RESEARCH

New approach fits multivariate genomic prediction models efficiently

Alencar Xavier[1,2*†] and David Habier[1]

*Correspondence:
alencar.xavier@corteva.com
[1] Biostatistics, Corteva Agrisciences, 8305 NW 62nd Ave, 50131, Johnston, Iowa, USA
[2] Department of Agronomy,Purdue University, 915 W State St, 47907, West Lafayette, Indiana, USA
Full list of author information is available at the end of the article
†Equal contributor

Abstract

Background: Fast, memory-efficient, and reliable algorithms for estimating genomic estimated breeding values (GEBVs) for multiple traits and environments are needed to make timely decisions in plant breeding. Multivariate genomic prediction exploits genetic correlations among traits and environments to increase accuracy of GEBVs compared to univariate methods. These genetic correlations are estimated simultaneously with GEBVs, because they are specific to year, environment, and management. However, estimating genetic parameters is computationally demanding with Restricted Maximum Likelihood (REML) and Bayesian samplers for tens to hundreds of traits and environments. Techniques such as canonical transformation or orthogonalization cannot be used for unbalanced experimental designs that are common in plant breeding.

Methods: We propose a multivariate Randomized Gauss-Seidel algorithm for simultaneous estimation of model effects and genetic parameters. Two efficient methods for estimating genetic parameters, which were proposed earlier, were combined with a Gauss-Seidel (GS) solver. They were called *Tilde-Hat*-GS (THGS) and *Pseudo-Expectation*-GS (PEGS). Simulations of balanced and unbalanced experimental designs were used to study runtime, bias and accuracy of GEBVs, and bias and standard error of estimated heritabilities and genetic correlations. Statistics of THGS and PEGS were compared to those from REML. Multivariate models were evaluated that fitted 10 to 400 response variables, 1,279 to 42,034 markers, and 5,990 to 1,85 million observations.

Results: Runtime of PEGS and THGS was a fraction of REML. Models with 100 response variables ran under 30 minutes. Accuracies of GEBVs were slightly lower than those from REML, but higher than those from the univariate approach, which shows that THGS and PEGS exploited genetic correlations. For 500 to 600 observations per response variable, biases of heritability and genetic correlations of THGS and PEGS were small, but standard errors of genetic correlations were higher than for REML. Bias and standard error decreased as sample size increased. For balanced designs, GEBVs and estimated genetic correlations of THGS were unbiased when only an intercept was modeled, and either Principal components or eigenvectors of genotype scores were fitted.

Conclusions: THGS and PEGS are fast and memory-efficient algorithms for multivariate genomic prediction for both balanced and unbalanced experimental designs. They are scalable for increasing number of environments and markers. Bias of GEBVs is small and accuracy of GEBVs comparable to REML. Estimated genetic parameters have little bias, but their standard errors are larger than for REML. More studies are needed to evaluate the proposed methods for datasets that contain selection.

Keywords: Multi-Trait; Accuracy; Genetic correlation; Tilde-Hat

1 Background

Genomic prediction [1] uses genetic markers across the genome to predict complex diseases in humans and breeding values in animals and plants [2, 3]. Multivariate genomic prediction [4] exploits genetic correlations among response variables to increase prediction accuracy for each variable [5] compared to univariate analyses. In plant breeding, these response variables come from different quantitative traits that are measured in different field locations and years. Variance components and genetic correlations are estimated simultaneously with breeding values, because they vary across years, locations, and management. In animal breeding, in contrast, variance components are estimated infrequently within a breeding program and are used to solve mixed-model equations repeatedly over years.

The estimation of variances and covariances can be computationally demanding with standard multivariate approaches for trials with multiple quantitative traits and environments. In Restricted Maximum Likelihood (REML) analyses, large and dense mixed-model equations need to be stored in memory and inverted repeatedly. In Bayesian analyses, model effects need to be sampled for thousands of MCMC iterations. This becomes time-consuming with an increasing number of response variables, because increasingly large matrices need to be inverted and factorized in each iteration. Canonical transformation [6] or diagonalization of genomic relationship matrices [7] are only applicable to balanced experimental designs when individuals are phenotyped in all environments and for all quantitative traits. In plant breeding, however, unbalanced experimental designs are common. A solution would be to estimate genetic correlations for pairs of environments using bivariate models, but this also requires considerable computation resources. Moreover, the heritabilities of harvest yield are often low (0.1-0.2), so that precision of estimated variance components for yield can be increased by analyzing it together with higher heritable traits.

Fast and reliable algorithms are economically important in plant breeding enterprises to make timely decisions and advance the breeding pipeline. With any delays during harvest season, e.g., due to weather, only a few hours may be available for selection decisions. If a breeder misses a deadline to request either new breeding crosses from nurseries or seed of selected individuals or seed of test-crosses, the generation interval increases, genetic gain per year decreases, and product launches are delayed.

To speed up computations and provide estimated breeding values on time, we propose to combine a Randomized Gauss-Seidel [8, 9] solver for updating the effects of a multivariate model with an efficient approach for updating variances and covariances in each iteration of the algorithm. This approach calculates quadratic forms of random effects that resemble those used in REML but are equated to expectations that are easier to compute, as first proposed by [10, 11]. Similar approximations have been proposed over the years as depicted in [12], who compared their *Tilde-Hat* approach to methods of Schaeffer [13] and Henderson [14].

Statistical models that fit either a genomic relationship matrix or marker effects have been proposed for genomic prediction [2]. The latter is favored when the number of individuals exceeds the number of markers. In closed breeding programs, effective population sizes are such that a moderate number of markers, e.g. 10,000,

is sufficient to estimate breeding values using training datasets with a larger number of individuals, e.g. 100,000.

The objective of this study is to present and evaluate a multivariate ridge regression approach that uses jointly a Randomized Gauss-Seidel solver to estimate marker effects and the methods of either VanRanden [12] or Schaeffer [13] to estimate variances and covariances. Bias and accuracy of genomic estimated breeding values (GEBVs) and runtime are studied by simulation of different scenarios using a wheat dataset from CIMMYT's Global Wheat Program and a soybean dataset from the SoyNAM project. The proposed methods are compared to standard software implementations of REML and univariate analyses to show that the approximations harness the benefits of multivariate models for prediction accuracy. Bayesian Gibbs sampling was added to compare runtime. To understand and interpret differences in bias and accuracy of GEBVs between methods, bias and standard errors of estimated heritabilities and genetic correlations were evaluated.

2 Methods

2.1 Statistical model

The multivariate ridge regression model can be written as

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{b} + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{y}$ is a vector of phenotypes from K environments, which can be partitioned into $\boldsymbol{y}' = [\boldsymbol{y}'_1\ \boldsymbol{y}'_2\ \dots \boldsymbol{y}'_K]$, and each vector $\boldsymbol{y}'_k$ has length n_k; $\mathbf{X} = \oplus_{k=1}^{K}\mathbf{X}_k$, $\oplus$ denotes the direct sum operator, $\mathbf{X}_k$ is an $n_k \times r_k$ matrix with full column rank of r_k fixed effects; $\boldsymbol{b}' = [\boldsymbol{b}'_1\ \boldsymbol{b}'_2\ \dots \boldsymbol{b}'_K]$ is a vector of fixed effects for all environments, and each vector $\boldsymbol{b}'_k$ has length r_k; $\mathbf{Z} = \oplus_{k=1}^{K}\mathbf{Z}_k$, $\mathbf{Z}_k$ is an $n_k \times m$ matrix that contains marker scores of n_k individuals with phenotypes in environment k and m markers; $\boldsymbol{\beta}' = [\boldsymbol{\beta}'_1\ \boldsymbol{\beta}'_2\ \dots \boldsymbol{\beta}'_K]$ is an $(m \cdot K)$-vector of random marker effects for all environments, and each vector $\boldsymbol{\beta}'_k$ has length m; $\boldsymbol{e}' = [\boldsymbol{e}'_1\ \boldsymbol{e}'_2\ \dots \boldsymbol{e}'_K]$ is a vector of residuals, and each vector $\boldsymbol{e}'_k$ has length n_k. Marker effects are assumed multivariate-normal distributed with mean zero and variance-covariance matrix $Var(\boldsymbol{\beta}) = \boldsymbol{\Sigma}_\beta \otimes \mathbf{I}_m$, where $\boldsymbol{\Sigma}_\beta$ is a $K \times K$ matrix of genetic variances of marker effects, $\sigma^2_{\beta_k}$, on the diagonal, and genetic covariances between marker effects from different environments, $\sigma_{\beta_{kk'}}$, on the off-diagonal, $\otimes$ is the Kronecker product operator, and $\mathbf{I}_m$ is an identity matrix of dimension m. Residuals are assumed uncorrelated between environments, and normal distributed with mean zero and variance $Var(\boldsymbol{e}) = \oplus_{k=1}^{K}\mathbf{I}_k\sigma^2_{e_k}$.

2.2 Solving fixed effects and marker effects

The mixed-model equations can be written as

$$\begin{bmatrix} \mathbf{X}'_1\mathbf{X}_1\sigma^{-2}_{e_1} & \dots & \mathbf{0} & \mathbf{X}'_1\mathbf{Z}_1\sigma^{-2}_{e_1} & \dots & \mathbf{0} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{X}'_K\mathbf{X}_K\sigma^{-2}_{e_K} & \mathbf{0} & \dots & \mathbf{X}'_K\mathbf{Z}_K\sigma^{-2}_{e_K} \\ \mathbf{Z}'_1\mathbf{X}'_1\sigma^{-2}_{e_1} & \dots & \mathbf{0} & \mathbf{Z}'_1\mathbf{Z}_1\sigma^{-2}_{e_1} + \mathbf{I}_m\sigma^{11}_\beta & \dots & \mathbf{I}_m\sigma^{1K}_\beta \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \dots & \mathbf{Z}'_K\mathbf{X}'_K\sigma^{-2}_{e_K} & \mathbf{I}_m\sigma^{K1}_\beta & \dots & \mathbf{Z}'_K\mathbf{Z}_K\sigma^{-2}_{e_K} + \mathbf{I}_m\sigma^{KK}_\beta \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{b}}_1 \\ \vdots \\ \hat{\boldsymbol{b}}_k \\ \hat{\boldsymbol{\beta}}_1 \\ \vdots \\ \hat{\boldsymbol{\beta}}_K \end{bmatrix} = \begin{bmatrix} \sigma^{-2}_{e_1}\mathbf{X}'_1\boldsymbol{y}_1 \\ \vdots \\ \sigma^{-2}_{e_K}\mathbf{X}'_k\boldsymbol{y}_K \\ \sigma^{-2}_{e_1}\mathbf{Z}'_1\boldsymbol{y}_1 \\ \vdots \\ \sigma^{-2}_{e_K}\mathbf{Z}'_K\boldsymbol{y}_K \end{bmatrix},$$

where σ_β^{ij} is the element at position ij of $\mathbf{\Sigma}_\beta^{-1}$.

The iterative Gauss-Seidel method with residual updates, as presented in [15], was used to solve the mixed-model equations without setting them up explicitly, while updating variances and covariances in each iteration. We define $\hat{\boldsymbol{e}} = [\hat{\boldsymbol{e}}_1\ \hat{\boldsymbol{e}}_2\ \ldots\ \hat{\boldsymbol{e}}_K]$ to be the vector of estimated residuals, which is initialized as $\hat{\boldsymbol{e}}^{(0)} = [\boldsymbol{y}_1'\ \boldsymbol{y}_2'\ \ldots\ \boldsymbol{y}_K']$. The estimated fixed effect j of environment k is updated in iteration t by

$$\hat{b}_{jk}^{(t+1)} = \frac{\boldsymbol{x}_{jk}'\hat{\boldsymbol{e}}_k}{\boldsymbol{x}_{jk}'\boldsymbol{x}_{jk}},$$

and before moving to the next fixed effect, the residual vector is updated by

$$\hat{\boldsymbol{e}}_k^{(new)} = \hat{\boldsymbol{e}}_k^{(old)} - \boldsymbol{x}_{jk}\hat{b}_{jk}^{(t+1)}.$$

For updating estimated marker effects, we define $\hat{\dot{\boldsymbol{\beta}}}_j^{\prime(t)} = [\hat{\beta}_{j1}^{(t)}\ \hat{\beta}_{j2}^{(t)}\ \ldots\ \hat{\beta}_{jK}^{(t)}]$ to be the vector of estimated marker effects for marker j and all K environments in iteration t, $\dot{\mathbf{Z}}_j = \oplus_{k=1}^{K}\boldsymbol{z}_{jk}$ to be a matrix containing scores for marker j, $\boldsymbol{z}_{jk}$ to be an n_k column vector for scores at marker j and environment k, and $\hat{\mathbf{\Sigma}}_e^{(t)} = Diag\{\hat{\sigma}_{e_1}^{2(t)}, \hat{\sigma}_{e_2}^{2(t)}, \ldots, \hat{\sigma}_{e_K}^{2(t)}\}$ to be a diagonal matrix of estimated residual variances from all environments. Estimated effects for marker j are initialized to zero and updated by

$$\hat{\dot{\boldsymbol{\beta}}}_j^{(t+1)} = (\hat{\mathbf{\Sigma}}_e^{-1(t)}\dot{\mathbf{Z}}_j'\dot{\mathbf{Z}}_j + \hat{\mathbf{\Sigma}}_\beta^{-1(t)})^{-1}\hat{\mathbf{\Sigma}}_e^{-1(t)}\dot{\mathbf{Z}}_j'(\dot{\mathbf{Z}}_j\hat{\dot{\boldsymbol{\beta}}}_j^{(t)} + \hat{\boldsymbol{e}}), \tag{2}$$

and before moving to the next marker, the residual vector is updated as

$$\hat{\boldsymbol{e}}^{(new)} = \hat{\boldsymbol{e}}^{(old)} - \dot{\mathbf{Z}}_j'(\hat{\dot{\boldsymbol{\beta}}}_j^{(t+1)} - \hat{\dot{\boldsymbol{\beta}}}_j^{(t)}).$$

The term $\hat{\mathbf{\Sigma}}_e^{-1(t)}\dot{\mathbf{Z}}_j'\dot{\mathbf{Z}}_j$ of equation (2) is a $K \times K$ diagonal matrix with elements $\{\hat{\sigma}_{e1}^{-2(t)}\boldsymbol{z}_{j1}'\boldsymbol{z}_{j1}, ..., \hat{\sigma}_{eK}^{-2(t)}\boldsymbol{z}_{jK}'\boldsymbol{z}_{jK}\}$, and the term $\hat{\mathbf{\Sigma}}_e^{-1(t)}\dot{\mathbf{Z}}_j'(\dot{\mathbf{Z}}_j\hat{\dot{\boldsymbol{\beta}}}_j^{(t)} + \hat{\boldsymbol{e}})$ can be computed as a vector of length K with elements $[\hat{\sigma}_{e_1}^{-2(t)}(\boldsymbol{z}_{j1}'\boldsymbol{z}_{j1}\hat{\beta}_{j1}^{(t)} + \boldsymbol{z}_{j1}'\hat{\boldsymbol{e}}_1), ..., \hat{\sigma}_{e_K}^{-2(t)}(\boldsymbol{z}_{jK}'\boldsymbol{z}_{jK}\hat{\beta}_{jK}^{(t)} + \boldsymbol{z}_{jK}'\hat{\boldsymbol{e}}_K)]$. Values of $\boldsymbol{z}_{jk}'\boldsymbol{z}_{jk}$ are calculated before iterations start for all combinations of markers (j) and environments (k).

To increase convergence rate, the order in which the marker effects are updated is randomized in each iteration. This approach is referred to as Randomized Gauss-Seidel [8, 9].

2.3 Solving variances and covariances

Genetic variances and covariances were updated by using the method proposed by either [12] or [13], called *Tilde-Hat* (TH) and *Pseudo Expectation* (PE), respectively. Both methods use the quadratic form $\tilde{\boldsymbol{\beta}}_k^{\prime(t)}\hat{\boldsymbol{\beta}}_k^{(t)}$, where $\hat{\boldsymbol{\beta}}_k^{(t)}$ contains all estimated marker effects for environment k in iteration t, and

$$\tilde{\boldsymbol{\beta}}_k^{(t)} = \mathbf{D}_k^{-1(t)}\mathbf{Z}_k'\mathbf{M}_k\boldsymbol{y}_k. \tag{3}$$

The two methods differ in matrix $\mathbf{D}_k^{-1(t)}$: In PE, $\mathbf{D}_k^{(t)} = \mathbf{I}_m$, whereas in TH,

$$\mathbf{D}_k^{(t)} = Diag\{\mathbf{Z}_k'\mathbf{M}_k\mathbf{Z}_k\hat{\sigma}_{e_k}^{-2(t)} + \mathbf{I}_m\hat{\sigma}_{\beta}^{kk(t)}\}, \tag{4}$$

which denotes a diagonal matrix, and $\mathbf{M}_k = \mathbf{I}_k - \mathbf{X}_k(\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k'$. As $\mathbf{D}_k^{(t)}$ is diagonal, $\mathbf{M}_k$ does not have to be explicitly generated, but only the diagonal of $\mathbf{Z}_k'\mathbf{M}_k\mathbf{Z}_k$ needs to be computed once before iterations start and stored. This computation can be done efficiently as shown in Appendix A. When the intercept is the only fixed effect, and both $\boldsymbol{y}_k$ and the columns of $\mathbf{Z}_k$ are centered, then $\mathbf{M}_k$ can be omitted.

The estimated genetic and residual variances for environment k were initialized to $\hat{\sigma}_{\beta_k}^{2(0)} = 0.5 \cdot \sigma_{y_k}^2/(m \cdot \overline{\sigma^2}_{Z_k})$ and $\hat{\sigma}_{e_k}^{2(0)} = 0.5 \cdot \sigma_{y_k}^2$, respectively, where $\sigma_{y_k}^2$ is the sample variance of phenotypes and $\overline{\sigma^2}_{Z_k} = \frac{1}{m}\sum_{j=1}^{m}\sigma_{Z_{k_j}}^2$ is the average of marker-score variances across the m columns of $\mathbf{Z}_k$. Estimated genetic covariances were initialized to zero. The estimated variance of marker effects for environment k is updated by

$$\hat{\sigma}_{\beta_k}^{2(t+1)} = \frac{\tilde{\boldsymbol{\beta}}_k^{'(t)}\hat{\boldsymbol{\beta}}_k^{(t)}}{tr(\mathbf{D}_k^{-1(t)}\mathbf{Z}_k'\mathbf{M}_k\mathbf{Z}_k)}, \tag{5}$$

where $\mathbf{Z}_k$ contains marker scores for environment k, $tr(\cdot)$ is the trace operator, and $tr(\mathbf{D}_k^{-1(t)}\mathbf{Z}_k'\mathbf{M}_k\mathbf{Z}_k)$ is the expected value of $\tilde{\boldsymbol{\beta}}_k^{'(t)}\hat{\boldsymbol{\beta}}_k^{(t)}$ as derived in [12] and in Appendix B. The estimated covariance between environments k and k' is updated by

$$\hat{\sigma}_{\beta_{kk'}}^{(t+1)} = \frac{\tilde{\boldsymbol{\beta}}_k^{'(t)}\hat{\boldsymbol{\beta}}_{k'}^{(t)} + \tilde{\boldsymbol{\beta}}_{k'}^{'(t)}\hat{\boldsymbol{\beta}}_k^{(t)}}{tr(\mathbf{D}_k^{-1(t)}\mathbf{Z}_k'\mathbf{M}_k\mathbf{Z}_k) + tr(\mathbf{D}_{k'}^{-1(t)}\mathbf{Z}_{k'}'\mathbf{M}_{k'}\mathbf{Z}_{k'})}, \tag{6}$$

as proposed by [13] and derived in section 1 of the supplement, and residual variances are updated by

$$\hat{\sigma}_{e_k}^{2(t+1)} = \frac{(\mathbf{M}_k\boldsymbol{y}_k)'\hat{\boldsymbol{e}}_k}{n_k - r_k} \tag{7}$$

as in [15], where r_k is the number of linear independent columns of $\mathbf{X}_k$.

Bending [16] of $\hat{\Sigma}_\beta$ was used after an iteration when it was not positive definite. The iterative scheme is repeated until mean-squared convergence of 10^{-8} is reached for effects, variances, and covariances. The combination of the Randomized Gauss-Seidel solver with either of the two methods for variance component estimation, i.e., TH or PE, is referred to here as THGS and PEGS, respectively. An implementation of PEGS is provided in the R package bWGR (2.0), function **mrr** [17], and is shown in section 6 of the supplement.

2.4 Exact THGS

For balanced experimental designs, when the intercept is the only fixed effect, and either a Principal components [18] or eigenvector regression [19, 20, 21] is used,

THGS is exact. This is demonstrated in Appendix C. By either using a singular-value decomposition of $\mathbf{Z}_k$ or an eigenvalue decomposition (EVD) of $\mathbf{Z}_k'\mathbf{Z}_k$, a matrix of eigenvectors, $\mathbf{U}_k$, and a diagonal matrix of eigenvalues, $\mathbf{\Lambda}_k$, can be calculated. By fitting $\check{\mathbf{Z}}_k = \mathbf{Z}_k\mathbf{U}_k$ rather than $\mathbf{Z}_k$ in model (1), $\mathbf{Z}_k'\mathbf{M}_k\mathbf{Z}_k$ in equation (4) becomes a diagonal matrix of eigenvalues, $\mathbf{\Lambda}_k$. Thus, $\mathbf{D}_k^{(t)}$ in equations (5) and (6) can be written as

$$\mathbf{D}_k^{(t)} = \mathbf{\Lambda}_k\hat{\sigma}_{e_k}^{-2(t)} + \mathbf{I}_m\hat{\sigma}_\beta^{kk(t)}. \tag{8}$$

This does not apply to PEGS, because it uses $\mathbf{D}_k^{(t)} = \mathbf{I}_m$.

2.5 Alternative methods

As a gold standard for low bias and standard error of both GEBVs and variance components, Empirical Genomic Best Linear Unbiased Predictions (GBLUP) [22] were obtained by REML [23] for balanced experimental designs as follows. The genomic relationship matrix ($\mathbf{G}$) was diagonalized and the statistical model was transformed by the eigenvectors of an eigenvalue decomposition of $\mathbf{G}$ [7] (see Appendix D). Eigenvectors of the smallest eigenvalues, which explained the last 1% of the variation in $\mathbf{G}$ were neglected [24]. The transformed model was evaluated by ASREML-R [25]. For unbalanced experimental designs, neither ASREML 4.2 nor AIREMLF90 nor REMLF90 returned results for the full multivariate models in this simulation study. Thus, to obtain an upper bound of accuracy of GEBVs, GBLUPs were calculated using the true simulated varaince components. This method was called True Value-Gauss-Seidel (TVGS).

Runtimes of the proposed and other methods were compared only for the balanced designs. In addition to the REML approach described above, $\mathbf{G}$ was used in its natural, dense form and 0.01 was added to its diagonal to render it positive definite. The Expectation Maximization (EM) REML algorithm of REMLF90 [26] and the Average Information (AI) REML algorithms of ASREML 4.2 [23, 25] and AIREMLF90 [27] were used with their options for dense equations operations *!gdense* and *use_yams*, respectively. Additionally, the Gibbs sampler of GIBBSF90 was run for comparison.

Univariate THGS (UV-THGS), which analyzes phenotypes of only one environment at a time with the Randomised Gauss-Seidel solver and TH, was run to evaluate the increase in accuracy of GEBVs with multivariate THGS over univariate THGS. Table 1 summarizes the methods in this study.

Table 1: Summary of methods.

	TVGS	PEGS	THGS	UV-THGS	REML
Effect type in the model	Marker	Marker	Marker	Marker	Polygenic
Multivariate	Yes	Yes	Yes	No	Yes
(Co)variance estimation*	True values	PE	TH	TH	REML
Orthogonalization	No	No	No	No	Yes

*PE: Pseudo Expectation; TH: Tilde-Hat.

2.6 Data and evaluation statistics

Phenotypic data for five scenarios were simulated to evaluate bias and accuracy of estimated genomic breeding values (GEBVs) within environments, runtime, and bias and standard error of estimated heritabilities and genetic correlations (Table 2). The genotypes used in the simulations come from a wheat [28, 29, 30, 31] and a soybean dataset [32, 33, 21], which have been used in multiple genomic prediction studies, and are available through the R packages BGLR and SoyNAM, respectively.

Scenario 1 contained simulated phenotypes from individuals that are all grown in the same ten environments, using 599 inbred lines from CIMMYT's Global Wheat Program [28, 29] genotyped at 1,279 DArT markers [34]. **Scenario 2** contained simulated phenotypes from different individuals grown in ten different environments, using 5,142 recombinant inbred lines from the SoyNAM project [35, 36] genotyped with 4,311 Single Nucleotide Polymorphism (SNP) markers. These lines were randomly split into ten different environments, and each line was observed in only a single environment. **Scenario 3** was used to study the evaluation statistics for an increasing number of soy inbred lines in each of the ten environments. Thus, each line could be present in multiple environments. **Scenario 4** was used to study runtime of PEGS and THGS for an increasing number of environments (response variables), i.e., 10, 50, 100, 200 and 400, using the SoyNAM dataset with 10% missing individuals at random in each environment. **Scenario 5** was used to study runtime with higher marker density, using the SoyNAM dataset and 42,034 SNPs that were obtained from the original SNPs plus a linkage disequilibrium-based imputation of SNPs as described in [36].

Table 2: Summary of simulated scenarios.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5
No. of environments (traits)	10	10	10	10-400	10-400
No. of environments per individual	10	1	0-10	0-400	0-400
No. of individuals per environment	599	514	250-3,000	4,628	4,628
% of individuals per environment	100%	10%	5-60%	90%	90%
No. of phenotypic records	5,990	51,420	30,000	1,851,120	1,851,120
No. of markers	1,279	4,311	4,311	4,311	42,034
Species	Wheat	Soy	Soy	Soy	Soy

Phenotypes were simulated by adding true genomic breeding values (TBVs) to residuals. TBVs of environment k were sampled as $\mathbf{Z}\boldsymbol{\beta}_k$, where $\mathbf{Z}$ contains marker scores of inbred lines from all environments and the true marker effects in $\boldsymbol{\beta}_k$ were taken from $\boldsymbol{\beta}' = [\boldsymbol{\beta}'_1\ \boldsymbol{\beta}'_2\ \ldots \boldsymbol{\beta}'_K]$. This vector was sampled from $N(\mathbf{0}, \boldsymbol{\Sigma}_\beta \otimes \mathbf{I}_m)$, where $\boldsymbol{\Sigma}_\beta = \alpha^{-1}\boldsymbol{\Sigma}_g$, $\alpha = \sum_{j=1}^{J} \sigma^2_{Z_j}$, $\sigma^2_{Z_j}$ is the variance of marker scores in $\mathbf{Z}$ at marker j, and $\boldsymbol{\Sigma}_g$ is the additive genetic variance-covariance matrix with 1 on the diagonal and genetic correlations on the off-diagonals. Residuals were sampled from $N(0, (1-h^2)h^{-2})$, where h^2 is the heritability in an environment. Three heritabilities (0.2, 0.5, and 0.8) and three ranges of genetic correlations, low (0.2-0.4), medium (0.4-0.6), and high (0.6-0.8) were considered. Correlations were sampled from a uniform distribution within each range. Each simulation scenario was replicated 100 times.

Bias and standard error of estimated heritabilities and genetic correlations were calculated as average and standard deviation, respectively, of estimated minus true

simulated values across replicates. GEBVs of environment k were calculated as $\mathbf{Z}_k\hat{\boldsymbol{\beta}}_k$, bias and accuracy of these GEBVs were calculated as the regression coefficient of TBV on GEBV and correlation between TBV and GEBV, respectively.

3 Results

3.1 Runtime

Average runtime of the different methods used in scenario 1 is presented in table 3. Multivariate PEGS and THGS took 0.4 and 0.3 seconds, respectively, univariate THGS aggregated across ten environments 0.2 seconds, and AI-REML using ASREML-R 3.3 seconds when the genomic relationship matrix was diagonalized by eigenvalue decomposition. Standard implementations of REML based on the dense genomic relationship matrix ranged from 109.8 to 1,250.7 seconds, whereas the Gibbs sampler took 559.8 seconds.

Table 3: Average runtime in seconds (s.e.) of the balanced experimental design in scenario 1 based on 100 replicates of the simulation.

Method	Software	Model[1]	Runtime
PEGS	-	RR	0.4 (0.0)
THGS	-	RR	0.3 (0.0)
UV-THGS	-	RR	0.2 (0.0)
AI-REML (EVD)[2]	ASREML-R	GBLUP	3.3 (0.3)
AI-REML	ASREML 4.2	GBLUP	272.6 (36.5)
AI-REML	AIREMLF90	GBLUP	109.8 (2.4)
EM-REML	REMLF90	GBLUP	1,250.7 (11.7)
Gibbs sampling[3]	GIBBS3F90	GBLUP	559.8 (9.6)

[1]RR: Ridge-Regression; GBLUP: Genomic Best Linear Unbiased Prediction.
[2] Eigenvalue decomposition (EVD). [3] 10,000 MCMC iterations.

Figure 1 shows convergence of the Gauss-Seidel solver with and without randomizing the order in which marker effects are updated for one replicate of scenario 2. The algorithm converged after 54 iterations with randomization, but required more than 3,000 iterations without randomization.

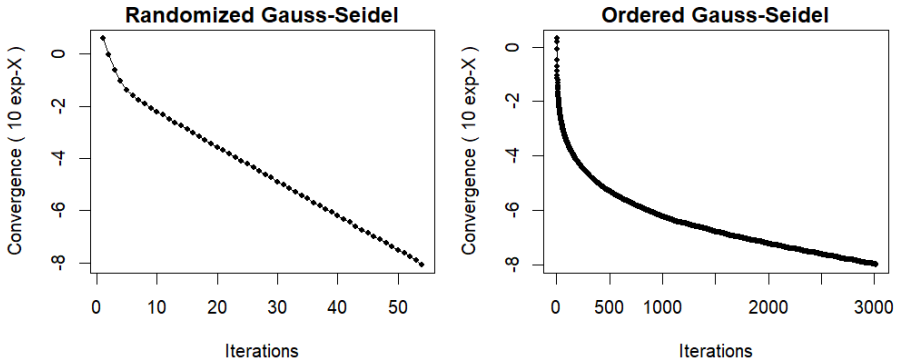

Figure 1: Convergence of the Gauss-Seidel solver with (left) and without (right) randomizing the order in which marker effects are updated for one replicate of the simulation of scenario 2.

Table 4 depicts average runtime in minutes for PEGS, THGS, and UV-THGS with and without randomizing the marker order in the Gauss-Seidel solver as well

as an increasing number of environments (scenario 4) and markers (scenario 5). PEGS and THGS had similar runtimes that were lower with randomization. Without randomization, the multivariate models that fitted 42,034 SNPs did not converge within 2,000 iterations. Runtimes of PEGS and THGS increased exponentially with number of environments from 0.2 minutes for ten environments to 448 minutes for 400 environments using 4,311 SNPs. Runtime of UV-THGS, in contrast, increased linearly from 0.1 to 4.3 minutes under the same conditions. With randomization, runtime increased with increasing number of markers from 0.2 minutes for 4,311 SNPs to 0.8 minutes for 42,034 SNPs and ten environments, and from 80.5 to 123.2 minutes for 200 environments. Without randomization, runtime increased to 3,057.3 minutes for 42,034 SNPs and 200 environments.

Table 4: Average runtime in minutes (s.e.) of the Gauss-Seidel solver with and without randomizing the order of markers for updating marker effects, with increasing number of SNPs and environments (envir.), and based on 10 replicates of scenarios 4 (4,311 SNPs) and 5 (42,034 SNPs).

Randomized	No. of SNPs	No. of envir.	PEGS	THGS	UV-THGS
Yes	4,311	10	0.2 (0)	0.2 (0)	0.1 (0)
Yes	4,311	50	3.5 (0.4)	3.5 (0.4)	0.6 (0)
Yes	4,311	100	14.4 (2)	14.4 (1.8)	1.1 (0)
Yes	4,311	200	80.5 (10.1)	79.2 (11)	2.3 (0.1)
Yes	4,311	400	459.3 (55.1)	448 (58)	4.3 (0.1)
No	4,311	10	5.5 (1)	5.4 (0.9)	1.9 (0.2)
No	4,311	50	44.9 (7)	44.6 (6.9)	9.3 (1.1)
No	4,311	100	120.9 (10.1)	123.7 (9.9)	20 (1.8)
No	4,311	200	361.1 (48.9)	364.6 (44.4)	39.3 (2.8)
No	4,311	400	1,261.8 (115.8)	1,261.7 (107.9)	74.1 (8.3)
Yes	42,034	10	0.8 (0.1)	0.8 (0)	1.2 (0.1)
Yes	42,034	50	9.9 (0.4)	12.5 (1.3)	5.7 (0.4)
Yes	42,034	100	36.4 (1.4)	29.2 (2.7)	11.3 (0.6)
Yes	42,034	200	123.2 (17.1)	119.7 (10.1)	22.5 (2)
Yes	42,034	400	730 (64.4)	802.2 (118.2)	46.4 (4.1)
No	42,034	10	64* (14.7)	64.2* (16)	14.5 (5.1)
No	42,034	50	540.2* (38.3)	536* (26.8)	106.5 (63.2)
No	42,034	100	1,109.6* (71.5)	1,148.1* (109.3)	181.4 (40.6)
No	42,034	200	3,057.3* (292.7)	3,001.2* (259)	310.3 (114.8)

* Did not converge within 2,000 iterations.

3.2 Accuracy and bias of GEBVs

Accuracy of GEBVs increased with increasing heritability and genetic correlation, as expected (Figure 2). It was 0.03-0.09 higher for multivariate approaches than univariate THGS when heritability was low and genetic correlation medium to high (Figure 2, a and b, lower left panels). For most genetic parameters in scenario 1, REML provided 0.01 higher accuracy than PEGS and THGS. For low heritability and low genetic correlations, however, REML had 0.02 higher accuracy and UV-THGS was as accurate as PEGS and THGS (Figure 2a, upper left panel). The latter was also true for scenario 2. Upon further simulations of scenario 1 for low heritability and low genetic correlations, accuracies of PEGS and THGS became larger than UV-GS and approached those of REML with increasing number of environments (Section 2 of the supplement). Compared to TVGS, even REML tended

to have lower accuracies for low heritability and low genetic correlation (Figure 2a, upper left panel). The differences between TVGS and both PEGS and THGS were similar in scenarios 1 and 2 (Figure 2, a vs. b). PEGS and THGS were not significantly different in scenarios 1 and 2.

Regression coefficients of TBV on GEBV are shown in Figure 3. For scenario 1 and low heritability, it was 1 for PEGS and THGS, close to 1 for REML, and significantly above 1 for UV-THGS. This bias for UV-THGS decreased with increasing heritability. For medium to high heritabilities, however, PEGS and THGS slightly underestimated (values > 1) the TBVs, whereas REML was usually unbiased with value 1 (Figure 3a). The bias for PEGS and THGS decreased with increasing genetic correlation. In scenario 2 (Figure 3b), PEGS and THGS slightly overestimated TBVs (values <1) for low heritability, but slightly underestimated TBVs (values >1) for medium to high heritabilities.

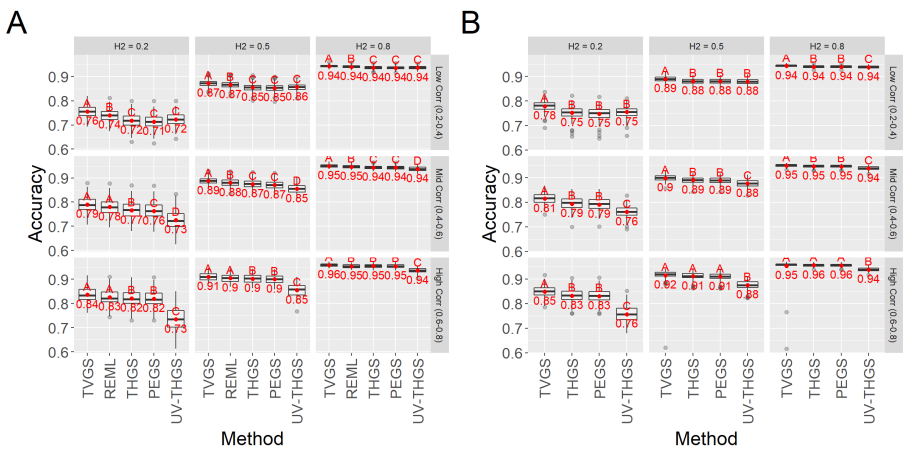

Figure 2: Accuracy of GEBVs for scenario 1 (a, wheat dataset) and scenario 2 (b, soybean dataset) for different true heritabilities (columns) and genetic correlations (rows), based on 100 replicates of the simulation. Letters indicate Tukey's test of multiple comparisons ($\alpha = 0.05$).

3.3 Bias and standard error of estimated parameters

Figure 4 shows bias of estimated heritabilities for scenarios 1 and 2 and different genetic parameters. In both scenarios, heritabilities tended to be downward biased. The bias of PEGS and THGS was smallest or even zero for low heritability and medium to high genetic correlations (Figure 4, bottom left panels). Their biases decreased with increasing genetic correlations. The bias of UV-THGS tended to be lower than PEGS and THGS. REML provided the least biased heritability estimates in scenario 1.

Figure 5 shows standard errors of estimated heritabilities for scenarios 1 and 2 and different genetic parameters. Standard errors were higher for scenario 1 than 2, higher for medium heritability than low and high heritabilities, highest for low genetic correlations, and decreased with increasing genetic correlation. They were between 60-100% higher for PEGS and THGS than for REML.

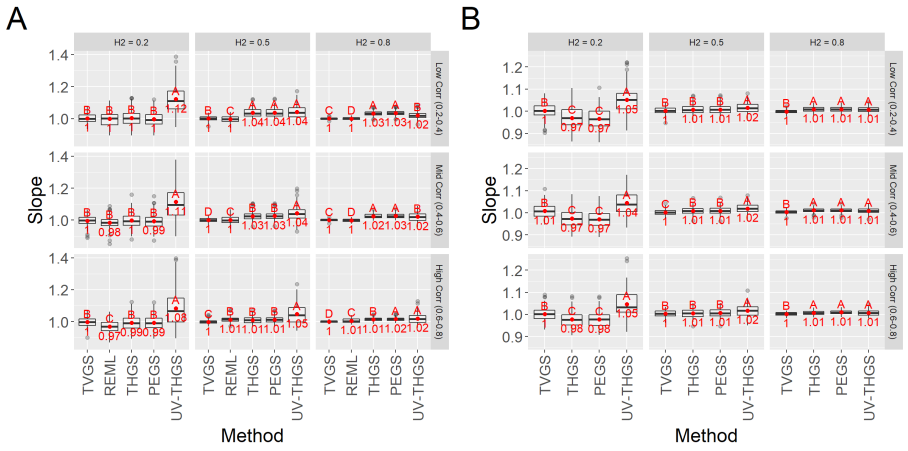

Figure 3: Regression of true breeding values on GEBVs (slope) for scenario 1 (a, wheat dataset) and scenario 2 (b, soybean dataset) for different true heritabilities (columns) and genetic correlations (rows), based on 100 replicates of the simulation. Letters indicate Tukey's test of multiple comparisons ($\alpha = 0.05$).

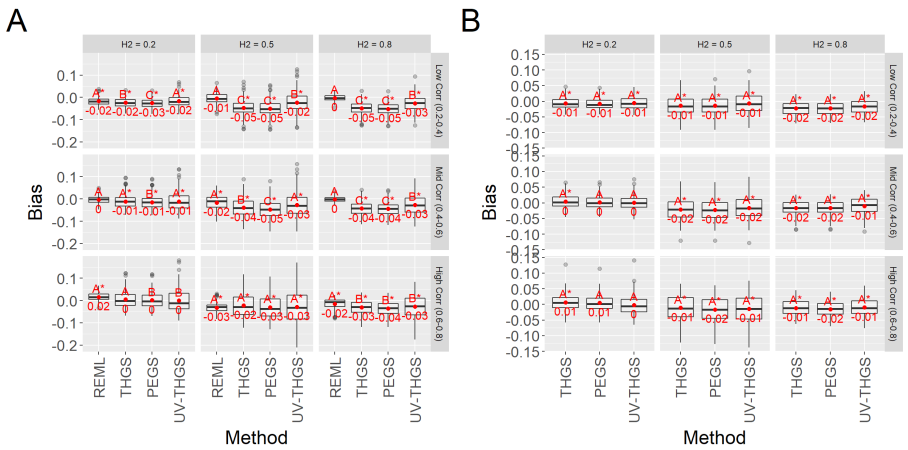

Figure 4: Bias of estimated heritability for scenario 1 (a, wheat dataset) and scenario 2 (b, soybean dataset) for different true heritabilities (columns) and true genetic correlations (rows), based on 100 replicates of the simulation. Letters indicate Tukey's test of multiple comparisons ($\alpha = 0.05$). Asterisk indicates that the mean is significantly different from zero ($\alpha = 0.05$).

Figures 6 and 7 show bias and standard error of estimated genetic correlations for scenarios 1 and 2. Bias tended to be low for PEGS and THGS in scenario 2, except for low heritability and high genetic correlations (Figure 6b, lower left panel). In scenario 1 and for high genetic correlations (Figure 6a, lower left panel), REML had large biases with absolute values of up to 0.08, compared to 0.01 for THGS. Otherwise, REML and the proposed methods had similar biases, and they were not significantly different for PEGS and THGS. As standard software for REML did not return results for the full model and the unbalanced designs in scenario 2, bivariate models were ran and estimated genetic correlations are given in section 3 of the supplement.

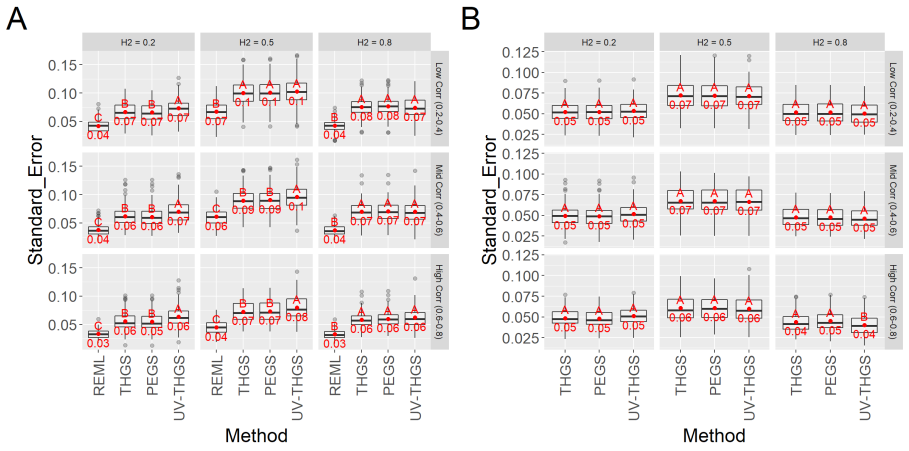

Figure 5: Standard error of estimated heritability for scenario 1 (a, wheat dataset) and scenario 2 (b, soybean dataset) for different true heritabilities (columns) and true genetic correlations (rows), based on 100 replicates of the simulation. Letters indicate Tukey's test of multiple comparisons ($\alpha = 0.05$).

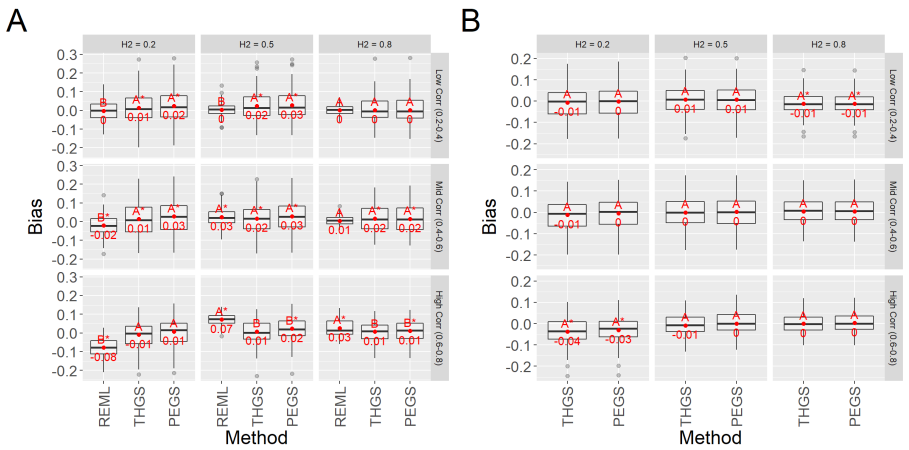

Figure 6: Bias of estimated genetic correlations for scenario 1 (a, wheat dataset) and scenario 2 (b, soybean dataset) for different true heritabilities (columns) and true genetic correlations (rows), and based on 100 replicates of the simulation. Letters indicate Tukey's test of multiple comparison ($\alpha = 0.05$). Asterisk indicates that the mean is significantly different from zero ($\alpha = 0.05$).

Standard errors of estimated genetic correlations decreased with increasing heritability and genetic correlations (Figure 7). Values of PEGS and THGS were always similar, but higher than for REML for low to medium genetic genetic correlations. For high genetic correlations, standard errors were similar for all methods. Standard errors were lower for scenario 2 than 1.

As the number of observations per environment increased in scenario 3, standard errors of estimated genetic parameters decreased, bias of estimated genetic correlations decreased, but bias of heritabilities did not approach zero even with 3,000 observations per environment (Table 5). Section 4 of the supplement demonstrates the outcome when all 5,142 individuals were observed in all environments: heritabil-

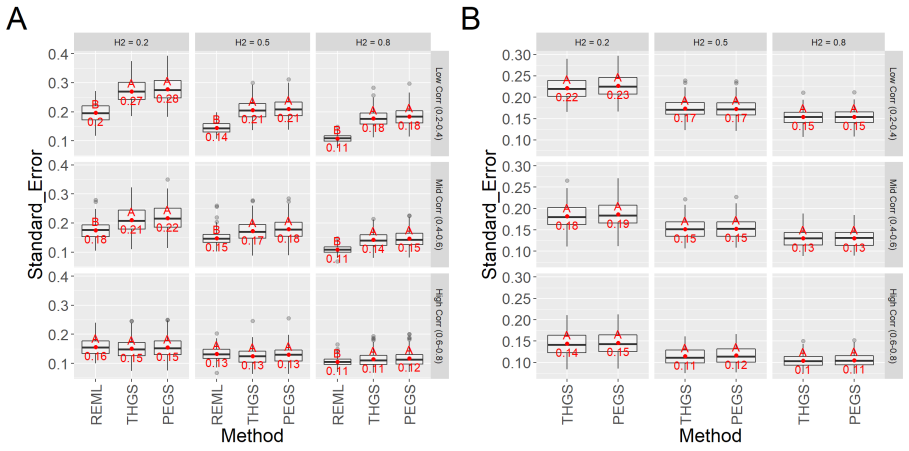

Figure 7: Standard error of estimated genetic correlations for scenario 1 (a, wheat dataset) and scenario 2 (b, soybean dataset) for different true heritabilities (columns) and true genetic correlations (rows), based on 100 replicates of the simulation. Letters indicate Tukey's test of multiple comparisons ($\alpha = 0.05$). Asterisk indicates that the mean is significantly different from zero ($\alpha = 0.05$).

ities estimated with THGS were unbiased, and genetic correlations estimated with PEGS or THGS were unbiased.

Table 5: Accuracy of GEBVs, regression of TBV on GEBV (Slope), and bias and standard error (SE) of estimated heritabilities ($\hat{h}^2$) and genetic correlations (GC) with increasing number of observations per environment (Obs/Env) in scenario 3, based on 100 replicates of the simulation. Standard errors of statistics are in parenthesis.

Method	Obs/Env	Accuracy	Slope	Bias of $\hat{h}^2$	SE of $\hat{h}^2$	Bias of GC	SE of GC
PEGS	250	0.82 (0.03)	0.98 (0.03)	-0.01 (0.03)	0.07 (0.01)	-0.01 (0.06)	0.17 (0.02)
PEGS	3000	0.96 (0.03)	1.00 (0.03)	-0.01 (0.03)	0.04 (0.01)	0.00 (0.06)	0.13 (0.02)
THGS	250	0.82 (0.03)	0.98 (0.04)	0.00 (0.03)	0.07 (0.01)	-0.02 (0.06)	0.17 (0.02)
THGS	3000	0.96 (0.03)	1.00 (0.03)	-0.01 (0.03)	0.04 (0.01)	0.00 (0.06)	0.13 (0.02)
UV-THGS	250	0.79 (0.03)	1.04 (0.03)	-0.01 (0.03)	0.07 (0.01)	-	-
UV-THGS	3000	0.95 (0.03)	1.00 (0.04)	-0.01 (0.03)	0.04 (0.01)	-	-

3.4 Orthogonalization

Table 6 presents bias and accuracy of GEBVs as well as bias and standard error of estimated genetic parameters with and without using eigenvalue decomposition (EVD). THGS-EVD provided unbiased GEBVs (Slope = 1) and its accuracy was 0.01 higher than THGS and thus equal to the accuracy of REML. Estimated genetic correlations of THGS-EVD were unbiased and had lower standard errors than THGS. The accuracy of GEBVs from UV-THGS-EVD did not increase compared to UV-THGS, suggesting that the increase of accuracy for THGS-EVD resulted from a higher precision in estimating genetic correlations. PEGS and PEGS-EVD were not different.

Table 6: Accuracy of GEBVs, regression of TBV on GEBV (Slope) as well as bias and standard error (SE) of estimated heritabilities ($\hat{h}^2$) and genetic correlations (GC) with eigenvalue decomposition (EVD) and without, based on 100 replicates of the simulation of scenario 1.

Method	Accuracy	Slope	Bias of $\hat{h}^2$	SE of $\hat{h}^2$	Bias of GC	SE of GC
REML-EVD	0.87 (0.02)	1.00 (0.03)	-0.01 (0.02)	0.04 (0.01)	0.00 (0.04)	0.14 (0.03)
PEGS	0.86 (0.02)	1.02 (0.03)	-0.03 (0.04)	0.07 (0.02)	0.02 (0.08)	0.18 (0.04)
PEGS-EVD	0.86 (0.02)	1.02 (0.03)	-0.04 (0.04)	0.07 (0.02)	0.02 (0.08)	0.18 (0.04)
THGS	0.86 (0.02)	1.02 (0.03)	-0.03 (0.04)	0.07 (0.02)	0.01 (0.08)	0.17 (0.04)
THGS-EVD	0.87 (0.02)	1.00 (0.03)	-0.02 (0.03)	0.05 (0.01)	0.00 (0.04)	0.13 (0.02)
UV-THGS	0.84 (0.04)	1.06 (0.09)	-0.02 (0.05)	0.08 (0.02)	-	-
UV-THGS-EVD	0.84 (0.03)	1.03 (0.04)	-0.03 (0.03)	0.05 (0.01)	-	-

4 Discussion

Our main goal was to find an algorithm for multivariate genomic prediction, which is efficient in runtime and memory, applicable to unbalanced experimental designs, and exploits genetic correlations between environments to increase accuracy of GEBVs compared to univariate analyses. We proposed two algorithms, PEGS and THGS, that use Randomized Gauss-Seidel to solve marker effects and simultaneously estimate variance components with methods developed by [13] and [12], respectively. Simulations were conducted to evaluate bias and accuracy of GEBVs within environment and compare them to those obtained by REML and a univariate approach. Bias and standard error of estimated heritabilities and genetic correlations were also evaluated to interpret the differences in bias and accuracy of GEBVs between methods (Table 1).

PEGS and THGS are fast and memory-efficient algorithms for both balanced and unbalanced experimental designs, and had much lower runtime than REML using standard software implementations (Tables 3 and 4). Moreover, they are scalable with number of environments and markers. The reasons for the speed-up are that equations are solved by Randomized Gauss-Seidel and that expectations of quadratic forms, shown in the denominator of equations (5) and (6), are inexpensive to compute. These expectations do not require elements of the inverse of the left-hand side of the mixed-model equations as shown in [13]. Therefore, the system of equations essentially reduces to a $K \times K$ problem (equation 2) with complexity $O(K^3)$. When hundreds to thousands of response variables were to be fitted at once, it is possible to linearize operations through a full-conditional multivariate Gauss-Seidel algorithm presented in Appendix E.

The number of iterations to convergence (Figure 1) and thereby runtime of PEGS and THGS decreased greatly by randomizing the marker order for updating marker effects (Table 4). This may be due to reducing dependencies among markers that stem from high linkage disequilibrium between adjacent markers on the same chromosome. With an increasing number of environments and markers, PEGS and THGS have reasonably low runtimes (Table 4, with randomization), which allows breeders to make decisions on time, and rerun genetic evaluations as data become available during harvest season.

For balanced designs, the number of iterations to convergence can be further reduced by modeling the eigenvectors of genotype scores, which completely removes

dependencies among model effects. In addition, THGS becomes an exact method that yields unbiased estimates of genetic correlations and GEBVs (section 2.4), and reduces the bias of estimated heritabilities as can be demonstrated for scenario 1 (Table 6). Matrix decomposition is also useful to analyze high-dimensional datasets with many factors ($P >> N$ problem), and to fit one or multiple kernels of different types within a multivariate ridge regression model. For example, for modeling dominance, epistasis [37], Gaussian or Arc-cosine relationships [21, 38]. The computing costs for matrix decomposition to obtain those eigenvectors, however, may outweigh the benefits for THGS with an increasing number of individuals and markers in the analysis.

The trade-off for higher speed is a slightly lower accuracy of GEBVs of 0.01 compared to REML under realistic conditions when heritability was low and genetic correlations between environments were medium to high (Figure 2a). PEGS and THGS exploited genetic correlations between environments under these conditions and had higher accuracy of GEBVs than the univariate approach (Figure 2a and b). Only in the worst case, when all heritabilities and all genetic correlations between environments were low, the benefit of multivariate genomic prediction for achieving a higher accuracy than the univariate approach vanished with PEGS and THGS (Figure 2a and b). The reason is a lower precision in estimating genetic correlations as their standard errors were notably higher for PEGS and THGS than for REML (Figure 7). Moreover, PEGS and THGS slightly underestimated heritabilities and slightly overestimated genetic correlations. The bias of GEBVs, however, was close to zero and approached zero with increasing number of individuals per environment (Figure 3, Table 5, Appendix).

Residuals were treated as uncorrelated between environments for three reasons. First, the phenotypes come from different plants that are assumed to have uncorrelated environmental effects in their residuals. Second, epistatic effects, which are not captured by the marker effects in the model of equation (1), are assumed to have small covariances between environments. Third, the PEGS and THGS algorithms are faster because the absorption matrix $\mathbf{M}$, which is used in equations 3-7, is block-diagonal with one block per environment, $\mathbf{M}_k$. In addition, fewer computations are required to update estimated marker effects when the residual covariance matrix is diagonal (see equation 2). If phenotypes come from multiple quantitative traits, residual covariances may need to be modeled to avoid further bias in estimated genetic parameters and GEBVs, which may increase runtime [13] and offset the computational advantage compared to REML. However, these covariances could be modeled with an additional random term constructed by the cross-product of sparse 0/1-incidence matrices for genotypes from different environments. Otherwise, the effect of neglecting residual covariances on bias of genetic parameters and GEBVs could be evaluated on a case-by-case basis.

Estimated variances and covariances from the methods PE and TH are unbiased when the mixed-model equations are weighted by the true variances and covariances as shown in section 1 of the supplement and Appendix B. In practice, however, an iterative procedure starts with best guesses for genetic parameters, and thus estimates are not expected to be unbiased, which is identical to REML or iterative MIVQUE [39]. As discussed in [12], estimates may be further biased when populations are under selection. This was not considered here because plant-breeding

datasets used for genomic prediction may not contain selection if they are properly augmented by unselected genotypes or designed to maximize prediction accuracy [40, 41]. Furthermore, data may come from a single selection stage and thereby do not contain selection information. Yet, Ouweltjes et al. [42] and VanRaden and Jung [12] found that PE can be more suitable than TH for selected data, but both methods are more biased than REML. These studies were performed using pedigree information and the bias was attributed to neglecting off-diagonals of the relationship matrix. To better understand this, the original quadratic form, $\hat{\boldsymbol{\beta}}_k'\hat{\boldsymbol{\beta}}_k$, can be compared to $\tilde{\boldsymbol{\beta}}_k'\hat{\boldsymbol{\beta}}_k$ from equation (5). For ease of explanation, only the univariate case and the method PE with $\tilde{\boldsymbol{\beta}}_k = \mathbf{Z}_k'\mathbf{M}_k\boldsymbol{y}_k$ is considered here. Using BLUP formulas [43], the quadratic forms can be written as

$$\hat{\boldsymbol{\beta}}_k'\hat{\boldsymbol{\beta}}_k = (\boldsymbol{y}_k - \mathbf{X}\hat{\boldsymbol{b}}_{\mathrm{GLS}_k})'\mathbf{V}_k^{-1}\mathbf{Z}_k\sigma^2_{\beta_k}\sigma^2_{\beta_k}\mathbf{Z}_k'\mathbf{V}_k^{-1}(\boldsymbol{y}_k - \mathbf{X}\hat{\boldsymbol{b}}_{\mathrm{GLS}_k}), \tag{9}$$

and

$$\begin{aligned}\tilde{\boldsymbol{\beta}}_k'\hat{\boldsymbol{\beta}}_k &= \boldsymbol{y}_k'\mathbf{M}_k\mathbf{Z}_k\hat{\boldsymbol{\beta}}_k \\ &= (\boldsymbol{y}_k - \mathbf{X}\hat{\boldsymbol{b}}_{\mathrm{LS}_k})'\mathbf{Z}_k\sigma^2_{\beta_k}\mathbf{Z}_k'\mathbf{V}_k^{-1}(\boldsymbol{y}_k - \mathbf{X}\hat{\boldsymbol{b}}_{\mathrm{GLS}_k}),\end{aligned} \tag{10}$$

where $\mathbf{V}_k^{-1}$ is the inverse of the variance-covariance matrix of $\boldsymbol{y}_k$, $\mathbf{V}_k = \mathbf{Z}_k\mathbf{Z}_k'\sigma^2_{\beta_k} + \mathbf{I}\sigma^2_{e_k}$, and $\hat{\boldsymbol{b}}_{\mathrm{GLS}_k}$ and $\hat{\boldsymbol{b}}_{\mathrm{LS}_k}$ are the Generalized Least Squares and Least Squares estimators of $\boldsymbol{b}$, respectively. Thus in $\tilde{\boldsymbol{\beta}}_k$, the matrix $\mathbf{V}_k^{-1}$, which contains genomic relationships between individuals, i.e., $\mathbf{Z}_k\mathbf{Z}_k'$, is not used to weigh $\boldsymbol{y}_k$, neither for estimating fixed effects ($\hat{\boldsymbol{b}}_{\mathrm{LS}_k}$) nor random effects. However, THGS in combination with a Principal components or an eigenvector regression does not have this issue and can be used for balanced datasets with selection.

PEGS and THGS should be evaluated against alternative methods that are commonly used in plant breeding for modeling phenotypes from multiple environments. These are compound symmetry and extended factor analytic (XFA) models [44]. Compound symmetry models fit a term for the average genetic effect of an individual across environments and another term for the specific environmental effects for an individual. As each term is modeled with only one variance, this model assumes that the genetic correlations between all pairs of environments are identical. The difference between that single correlation and the true correlation between any one pair can be regarded as bias. The XFA model fits more parameters than the compound symmetry model to reduce this bias, but less parameters than an unstructured multivariate model that fits a correlation for each pair of environments. XFA models thereby balance bias and precision of estimated genetic correlations. Therefore, these two alternative models tend to bias estimated genetic correlations between environments and are expected to decrease accuracy of GEBVs compared to estimating genetic correlations between all pairs of environments.

The iterative algorithm of PEGS and THGS differs from that of REML and Bayesian Gibbs sampling. In each iteration of REML, the mixed-model equations are fully solved to obtain estimates of the model effects conditional on the current variance components of that iteration. The estimated model effects are then used to update the variance components and a new iteration begins unless the change in

variance components is small. In PEGS and THGS, in contrast, the model effects are merely updated, not solved, before variance components are updated and a new iteration begins. In Bayesian Gibbs sampling, similar computations are conducted in each iteration as in PEGS and THGS. However, rather than converging directly to a solution within a small number of iterations, the Gibbs algorithm samples from the posterior for thousands of iterations, and therefore must have longer runtimes.

5 Conclusion

PEGS and THGS are fast, memory-efficient, and reliable algorithms for genomic prediction for both balanced and unbalanced experimental designs. They are scalable with increasing number of response variables and markers. Their runtime is much lower than for REML and Gibbs sampling. For balanced designs, THGS provides unbiased GEBVs and estimated genetic correlations if only an intercept is modeled, and eigenvalue decomposition is feasible. Without eigenvalue decomposition, the accuracy of GEBVs of PEGS and THGS is slightly lower than REML, but higher than univariate THGS under realistic genetic correlations among environments. Estimated genetic parameters have little bias, but the standard errors are larger than for REML. More studies are needed to evaluate them for unbalanced datasets with selection.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Availability of data and materials
Genotypic data of the wheat dataset are available in the R package BGLR using the command `data(wheat,package='BGLR')`, and genotypic data of the SoyNAM dataset are available in the R package SoyNAM using the command `data <- SoyNAM::ENV()`. An implementation of PEGS is provided in the R package bWGR (2.0), function **mrr**.

Competing interests
The authors declare that they have no competing interests.

Funding
The authors are salaried researchers. No particular funding was provided for this research.

Author's contributions
AX and DH developed the methods, implemented the algorithms, planned the validations, wrote the manuscript.

Acknowledgements
Not applicable.

Author details
[1] Biostatistics, Corteva Agrisciences, 8305 NW 62nd Ave, 50131, Johnston, Iowa, USA. [2] Department of Agronomy,Purdue University, 915 W State St, 47907, West Lafayette, Indiana, USA.

References

1. Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E.: Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**(4), 1819–1829 (2001)
2. de los Campos, G., Hickey, J.M., Pong-Wong, R., Daetwyler, H.D., Calus, M.P.: Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics **193**(2), 327–345 (2013)
3. Hickey, J.M., Chiurugwi, T., Mackay, I., Powell, W., Eggen, A., Kilian, A., Jones, C., Canales, C., Grattapaglia, D., Bassi, F., *et al.*: Genomic prediction unifies animal and plant breeding programs to form platforms for biological discovery. Nature genetics **49**(9), 1297–1303 (2017)
4. Calus, M.P., Veerkamp, R.F.: Accuracy of multi-trait genomic selection using different methods. Genetics Selection Evolution **43**(1), 1–14 (2011)
5. Jia, Y., Jannink, J.-L.: Multiple-trait genomic selection methods increase genetic value prediction accuracy. Genetics **192**(4), 1513–1522 (2012)
6. Meyer, K.: Maximum likelihood estimation of variance components for a multivariate mixed model with equal design matrices. Biometrics, 153–165 (1985)

7. Thompson, E., Shaw, R.: Pedigree analysis for quantitative traits: variance components without matrix inversion. Biometrics, 399–413 (1990)
8. Leventhal, D., Lewis, A.S.: Randomized methods for linear constraints: convergence rates and conditioning. Mathematics of Operations Research **35**(3), 641–654 (2010)
9. Ma, A., Needell, D., Ramdas, A.: Convergence properties of the randomized extended gauss–seidel and kaczmarz methods. SIAM Journal on Matrix Analysis and Applications **36**(4), 1590–1604 (2015)
10. Cunningham, E., Henderson, C.R.: An iterative procedure for estimating fixed effects and variance components in mixed model situations. Biometrics, 13–25 (1968)
11. Thompson, R.: Iterative estimation of variance components for non-orthogonal data. Biometrics, 767–773 (1969)
12. VanRaden, P., Jung, Y.: A general purpose approximation to restricted maximum likelihood: the tilde-hat approach. Journal of Dairy Science **71**(1), 187–194 (1988)
13. Schaeffer, L.: Pseudo expectation approach to variance component estimation. Journal of Dairy Science **69**(11), 2884–2889 (1986)
14. Henderson, C.: A simple method for unbiased estimation of variance components in the mixed model. J. Anim. Sci **51**(Suppl 1), 119 (1980)
15. Legarra, A., Misztal, I.: Computing strategies in genome-wide selection. Journal of dairy science **91**(1), 360–366 (2008)
16. Hayes, J., Hill, W.: Modification of estimates of parameters in the construction of genetic selection indices ('bending'). Biometrics, 483–493 (1981)
17. Xavier, A., Muir, W., Rainey, K.: bwgr: Bayesian whole-genome regression. Bioinformatics **36**(6), 1957–1959 (2019). doi:10.1093/bioinformatics/btz794
18. Trevor Hastie, J.F. Robert Tibshirani: The Elements of Statistical Learning - Data Mining, Inference, and Prediction. Springer, New York (2001)
19. de Los Campos, G., Gianola, D., Rosa, G.J., Weigel, K.A., Crossa, J.: Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel hilbert spaces methods. Genetics Research **92**(4), 295–308 (2010)
20. Ødegård, J., Indahl, U., Strandén, I., Meuwissen, T.H.: Large-scale genomic prediction using singular value decomposition of the genotype matrix. Genetics Selection Evolution **50**(1), 1–12 (2018)
21. Xavier, A.: Technical nuances of machine learning: implementation and validation of supervised methods for genomic prediction in plant breeding. Crop Breeding and Applied Biotechnology **21** (2021)
22. Habier, D., Fernando, R., Dekkers, J.C.: The impact of genetic relationship information on genome-assisted breeding values. Genetics **177**(4), 2389–2397 (2007)
23. Johnson, D., Thompson, R.: Restricted maximum likelihood estimation of variance components for univariate animal models using sparse matrix techniques and average information. Journal of dairy science **78**(2), 449–456 (1995)
24. Pocrnic, I., Lourenco, D.A., Masuda, Y., Misztal, I.: Dimensionality of genomic information and performance of the algorithm for proven and young for different livestock species. Genetics Selection Evolution **48**(1), 1–9 (2016)
25. Gilmour, A., Gogel, B., Cullis, B., Thompson, R., Butler, D., Cherry, M., Collins, D., Dutkowski, G., Harding, S., Haskard, K., et al.: Asreml user guide release 4.1 structural specification. VSN Int Ltd (2015)
26. Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T., Lee, D., *et al.*: Blupf90 and related programs (bgf90). In: Proceedings of the 7th World Congress on Genetics Applied to Livestock Production, vol. 28 (2002). Montpellier
27. Masuda, Y., Baba, T., Suzuki, M.: Application of supernodal sparse factorization and inversion to the estimation of (co) variance components by residual maximum likelihood. Journal of animal breeding and genetics **131**(3), 227–236 (2014)
28. Crossa, J., Campos, G.d.l., Pérez, P., Gianola, D., Burgueno, J., Araus, J.L., Makumbi, D., Singh, R.P., Dreisigacker, S., Yan, J., *et al.*: Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics **186**(2), 713–724 (2010)
29. Gianola, D., Okut, H., Weigel, K.A., Rosa, G.J.: Predicting complex quantitative traits with bayesian neural networks: a case study with jersey cows and wheat. BMC genetics **12**(1), 1–14 (2011)
30. Gianola, D., Fernando, R.L., Schön, C.-C.: Inferring trait-specific similarity among individuals from molecular markers and phenotypes with bayesian regression. Theoretical Population Biology **132**, 47–59 (2020)
31. Gianola, D., Fernando, R.L.: A multiple-trait bayesian lasso for genome-enabled analysis and prediction of complex traits. Genetics **214**(2), 305–331 (2020)
32. Xavier, A., Muir, W.M., Rainey, K.M.: Assessing predictive properties of genome-wide selection in soybeans. G3: Genes, Genomes, Genetics **6**(8), 2611–2616 (2016)
33. Xavier, A.: Efficient estimation of marker effects in plant breeding. G3: Genes, Genomes, Genetics **9**(11), 3855–3866 (2019)
34. Marone, D., Panio, G., Ficco, D., Russo, M.A., De Vita, P., Papa, R., Rubiales, D., Cattivelli, L., Mastrangelo, A.M.: Characterization of wheat dart markers: genetic and functional features. Molecular genetics and genomics **287**(9), 741–753 (2012)
35. Xavier, A., Jarquin, D., Howard, R., Ramasubramanian, V., Specht, J.E., Graef, G.L., Beavis, W.D., Diers, B.W., Song, Q., Cregan, P.B., *et al.*: Genome-wide analysis of grain yield stability and environmental interactions in a multiparental soybean population. G3: Genes, Genomes, Genetics **8**(2), 519–529 (2018)
36. Diers, B.W., Specht, J., Rainey, K.M., Cregan, P., Song, Q., Ramasubramanian, V., Graef, G., Nelson, R., Schapaugh, W., Wang, D., *et al.*: Genetic architecture of soybean yield and agronomic traits. G3: Genes, Genomes, Genetics **8**(10), 3367–3375 (2018)
37. Xu, S.: Mapping quantitative trait loci by controlling polygenic background effects. Genetics **195**(4), 1209–1222 (2013)
38. Montesinos-López, A., Montesinos-López, O.A., Montesinos-López, J.C., Flores-Cortes, C.A., de la Rosa, R.,

Crossa, J.: A guide for kernel generalized regression methods for genomic-enabled prediction. Heredity **126**(4), 577–596 (2021)
39. Searle, S.R., Casella, G., McCulloch, C.E.: Prediction of random variables. In: Variance Components, pp. 367–377. John Wiley and Sons, Inc., New York (1992). doi:10.1002/9780470316856.ch7
40. Habier, D.: Improved Molecular Breeding Methods. Google Patents. WO2015100236A1 (1988). https://patents.google.com/patent/WO2015100236A1/en
41. Rincent, R., Charcosset, A., Moreau, L.: Predicting genomic selection efficiency to optimize calibration set and to assess prediction accuracy in highly structured populations. Theoretical and Applied Genetics **130**(11), 2231–2247 (2017)
42. Ouweltjes, W., Schaeffer, L., Kennedy, B.: Sensitivity of methods of variance component estimation to culling type of selection. Journal of Dairy Science **71**(3), 773–779 (1988)
43. Searle, S.R., Casella, G., McCulloch, C.E.: Mixed model prediction (blup). In: Variance Components, pp. 269–277. John Wiley and Sons, Inc., New York (1992). doi:10.1002/9780470316856.ch7
44. Meyer, K.: Factor-analytic models for genotype× environment type problems and structured covariance matrices. Genetics Selection Evolution **41**(1), 1–11 (2009)
45. Searl, S.R.: Linear Models, p. 65. John Wiley Sons, Inc., New York (1971)

Appendix A: Efficient calculation of $\mathbf{Z}_k'\mathbf{M}_k\mathbf{Z}_k$ and $\mathbf{M}_k\boldsymbol{y}_k$

Only the diagonal elements of $\mathbf{Z}_k'\mathbf{M}_k\mathbf{Z}_k$ are needed as matrix $\mathbf{D}_k$ is diagonal (equation 4). They can be computed one at a time for environment k and marker j as

$$\boldsymbol{z}_{jk}'\mathbf{M}_{jk}\boldsymbol{z}_k = \boldsymbol{z}_{jk}'\boldsymbol{z}_{jk} - \boldsymbol{z}_{jk}'\mathbf{X}_k(\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k'\boldsymbol{z}_{jk}$$

where $(\mathbf{X}_k'\mathbf{X}_k)^{-1}$ is computed once before iterations start. Likewise, $\mathbf{M}_k\boldsymbol{y}_k$ of equation 7 can be obtained once as

$$\mathbf{M}_k\boldsymbol{y}_k = \boldsymbol{y}_k - \mathbf{X}_k(\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k'\boldsymbol{y}_k = \boldsymbol{y}_k - \mathbf{X}_k\hat{\boldsymbol{b}}_{LSk},$$

where $\hat{\boldsymbol{b}}_{LSk}$ denotes the Least Squares estimate of $\boldsymbol{b}$.

Appendix B: Expected value of $\tilde{\boldsymbol{\beta}}_k'\hat{\boldsymbol{\beta}}_k$

Let $\tilde{\boldsymbol{\beta}}_k = \mathbf{D}_k^{-1}\mathbf{Z}_k'\mathbf{M}_k\boldsymbol{y}_k$ and $\mathbf{M}_k = \mathbf{I}_k - \mathbf{X}_k(\mathbf{X}_k'\mathbf{X}_k)^{-1}\mathbf{X}_k$, as defined in section 2.3, and let $\hat{\boldsymbol{\beta}}_k = \sigma^2_{\beta_k}\mathbf{Z}_k'\mathbf{P}_k\boldsymbol{y}_k$ be the Best Linear Unbiased Predictor (BLUP) of $\boldsymbol{\beta}$ [43], where $\mathbf{P}_k = \mathbf{V}_k^{-1}[\mathbf{I}_k - \mathbf{X}_k(\mathbf{X}_k'\mathbf{V}_k^{-1}\mathbf{X}_k)^{-1}\mathbf{X}_k\mathbf{V}_k^{-1}]$ and $E(\hat{\boldsymbol{\beta}}) = \mathbf{0}$. Then, the expected value of the bilinear form $\tilde{\boldsymbol{\beta}}_k'\hat{\boldsymbol{\beta}}_k$ [45] is

$$\begin{aligned} E(\tilde{\boldsymbol{\beta}}_k'\hat{\boldsymbol{\beta}}_k) &= tr(Cov(\tilde{\boldsymbol{\beta}}_k, \hat{\boldsymbol{\beta}}_k')) + E(\tilde{\boldsymbol{\beta}}_k)'E(\hat{\boldsymbol{\beta}}_k) \\ &= tr(\mathbf{D}_k^{-1}\mathbf{Z}_k'\mathbf{M}_k\mathbf{V}_k\mathbf{P}_k\mathbf{Z}_k\sigma^2_{\beta_k}) \\ &= tr(\mathbf{D}_k^{-1}\mathbf{Z}_k'\mathbf{M}_k\mathbf{Z}_k)\sigma^2_{\beta_k}, \end{aligned}$$

because $\mathbf{M}_k\mathbf{V}_k\mathbf{P}_k = \mathbf{M}_k$. Hence,

$$\hat{\sigma}^2_{\beta_k} = \frac{\tilde{\boldsymbol{\beta}}_k'\hat{\boldsymbol{\beta}}_k}{tr(\mathbf{D}_k^{-1}\mathbf{Z}_k'\mathbf{M}_k\mathbf{Z}_k)},$$

and $E(\hat{\sigma}^2_{\beta_k}) = \sigma^2_{\beta_k}$. The extension to using $\hat{\boldsymbol{\beta}}_k$ from a multivariate BLUP is presented in section 1 of the supplement.

Appendix C: Equivalence of $\hat{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\beta}}$ using EVD

Let the eigenvalue decomposition of $\mathbf{Z}_k'\mathbf{Z}_k$ be $\mathbf{U}_k\boldsymbol{\Lambda}_k\mathbf{U}_k'$, where $\mathbf{U}_k$ is an orthonormal matrix of eigenvectors with the property $\mathbf{U}_k'\mathbf{U}_k = \mathbf{U}_k\mathbf{U}_k' = \mathbf{I}_m$, and $\boldsymbol{\Lambda}_k$ is a diagonal matrix of eigenvalues. The Principal component regression [18] can be written as

$$\begin{aligned} \mathbf{y}_k &= \mathbf{1}\mu_k + \mathbf{Z}_k\mathbf{U}_k\mathbf{U}_k'\boldsymbol{\beta}_k + \boldsymbol{e}_k \\ &= \mathbf{1}\mu_k + \check{\mathbf{Z}}_k\check{\boldsymbol{\beta}}_k + \boldsymbol{e}_k, \end{aligned}$$

where $\tilde{\mathbf{Z}}_k = \mathbf{Z}_k\mathbf{U}_k$ and $\check{\boldsymbol{\beta}}_k = \mathbf{U}_k'\boldsymbol{\beta}_k$. Let the estimate of $\check{\boldsymbol{\beta}}_k$ be $\tilde{\boldsymbol{\beta}}_k = \mathbf{D}_k^{-1}\check{\mathbf{Z}}_k'\boldsymbol{y}_k$ similar to equation (8), where $\mathbf{M}_k$ was omitted because $\mathbf{Z}_k$ and $\boldsymbol{y}_k$ are assumed

centered. Then, defining $\lambda_k = \sigma^2_{e_k}/\sigma^2_{\beta_k}$, and using $(\mathbf{U}_k)^{-1} = \mathbf{U}'_k$ and $(\mathbf{U}'_k)^{-1} = \mathbf{U}_k$,

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_k &= (\mathbf{Z}'_k\mathbf{Z}_k + \mathbf{I}_m\lambda_k)^{-1}\mathbf{Z}'_k\boldsymbol{y}_k \\
&= \mathbf{U}_k\tilde{\boldsymbol{\beta}}_k \\
&= \mathbf{U}_k\mathbf{D}_k^{-1}\check{\mathbf{Z}}'_k\boldsymbol{y}_k \\
&= \mathbf{U}_k(\boldsymbol{\Lambda}_k + \mathbf{I}_m\lambda_k)^{-1}\check{\mathbf{Z}}_k\boldsymbol{y}_k \\
&= \mathbf{U}_k[\mathbf{U}'_k\mathbf{U}_k(\boldsymbol{\Lambda}_k + \mathbf{I}_m\lambda_k)\mathbf{U}'_k\mathbf{U}_k]^{-1}\check{\mathbf{Z}}'_k\boldsymbol{y}_k \\
&= \mathbf{U}_k[\mathbf{U}'_k(\mathbf{U}_k\boldsymbol{\Lambda}_k\mathbf{U}'_k + \mathbf{I}_m\lambda_k)\mathbf{U}_k]^{-1}\check{\mathbf{Z}}'_k\boldsymbol{y}_k \\
&= \mathbf{U}_k\mathbf{U}'_k(\mathbf{Z}'_k\mathbf{Z}_k + \mathbf{I}_m\lambda_k)^{-1}\mathbf{U}_k\tilde{\mathbf{Z}}'_k\boldsymbol{y}_k \\
&= (\mathbf{Z}'_k\mathbf{Z}_k + \mathbf{I}_m\lambda_k)^{-1}\mathbf{U}_k\mathbf{U}'_k\mathbf{Z}'_k\boldsymbol{y}_k \\
&= (\mathbf{Z}'_k\mathbf{Z}_k + \mathbf{I}_m\lambda_k)^{-1}\mathbf{Z}'_k\boldsymbol{y}_k.
\end{aligned}$$

Appendix D: Polygenic model using EVD

The model can be written as

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{b} + \boldsymbol{g} + \boldsymbol{e}, \tag{11}$$

where $\boldsymbol{y}$, $\mathbf{X}$, $\boldsymbol{b}$, and $\boldsymbol{e}$ are defined as in statistical model 2.1, and $\boldsymbol{g}$ is a vector of breeding values that can be partitioned into $\boldsymbol{g}' = [\boldsymbol{g}'_1\ \boldsymbol{g}'_2\ \ldots\ \boldsymbol{g}'_K]$. It is assumed multivariate normal-distributed with mean zero and variance $\boldsymbol{\Sigma}_g \otimes \mathbf{G}$, where $\boldsymbol{\Sigma}_g$ is a K x K variance-covariance matrix of breeding values for K environments and $\mathbf{G}$ is the genomic relationship matrix. The eigenvalue decomposition of this matrix can be written as $\mathbf{G} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}'$, where $\mathbf{U}$ contains orthogonal eigenvectors and $\boldsymbol{\Lambda}$ is a diagonal matrix that contains eigenvalues. To diagonalize $\mathbf{G}$, model 11 was tranformed by $\mathbf{T} = \mathbf{1}_K \otimes \mathbf{U}'$, where $\mathbf{1}_K$ is a K vector of ones, hence

$$\begin{aligned}
\mathbf{T}\boldsymbol{y} &= \mathbf{T}\mathbf{X}\boldsymbol{b} + \mathbf{T}\boldsymbol{g} + \mathbf{T}\boldsymbol{e} \\
&= \tilde{\mathbf{X}}\boldsymbol{b} + \tilde{\boldsymbol{g}} + \tilde{\boldsymbol{e}},
\end{aligned}$$

where $\tilde{\boldsymbol{g}} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_g \otimes \boldsymbol{\Lambda})$ and $\tilde{\boldsymbol{e}} \sim N(\mathbf{0}, \oplus_{i=1}^{K}\mathbf{I}\sigma^2_{e_k})$.

Appendix E: Full-conditional Gauss-Seidel solution

Equation (2) can be rearranged to reduce the multivariate Gauss-Seidel solver into a univariate algorithm, as an extension of the algorithm in [15]. This circumvents the inverse in equation (2), but may have slower convergence. The estimated effect of marker j and environment k is updated as

$$\hat{\beta}^{(t+1)}_{jk}|\hat{\boldsymbol{\beta}}^{(t)}_j, \hat{\boldsymbol{\Sigma}}_\beta = \frac{\boldsymbol{z}'_{jk}\hat{\boldsymbol{e}}_k + \boldsymbol{z}'_{jk}\boldsymbol{z}_{jk}\hat{\beta}^{(t)}_{jk} - \hat{\sigma}^2_{e_k}\sum_{l=1,l\neq k}^{K}\hat{\boldsymbol{\Sigma}}^{-1}_{\beta_{kl}}\cdot\hat{\beta}^{(t)}_{jl}}{\boldsymbol{z}'_{jk}\boldsymbol{z}_{jk} + \hat{\sigma}^2_{e_k}\hat{\sigma}^{kk}_\beta}$$

where $\hat{\sigma}^{kk}_\beta$ is the kk element of $\hat{\boldsymbol{\Sigma}}^{-1}_\beta$. The update of $\hat{\beta}^{(t+1)}_{jk}$ is followed by the update of residuals of environment k as

$$\hat{\boldsymbol{e}}^{(new)}_k = \hat{\boldsymbol{e}}^{(old)}_k - \boldsymbol{z}_{jk}(\hat{\beta}^{(t+1)}_{jk} - \hat{\beta}^{(t)}_{jk}).$$