

For non-overlapping individuals: $\epsilon^*_{0i} \sim \mathcal{N}(0, \sigma^2_{0e}), \quad \epsilon^*_{mj} \sim \mathcal{N}(0, \sigma^2_{me})$

For overlapping individuals: $\begin{bmatrix} \tilde{\epsilon}_{0i} \\ \tilde{\epsilon}_{mi} \end{bmatrix} \sim \mathcal{BN}(\mathbf{0}, \mathbf{V}_e)$

where $\mathbf{V}_e = \begin{bmatrix} \sigma^2_{0e} & \rho_e \sigma_{0e} \sigma_{me} \\ \rho_e \sigma_{0e} \sigma_{me} & \sigma^2_{me} \end{bmatrix}$



ρ_e models environmental correlation in overlap region

Deterministic Inference Algorithm (1/2)

Scalable Alternative to MCMC

Key Innovation: Replace 4^p dimensional MCMC search with deterministic approximation

GECKO Parameter Estimation:

- Heritability (h^2): Method-of-moments estimator
- Genetic correlation (ρ_g): Cross-trait LD score regression
- Environmental correlation (ρ_e): Sample overlap-based estimation

Deterministic Inference Algorithm (2/2)

Scalable Alternative to MCMC

Algorithm 1 mtPGS Deterministic Algorithm

Input: GWAS summary statistics, LD matrix $\mathbf{R}$
Estimate h^2, ρ_g, ρ_e using GECKO
for each trait pair (target + relevant) **do**
    Identify large-effect SNPs via C+T procedure:
        Parameters: p-threshold $\in \{10^{-5}, 10^{-6}, 10^{-7}, 10^{-8}\}$
        LD threshold $r^2 \in \{0.1, 0.2, 0.25\}$, window = 1000kb
    Partition SNPs: $\boldsymbol{\beta} = [\boldsymbol{\beta}_l^T, \boldsymbol{\beta}_s^T]^T$
    Solve linear system for $\hat{\boldsymbol{\beta}}_l$, $\hat{\boldsymbol{\beta}}_s$:
        $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{X} + \mathbf{D}^{-1})^{-1}\mathbf{X}^T\boldsymbol{\Sigma}^{-1}\mathbf{y}$
    Construct PGS: $s_m = \mathbf{x}^T\hat{\boldsymbol{\beta}}_0$
end for
Combine PGSs: $s = \sum_m w_m s_m$ (weights via cross-validation)
Output: Final polygenic score s

Analytical Solution Details

Closed-Form Expressions for Effect Size Estimation

Linear System Solution: $\hat{\beta} = (X^T\Sigma^{-1}X + D^{-1})^{-1}X^T\Sigma^{-1}y$

where:

- Σ: Phenotypic covariance matrix, D: Prior precision matrix (mixture-dependent)
- Analytical forms for large/small effect SNPs

Large Effect SNPs: $\begin{bmatrix}\hat{\beta}_{0l}\\ \hat{\beta}_{ml}\end{bmatrix} = \left[\left(n_{ms}\hat{\mathbf{V}}_e^{-1} + \begin{bmatrix}\hat{\sigma}_{0e}^{-2}n_0^* & 0\\ 0 & \hat{\sigma}_{me}^{-2}n_m^*\end{bmatrix}\right)\otimes \mathbf{S}_{ll}\right]^{-1}\mathbf{z}_l$

Small Effect SNPs:

$$\begin{bmatrix}\hat{\beta}_{0s}\\ \hat{\beta}_{ms}\end{bmatrix} = \left[2(\hat{\mathbf{V}}_g\otimes \mathbf{I}_{p_s})^{-1} + \left(n_{ms}\hat{\mathbf{V}}_e^{-1} + \begin{bmatrix}\hat{\sigma}_{0e}^{-2}n_0^* & 0\\ 0 & \hat{\sigma}_{me}^{-2}n_m^*\end{bmatrix}\right)\otimes \mathbf{S}_{ss}\right]^{-1}\mathbf{z}_s^{adj}$$

where:

- $\mathbf{S}_{ll}, \mathbf{S}_{ss}$: LD matrices for large/small effect SNPs, $\mathbf{z}_l, \mathbf{z}_s^{adj}$: (Adjusted) Z-score vectors
- $\hat{\mathbf{V}}_g = \begin{bmatrix}\hat{\sigma}_{0s}^2 & \hat{\rho}_g\hat{\sigma}_{0s}\hat{\sigma}_{ms}\\ \hat{\rho}_g\hat{\sigma}_{0s}\hat{\sigma}_{ms} & \hat{\sigma}_{ms}^2\end{bmatrix}$: Genetic covariance
- $\hat{\mathbf{V}}_e = \begin{bmatrix}\hat{\sigma}_{0e}^2 & \hat{\rho}_e\hat{\sigma}_{0e}\hat{\sigma}_{me}\\ \hat{\rho}_e\hat{\sigma}_{0e}\hat{\sigma}_{me} & \hat{\sigma}_{me}^2\end{bmatrix}$: Environmental covariance

Computational Optimizations

Achieving Linear Scalability in SNP Number

1. Block-Diagonal LD Approximation:

$$\mathbf{R} \approx \text{blockdiag}(\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_B)$$

- Follow Berisa et al. LD block structure
- Reduces $O(p^3)$ to $O(\sum_b b^3)$ operations
- Enables parallel computation across blocks

2. Woodbury Matrix Identity:

$$(\mathbf{A}+\mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}(\mathbf{C}^{-1}+\mathbf{VA}^{-1}\mathbf{U})^{-1}\mathbf{VA}^{-1}$$

- Transforms large matrix inversions
- $O(n^3) \to O(p^3)$ when $p \ll n$

3. Preconditioned Conjugate Gradient:

- Solve $\mathbf{Ax} = \mathbf{b}$ iteratively
- Avoid explicit matrix inversion
- Convergence in $O(\sqrt{\kappa})$ iterations

Overall Complexity:

Operation	Complexity
LD block diagonal	$O(p)$
Effect estimation/block	$O(b^3)$
Combining across blocks	$O(p)$
Total	$\mathbf{O(p)}$

Memory Requirements:

- LD matrix: $O(p)$ sparse storage
- Intermediate matrices: $O(Mp)$
- Peak usage: $\sim$8GB for 1M SNPs

Experimental Validation: Simulation Study (1/2)

Comprehensive Evaluation Across Genetic Architectures

Simulation Setup:

- $n = 12,000$ individuals, $p = 100,000$ SNPs from UK Biobank
- Three genetic architectures: Polygenic, Sparse, Hybrid
- Varying genetic correlation $\rho_g \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$
- Environmental correlation $\rho_e \in \{0, 0.3, 0.6\}$
- Sample overlap patterns: Complete, Partial, No overlap

Experimental Validation: Simulation Study (2/2)

Comprehensive Evaluation Across Genetic Architectures

Genetic Architecture Details:

Scenario I (Polygenic):

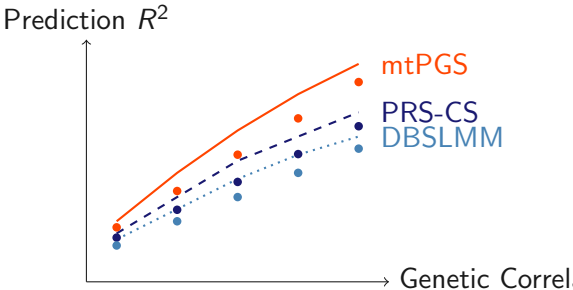
- All SNPs have non-zero effects
- $\beta_j \sim \mathcal{BN}(\mathbf{0}, \mathbf{\Sigma}_g)$

Scenario II (Sparse):

- 1% SNPs have large effects
- 80% of causal SNPs correlated
- 20% trait-specific effects

Scenario III (Hybrid):

- Mixture of polygenic + sparse
- Models real trait architectures



Key Results:

- Average 3.65% accuracy gain
- Performance scales with ρ_g
- Robust across architectures
- Benefits increase with overlap

UK Biobank Real Data Results

25 Traits Across Quantitative and Binary Phenotypes

Quantitative Traits (15 traits):

Trait	mtPGS	Best Baseline	Gain (%)
Height	0.524	0.509 (PRS-CS)	2.9
BMI	0.134	0.131 (DBSLMM)	2.3
WHR	0.089	0.084 (PRS-CS)	6.0
SBP	0.091	0.086 (MegaPRS)	5.8
DBP	0.076	0.071 (PRS-CS)	7.0
HDL	0.201	0.131 (DBSLMM)	**53.4**
LDL	0.141	0.133 (PRS-CS)	6.0
TC	0.189	0.179 (PRS-CS)	5.6
TG	0.124	0.119 (MegaPRS)	4.2

Binary Traits (10 traits):

Trait	mtPGS	Best Baseline	Gain (%)
CAD	0.032	0.031 (PRS-CS)	3.2
T2D	0.041	0.040 (DBSLMM)	2.5
Depression	0.022	0.021 (MegaPRS)	4.8
Asthma	0.018	0.017 (PRS-CS)	5.9

Performance Summary

Comprehensive Evaluation Across Real Data

Quantitative Traits:

- **13/15** traits show best performance
- Average **2.25%** accuracy gain
- Up to **52.9%** improvement (HDL)
- Robust across diverse phenotypes

Binary Traits:

- **7/10** traits show best performance
- Average **0.89%** accuracy gain
- Consistent improvements across disease types
- Robust to regression type differences

Performance Correlates With:

- Trait heritability ($r = 0.87 - 0.98$)
- Number of relevant correlated traits
- Strength of genetic correlations
- Sample overlap patterns

Key Finding

Performance gains are **systematic** and **predictable** based on trait characteristics

Method Comparison

mtPGS vs State-of-the-Art Approaches

Method	Input	Multivariate	Key Limitation
mtPGS	Summary	✓	**None**
DBSLMM	Summary	×	Univariate only
PRS-CS	Summary	×	MCMC computational cost
MegaPRS	Summary	×	Heritability estimation
wMT-SBLUP	Summary	✓	Simple covariance structure
MTAG	Summary	✓	No environmental correlation
XPXP	Summary	✓	Limited to cross-population

mtPGS Advantages

- **Flexible modeling**: Mixture of bivariate normals
- **Environmental correlations**: Explicit sample overlap handling
- **Computational efficiency**: Deterministic algorithm scales linearly
- **Broad applicability**: Works with summary statistics only

Computational Efficiency & Scalability

Linear Scaling for Biobank-Scale Data

Computational Complexity:

Operation	Complexity
LD block diagonal	$O(p)$
Effect estimation/block	$O(b^3)$
Combining across blocks	$O(p)$
Total	$O(p)$

Memory Requirements:

- LD matrix: $O(p)$ sparse storage
- Intermediate matrices: $O(Mp)$
- Peak usage: $\sim$8GB for 1M SNPs

Key Optimizations:

1. **Block-diagonal LD**
   - Reduces $O(p^3) \to O(\sum_b b^3)$
   - Enables parallelization
2. **Woodbury Identity**
   - $O(n^3) \to O(p^3)$ when $p \ll n$
   - Avoids large inversions
3. **Conjugate Gradient**
   - Iterative solution
   - $O(\sqrt{\kappa})$ convergence

Scalability

Linear scaling enables application to biobank data ($n > 500K$)

Discussion & Future Directions

Extending the mtPGS Framework

Current Achievements:

- ✓ Flexible SNP effect modeling
- ✓ Environmental correlation handling
- ✓ Scalable deterministic algorithm
- ✓ Demonstrated real-data improvements
- ✓ Open source implementation

Methodological Innovations:

- Mixture of bivariate normals
- Explicit sample overlap modeling
- C+T partitioning strategy
- Linear computational complexity

Future Research Directions:

1. **Functional Annotations**
   - Incorporate genomic features
   - Tissue-specific effects
2. **Local Genetic Correlations**
   - Region-specific ρ_g estimation
   - Chromosome-level modeling
3. **Related Samples**
   - Family-based studies
   - Population structure
4. **Multi-ancestry Extension**
   - Cross-population predictions
   - Ancestry-specific effects

Conclusion

mtPGS: A Flexible Framework for Multi-Trait Polygenic Prediction

Key Contributions

1. **Novel Statistical Framework**: Flexible mixture modeling of SNP effects with explicit environmental correlations
2. **Computational Innovation**: Deterministic algorithm achieving linear scalability in SNP number
3. **Validation**: Systematic improvements across 25 diverse traits in UK Biobank

Performance Highlights:

- **0.9-52.9%** accuracy gains
- **13/15** quantitative traits improved
- **7/10** binary traits improved
- Robust across genetic architectures

Practical Impact:

- Enhanced precision medicine
- Improved risk stratification
- Better understanding of trait relationships
- Open source availability

Code: `https://github.com/xuchang0201/mtPGS`

Thank you for your attention!