Efficient Computation of Genomic Prediction with Complex Genotype-by-Environment Interactions

A Preprint

Alencar Xavier
Principal Investigator, Corteva Agrisciences
Adjunct Associate Professor, Purdue University
8305 NW 62nd Ave, Johnston IA 50131
alencar.xavier@corteva.com

Daniel Runcie
Associate Professor, UC Davis
387 N Quad, Davis, CA 95616
deruncie@ucdavis.edu

David Habier
Senior Research Scientist, Corteva Agrisciences
8305 NW 62nd Ave, Johnston IA 50131
david.habier@corteva.com

May 15, 2024

Abstract

Understanding genotypic performance in different environments allows breeders to tailor varieties to specific conditions or to select for stability across environments. Genomic prediction models that capture genotype-by-environment interaction are instrumental in predicting site-specific performance by leveraging information among related individuals and correlated environments. However, the implementation of such models requires specialized algorithms or it is otherwise computationally prohibitive. This study reviews and describes the algorithms of efficient approaches to capture complex GxE patterns at scale. We rate the scalability of the different methods for their applications. Benchmarks of accuracy and computation time are performed on multiple simulated scenarios for dense and sparse testing. A set of recommendations are provided for implementation.

Keywords Accuracy · Genomic Prediction · Multivariate Models · Matrix Decomposition · Kernel Rotation

1 Introduction

Multi-environment trials constitute the main source of data for plant breeding decision-making. Genotype-by-environment interactions (GxE) occur when the performance of a genotype varies across trials due to different environmental conditions. Predicting genotype performance within environments is essential for enhanced adaptation and for the selection of superior genotypes. Understanding how genotypes perform in diverse environments allows breeders to tailor varieties to specific conditions or select genotypes that perform consistently well across different environments. This leads to more effective variety development and ensures new cultivars thrive in the intended conditions (Elias et al., 2016).

Genomic prediction (Meuwissen et al., 2001) plays a crucial role in predicting genotype performance within environments. It leverages relationship information to predict breeding values (Habier et al., 2007) and, by doing so, accelerates the breeding process by shortening breeding cycles while reducing phenotyping costs (Crossa et al., 2021). Genomic prediction models that capture complex GxE interactions can be more accurate by leveraging information among correlated environments (Xavier and Habier, 2022).

The limiting factor to implementing genomic models with terms that account for genotype-by-environment interaction is often computational (Heslot et al., 2014). Interaction models are inherently more complex than standard genomic models due to the number of parameters that need to be estimated (Hardner, 2017) with complex covariance structures

(Martini et al. [2020]). Compound symmetry models (*e.g.*, $G + G \times E$) assume constant interaction correlation among all pairs of environments (Cuevas et al. [2016]), interaction kernels (*e.g.*, $Var[\mathbf{g}_{G \times E}] = \mathbf{K_g} \# \mathbf{K_e} \sigma^2_{G \times E}$) require the Hadamard product of genetics and environment structures (Jarquín et al., 2014), whereas unstructured models (*e.g.*, $Var[\mathbf{g}] = \mathbf{\Sigma_g} \otimes \mathbf{G}$) require the computation of complex Kronecker products (Crossa et al., 2022, Bustos-Korts et al., 2016). For computational reasons, the analysis of MET often falls back to overly simplified parameterizations, such as Finlay–Wilkinson and GGE models (Malosetti et al., 2013). Fit complex models with multiple traits and environments requires new and more efficient algorithms. This is particularly important with the advent of high-throughput phenotyping data, where the number of trait-environment combinations increases substantially. Due to the number of traits being fit at once, this novel class of methods must be suitable to fit models that can be described "megavariate" mixed models.

The purpose of this study is to review computationally efficient methods that can be utilized for genomic prediction within environments while capturing complex GxE patterns. We rate the scalability of the different methods for their applications on varying numbers of genotypes, markers, and environments. Benchmarks of accuracy and computation time are performed on multiple simulated scenarios.

2 Environment-specific predictions

This section briefly introduces approaches that are computationally feasible at scale. Such methods enable genomic predictions at the single-environment level, which are necessary to capture complex genotype-by-environment interactions. For the given k^{th} environment, the phenotypes is modeled as:

$$\mathbf{y}_k = \mu_k + \mathbf{g}_k + \mathbf{e}_k \tag{1}$$

with variance

$$\begin{aligned} Var(\mathbf{y}_k) &= \mathbf{V}_k = \mathbf{G}_k + \mathbf{R}_k, \\ Var(\mathbf{g}_k) &= \mathbf{G}_k, \\ Var(\mathbf{e}_k) &= \mathbf{R}_k. \end{aligned} \tag{2}$$

Under an additive parameterization, the genetic variance is described by $\mathbf{G}_k = \mathbf{Z}_k \mathbf{Z}'_k \sigma^2_{\beta(k)}$. For simplicity, assume residuals to be independent and identically distributed as $\mathbf{R}_k = \mathbf{I}\sigma^2_{e(k)}$. The genetic term can be described as a linear combination of marker effects (Habier et al., 2007), as

$$\mathbf{g}_k = \mathbf{Z}_k \beta_k \tag{3}$$

where $Var(\beta_k) = \mathbf{I}\sigma^2_{\beta(k)}$. In the analysis of multiple environments, the covariance between pairs of environments is $Cov(\beta_k, \beta_{k'})$ defines the genotype-by-environment interactions, as

$$\sigma_{\beta(\mathbf{k},\mathbf{k'})} = \rho_{\beta(k,k')} \sigma_{\beta(k)} \sigma_{\beta(k')}, \tag{4}$$

where $\rho_{\beta(k,k')}$ is the genotype-by-environment correlation between environments k and k'. Under realistic conditions, every pair of environments has a unique correlation. Sophisticated approximations, such as factor analysis (Gilmour, 2019), have been proposed to handle unstructured covariances but are not computationally feasible as the data size and the number of parameters increase. Modeling approaches meant to accommodate larger and more complex datasets are presented next (Xavier and Habier, 2022, Runcie et al., 2021, Pocrnic et al., 2016).

2.1 Univariate by environment

The univariate model is the simplest approach to obtain environment-specific predictions. Environment-specific breeding values are attained by treating each environment as an independent trait, fitting location-year combinations from raw or spatially adjusted phenotypes (Möhring and Piepho, 2009, Piepho et al., 2012). Such models can provide accurate predictions for closely related populations, but the quality of such training set is often constrained by the population size (Xu, 2003) and genetic scope (Habier et al., 2013).

An efficient algorithm to fit this model is called the Pseudo-Expectation Gauss-Seidel (PEGS) method (Xavier and Habier, 2022). PEGS is an iterative model based on the pseudo-expectation (PE) estimator of variance components (Schaeffer, 1986) along with the estimation of coefficients using Gauss-Seidel residual update (GS, Legarra and Misztal, 2008) updated in random order within each iteration (Ma et al., 2015). PE is an efficient approximation of REML (VanRaden and Jung, 1988) that provides unbiased and invariant variance components (Xavier and Habier, 2022). In terms of implementation, PEGS resembles a non-MCMC version of the Bayesian Gibbs sampler. It estimates breeding values from an SNP-BLUP regression model as

$$\begin{aligned}\mathbf{y} &= \mu + \mathbf{g} + \mathbf{e} \\ &= \mu + \mathbf{Z}\beta + \mathbf{e},\end{aligned} \tag{5}$$

where $\beta \sim N(0, \mathbf{I}\sigma^2_{\beta})$ and $\mathbf{e} \sim N(0, \mathbf{I}\sigma^2_e)$. The variance components are estimated as

$$\sigma^2_{\beta} = \frac{\tilde{\beta}'\hat{\beta}}{\mathbf{Tr}(\tilde{\mathbf{Z}}\tilde{\mathbf{Z}})} \tag{6}$$

$$\sigma^2_e = \frac{\mathbf{y}'\mathbf{e}}{n-1} \tag{7}$$

where $\tilde{\mathbf{Z}}$ corresponds to the design matrix of marker effects with centralized columns, and $\tilde{\beta} = \tilde{\mathbf{Z}}\mathbf{y}$ or, alternatively, $\tilde{\beta} = \mathbf{Z}(\mathbf{y} - \bar{\mathbf{y}})$. When the intercept is the only fixed effect, it is also possible to simply the trace operation of the denominator as $\mathbf{Tr}(\tilde{\mathbf{Z}}\tilde{\mathbf{Z}}) = \sum_{j=1}^{J} \mathbf{Var}(\mathbf{z_j})$. The marker effects are solved with

$$\hat{\beta}_j^{(t+1)} = \frac{\mathbf{z}_j'\mathbf{e} + \hat{\beta}_j^t \mathbf{z}_j'\mathbf{z}_j}{\mathbf{z}_j'\mathbf{z}_j + \sigma^2_e \sigma^{-2}_{\beta}}, \tag{8}$$

and update of each vector of marker effects is followed by the update of the residuals as

$$\mathbf{e} = \mathbf{e} - \mathbf{z}_j'(\hat{\beta}_j^{(t+1)} - \hat{\beta}_j^{(t)}). \tag{9}$$

Similarly, the intercept updates, followed by the residual update, are achieved with

$$\mu^{(t+1)} = \mu^{(t)} + n^{-1}\sum_{i=1}^{I} \mathbf{e}_i \tag{10}$$

$$\mathbf{e} = \mathbf{e} - (\mu^{(t+1)} - \mu^{(t)}). \tag{11}$$

This solver is suitable for estimating coefficients and variance components for univariate models as well as multivariate parameterizations that do not explicitly involve the computation of covariances, including canonical transformation and structural equation models.

2.2 Canonical transformation

For complete data across environments, Canonical Transformation (CT) presents itself as an efficient framework to fit multiple traits when all traits have the same statistical model (Meyer, 1985, Konstantinov and Erasmus, 1993). Under CT, the matrix of phenotypes is converted into a series of orthogonal canonical traits through single-value decomposition, as

$$\begin{aligned}\mathbf{Y} &= \mathbf{UDV}' \\ &= \mathbf{QV}'\end{aligned} \tag{12}$$

where $\mathbf{Q} = \mathbf{UD}$, and $\mathbf{Y}$ is a matrix of phenotypes with rows and columns representing observations and traits, respectively. Each canonical trait is fit using an univariate model

$$\mathbf{q}_k = \mu_k + \mathbf{Z}_k\gamma_k + \mathbf{e}_k, \tag{13}$$

and the breeding values are recovered with

$$\begin{aligned}\mathbf{G} &= \mathbf{Z\Gamma V} \\ &= \mathbf{ZB}.\end{aligned} \tag{14}$$

Note $\mathbf{B} = \mathbf{\Gamma V}$ because $\mathbf{V}$ is rotating marker effects back to the natural scale of the phenotypes.

2.3 MegaLMM

MegaLMM (Runcie et al., 2021) extends the CT model with a more parsimonious latent representation of the phenotypes with the addition of trait-specific model terms. This enables efficient handling of missing values by MegaLMM, while permitting the model to accommodate more traits at a time. Before fitting the genomic model, latent spaces are inferred from a stochastic matrix decomposition of $\mathbf{Y}$ based on the following statistical model:

$$\mathbf{Y} = \mathbf{F}\mathbf{\Lambda} + \mathbf{J} \tag{15}$$

The matrix of phenotypes is decomposed into latent spaces $\mathbf{F}$ and rotation $\mathbf{\Lambda}$, with residuals $\mathbf{J}$ capturing what is not explained by $\mathbf{F}\mathbf{\Lambda}$. The $\mathbf{J}$ residual matrix contain genetic signal ($\mathbf{Z}\mathbf{\Delta}$) not captured by $\mathbf{F}\mathbf{\Lambda}$ and true error ($\mathbf{E}$), thus $\mathbf{J} = \mathbf{Z}\mathbf{\Delta} + \mathbf{E}$. Under this approach, each trait is fit under univariate settings as

$$\mathbf{y_k} = \mu_\mathbf{k} + \mathbf{F}\lambda_\mathbf{k} + \mathbf{Z_k}\delta_\mathbf{k} + \mathbf{e_k}, \tag{16}$$

and, subsequently, each latent space is modeled by

$$\mathbf{f_l} = \mu_\mathbf{l} + \mathbf{Z}\gamma_\mathbf{l} + \mathbf{e_l}. \tag{17}$$

The shared genetic signal is captured under $\mathbf{F}\lambda_\mathbf{k}$, whereas environment-specific is captured with $\mathbf{Z_k}\delta_\mathbf{k}$. Fitted value capturing all shared information, genetic or otherwise, can be obtained with $\hat{\mathbf{Y}} = \mathbf{1M} + \mathbf{F}\mathbf{\Lambda} + \mathbf{Z}\mathbf{\Delta}$. Multivariate marker effects are recovered with $\beta_\mathbf{k} = \mathbf{\Gamma}\mathbf{\Lambda}\lambda_\mathbf{k} + \delta_\mathbf{k}$, so that the complete matrix of breeding values can be recovered with

$$\begin{aligned} \mathbf{G} &= \mathbf{Z}\mathbf{\Gamma}\mathbf{\Lambda} + \mathbf{Z}\mathbf{\Delta} \\ &= \mathbf{Z}(\mathbf{\Gamma}\mathbf{\Lambda} + \mathbf{\Delta}) \\ &= \mathbf{Z}\mathbf{B}. \end{aligned} \tag{18}$$

The number of latent spaces is inferred while fitting the model, with strong variable selection. The original implementation of MegaLMM uses Markov Chain Monte Carlo, estimating $\hat{\mathbf{F}}$ and $\hat{\mathbf{\Lambda}}$ by alternating coefficient estimation between $(\mathbf{Y}|\mathbf{F}) = \mathbf{F}\mathbf{\Lambda} + \mathbf{J}$ and $(\mathbf{Y}'|\mathbf{\Lambda}) = \mathbf{\Lambda}\mathbf{F} + \mathbf{J}'$.

2.4 SEM

Structural equation modeling (SEM) can be utilized as a framework for multiple correlated response variables (Gianola and Sorensen, 2004) and complex trait networks (Valente et al., 2013). SEM differs from the MegaLMM model by explicitly parameterizing phenotypic traits in the model as opposed to using latent variables.

Under a fully-connected SEMs, any given phenotypic trait is fit as a function of other phenotypes. Thus,

$$\mathbf{Y} = 1\mu + \mathbf{\Psi}\mathbf{Y} + \mathbf{Z}\mathbf{\Delta} + \mathbf{E}, \tag{19}$$

where μ is the vector of intercepts for the traits, $\mu = \{\mu_\mathbf{1}\, \mu_\mathbf{2}, ..., \mu_\mathbf{k}\}$. The matrix $\mathbf{\Psi}$ has dimension $k \times k$ with elements quantifying the linear associations among traits, and with zeros on the diagonals. The final set of marker effects is estimated from

$$\mathbf{B} = (\mathbf{I} - \mathbf{\Psi})^{-\mathbf{1}}\mathbf{\Delta}. \tag{20}$$

SEM solution is straightforward when the modeling involves a small number of traits, under balanced settings.

2.5 MegaSEM

An intermediate parameterization between SEM and MegaLMM, herein referred to as MegaSEM, can be achieved through the latent representation of the genetic term. These latent spaces are notated as $\mathbf{F_0}$, these are derived from the single-value decomposition of genomic values estimated from univariate models, $\mathbf{G_0} = \mathbf{Z}\mathbf{B_0}$, using univariate marker effects $\mathbf{B_0} = \{\mathbf{b_{0(1)}}, \mathbf{b_{0(2)}}, \cdots, \mathbf{b_{0(K)}}\}$. The fitted genetic term is subsequently decomposed as

$$\begin{aligned} \mathbf{G_0} &= \mathbf{U_0}\mathbf{D_0}\mathbf{V_0'} \\ &= \mathbf{F_0}\mathbf{V_0'}. \end{aligned} \tag{21}$$

Under this proposed framework, latent spaces correspond to principal components fitted as either fixed or random effects, depending on the number of latent spaces utilized. One $\mathbf{F_0}$ has been computed, MegaSEM fits

$$\mathbf{y_k} = \mu_\mathbf{k} + \mathbf{F_0}\alpha_\mathbf{k} + \mathbf{Z_k}\delta_\mathbf{k} + \mathbf{e_k}, \tag{22}$$

which has the same form as equation 16. Like the MegaLMM model, the shared genetic signal is captured under $\mathbf{F_0}\alpha_\mathbf{k}$, whereas environment-specific is captured with $\mathbf{Z_k}\delta_\mathbf{k}$. The complete breeding value matrix is retrieved from

$$\begin{aligned}\mathbf{G} &= \mathbf{F_0 A} + \mathbf{Z\Delta}\\ &= \mathbf{ZB_0V_0A} + \mathbf{Z\Delta}\\ &= \mathbf{Z}(\mathbf{B_0V_0A} + \mathbf{\Delta})\\ &= \mathbf{ZB}.\end{aligned} \tag{23}$$

Note that $\mathbf{F_0} = \mathbf{ZB_0V_0}$ because, $\mathbf{G_0} = \mathbf{F_0V_0'}$ as $\mathbf{G_0} = \mathbf{ZB_0V_0}$, $\mathbf{G_0V_0} = \mathbf{F_0V_0'V_0}$, and $\mathbf{V_0'V_0} = \mathbf{I}$. As shown in equation 23, the multivariate-equivalent marker effect estimator ($\beta_\mathbf{k}$) for any given trait is a function of a linear combination of marker effects from all environments, in addition to an environment-specific estimator, $\beta_\mathbf{k} = \mathbf{B_0V_0'}\alpha_\mathbf{k} + \delta_\mathbf{k}$. A special case happens when all principal components of $\mathbf{G_0}$ are utilized (no dimensionality reduction), the model can further be simplified to

$$\mathbf{y_k} = \mu_\mathbf{k} + \mathbf{F_0}\alpha_\mathbf{k} + \mathbf{e_k} \tag{24}$$

where the environment-specific term is not necessary, provided the model is parameterized with all linear combinations of environments, yielding $\mathbf{B} = \mathbf{B_0V_0A}$. In both cases, the computational cost of MegaSEM consists of running an univariate model twice, first to estimate $\mathbf{B_0}$ and, subsequently, to estimate $\mathbf{B}$, in addition to the SVD of $\mathbf{G_0}$.

2.6 Multivariate PEGS

Xavier and Habier, 2022 introduced the multivariate version of PEGS. Under the multivariate framework, the PE genetic covariance component between any pair of environments k and k', and the residual variance for the k environment, have the following solution:

$$\mathbf{\Sigma}_{\beta(k,k')} = \frac{\tilde{\beta}'_\mathbf{k}\hat{\beta}_\mathbf{k'} + \tilde{\beta}'_\mathbf{k}\hat{\beta}_\mathbf{k'}}{\mathbf{Tr}(\tilde{\mathbf{Z}}'_\mathbf{k}\tilde{\mathbf{Z}}_\mathbf{k}) + \mathbf{Tr}(\tilde{\mathbf{Z}}'_\mathbf{k'}\tilde{\mathbf{Z}}_\mathbf{k'})} \tag{25}$$

$$\mathbf{\Sigma}_{e(k)} = \frac{\mathbf{y'_k e_k}}{n_k - 1} \tag{26}$$

where $\tilde{\mathbf{Z}}$ corresponds to the design matrix of marker effects with centralized columns, and $\tilde{\beta}_\mathbf{k} = \tilde{\mathbf{Z}}_\mathbf{k}\mathbf{y_k}$. The multivariate solution for updating coefficients, one marker at a time (j), is

$$\hat{\beta}_j^{(t+1)} = (\hat{\mathbf{\Sigma}}_e^{-1}\dot{\mathbf{Z}}_j'\dot{\mathbf{Z}}_j + \hat{\mathbf{\Sigma}}_\beta^{-1})^{-1}\hat{\mathbf{\Sigma}}_e^{-1}\dot{\mathbf{Z}}_j'(\dot{\mathbf{Z}}_j\hat{\beta}_j^{(t)} + \hat{\mathbf{e}}), \tag{27}$$

where $\dot{\mathbf{Z}}_j = \oplus_{k=1}^{K}\mathbf{z}_{jk}$. The update of each vector of marker effects is followed by the update of the residuals, as

$$\mathbf{e}_k = \mathbf{e}_k - \mathbf{Z}'_{j(k)}(\hat{\beta}_{j(k)}^{(t+1)} - \hat{\beta}_{j(k)}^{(t)}). \tag{28}$$

Depending on the number of environments, the inversion of $\mathbf{\Sigma}_\beta$ may require bending (Hayes and Hill, 1981, Meyer, 2019). Bending forces $\mathbf{\Sigma}_\beta$ to be positive-definite by adding a constant to its diagonal when this displays negative eigenvalues.

The non-vector operations involved in the multivariate PEGS are 1) the inversion of $\mathbf{\Sigma}_\beta$ and 2) solving the marker effects equation (27). The former issue can be mitigated by multiplying both sides of the equation by $\hat{\mathbf{\Sigma}}_\beta$, as proposed by Strandén and Garrick, 2009. Thus,

$$\begin{aligned}(\hat{\mathbf{\Sigma}}_e^{-1}\dot{\mathbf{Z}}_j'\dot{\mathbf{Z}}_j + \hat{\mathbf{\Sigma}}_\beta^{-1})\hat{\beta}_j^{(t+1)} &= \hat{\mathbf{\Sigma}}_e^{-1}\dot{\mathbf{Z}}_j'(\dot{\mathbf{Z}}_j\hat{\beta}_j^{(t)} + \hat{\mathbf{e}})\\ \hat{\mathbf{\Sigma}}_\beta(\hat{\mathbf{\Sigma}}_e^{-1}\dot{\mathbf{Z}}_j'\dot{\mathbf{Z}}_j + \hat{\mathbf{\Sigma}}_\beta^{-1})\hat{\beta}_j^{(t+1)} &= \hat{\mathbf{\Sigma}}_\beta\hat{\mathbf{\Sigma}}_e^{-1}\dot{\mathbf{Z}}_j'(\dot{\mathbf{Z}}_j\hat{\beta}_j^{(t)} + \hat{\mathbf{e}})\\ (\hat{\mathbf{\Sigma}}_\beta\hat{\mathbf{\Sigma}}_e^{-1}\dot{\mathbf{Z}}_j'\dot{\mathbf{Z}}_j + \mathbf{I})\hat{\beta}_j^{(t+1)} &= \hat{\mathbf{\Sigma}}_\beta\hat{\mathbf{\Sigma}}_e^{-1}\dot{\mathbf{Z}}_j'(\dot{\mathbf{Z}}_j\hat{\beta}_j^{(t)} + \hat{\mathbf{e}}).\end{aligned} \tag{29}$$

For the latter issue, one may address the inversion of the left-hand side by running an inner Gauss-Seidel, for the single site update of coefficients. Since the algorithm convergence is computed across coefficients for all marker-environment combinations, a single iteration of inner Gauss-Seidel per marker update suffices. Let

$$\begin{aligned}\mathbf{C} &= (\hat{\mathbf{\Sigma}}_\beta\hat{\mathbf{\Sigma}}_e^{-1}\dot{\mathbf{Z}}_j'\dot{\mathbf{Z}}_j + \mathbf{I})\\ \beta &= \hat{\beta}_j^{(t+1)}\\ \mathbf{r} &= \hat{\mathbf{\Sigma}}_\beta\dot{\mathbf{Z}}_j'\hat{\mathbf{\Sigma}}_e^{-1}(\dot{\mathbf{Z}}_j\hat{\beta}_j^{(t)} + \hat{\mathbf{e}}).\end{aligned} \tag{30}$$

aiming to solve $\mathbf{C}\upbeta = \mathbf{r}$. The single site update of the marker effect j, at trait k consists of

$$\upbeta_{\mathbf{k}} = \frac{\mathbf{r_k} - \mathbf{C_{k,-k}}\upbeta_{\mathbf{-k}}}{\mathbf{C_{k,k}}}. \tag{31}$$

2.7 Factor analytics

Multivariate PEGS provides simple solutions for unstructured covariance components (25). The covariance can be easily extended to simpler covariance structures, such as extended factor analytics (XFA). The XFA structure can be obtained by decomposing $\mathbf{\Sigma}_\upbeta$ using EVD, recomposing with a subset of eigenpairs, and reinstating the original diagonal elements to avoid changes in heritability. Let

$$\mathbf{\Sigma}_\upbeta = \mathbf{UD^2U}', \tag{32}$$

and consider that $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{D}}$ are subsets of $\mathbf{U}$ and $\mathbf{D}$. The XFA matrix is obtained with

$$\mathbf{\Sigma}_{\upbeta(\mathbf{XFA})} = \tilde{\mathbf{U}}\tilde{\mathbf{D}}^{\mathbf{2}}\tilde{\mathbf{U}}' + \mathbf{N}, \tag{33}$$

where $\mathbf{N}$ is a diagonal matrix that recovers the original diagonal elements of $\mathbf{\Sigma}_\upbeta$.

2.8 Kernels and rotations

Many modern genomic prediction methods are based on non-linear relationships that capture more variance than additive models (Montesinos-López et al., 2021, de los Campos et al., 2013, de Los Campos et al., 2010). Rotation of kernels through spectral decomposition, eigen-value (EVD) or singular-value decompositions (SVD), enables solving such models through the Gauss-Seidel framework (Legarra and Misztal, 2008). Rotations are also necessary when the number of parameters far exceeds the number of genotypes, as it can substantially reduce the dimensionality of the problem (Ødegård et al., 2018, Xavier and Habier, 2022). Genomic prediction based on any generalized kernel $\mathbf{K}$ can be described as

$$\begin{aligned} \mathbf{y} &= \mu + \mathbf{g} + \mathbf{e}, \\ \mathbf{g} &\sim N(0, \mathbf{K}\sigma_g^2), \\ \mathbf{e} &\sim N(0, \mathbf{I}\sigma_e^2), \end{aligned} \tag{34}$$

One can use EVD to decompose the kernel that describes genetic relationships as

$$\begin{aligned} \mathbf{K} &= \mathbf{UD^2U}' \\ &= (\mathbf{UD})(\mathbf{UD})' \\ &= \mathbf{QQ}', \end{aligned} \tag{35}$$

where $\mathbf{Q} = \mathbf{UD}$, so that one can reparameterize the model as

$$\begin{aligned} \mathbf{y} &= \mu + \mathbf{Q}\alpha + \mathbf{e}, \\ \alpha &\sim N(0, \mathbf{I}\sigma_g^2), \\ \mathbf{e} &\sim N(0, \mathbf{I}\sigma_e^2), \end{aligned} \tag{36}$$

yielding a breeding value solution $\mathbf{g} = \mathbf{Q}\alpha$, which preserves the probabilistic description of the original model as $\mathbf{QQ}'\sigma_g^2 = \mathbf{K}\sigma_g^2$. The matrix $\mathbf{Q}$ can be recovered by rotating the relationship matrix. Let the rotation matrix be defined as

$$\mathbf{R} = \mathbf{UD^{-1}}, \tag{37}$$

that enables the recovery of the design matrix with rotated parameters

$$\mathbf{Q} = \mathbf{GR}, \tag{38}$$

as their equivalence is shown through

$$\begin{aligned} \mathbf{Q} &= \mathbf{GR} \\ &= \mathbf{UD^2U'UD^{-1}} \\ &= \mathbf{UD^2D^{-1}} \\ &= \mathbf{UD} \end{aligned} \tag{39}$$

as $\mathbf{U'U} = \mathbf{I}$ and $\mathbf{D^2D^{-1}} = \mathbf{UD^{(2-1)}} = \mathbf{D}$.

For the computation of breeding values for any prediction set (PS), based on the relationship between ES and PS is known ($\mathbf{K_{PS,ES}}$), the rotation matrix computed from the genotypes observed in the estimation set ($\mathbf{R_{ES}}$), and regression coefficients estimated from the estimation set (ES). The design matrix is obtained with

$$\mathbf{Q_{PS|ES}} = \mathbf{K_{PS,ES}R_{ES}}, \tag{40}$$

and the breeding values are obtained with the regression coefficients estimated from the ES (eq. 36). Thus,

$$\hat{\mathbf{g}}_{\mathbf{PS}} = \mathbf{Q_{PS|ES}}\hat{\alpha}. \tag{41}$$

2.9 Rotation for reducing dimensions

When fitting a linear additive model ($\mathbf{g} = \mathbf{Z}\beta$) with a large number of parameters ($p >> n$), one can use the kernel trick to reduce the model dimensionality by kernalizing the genomic information with $\mathbf{K} = \mathbf{ZZ'}$, and subsequently decomposing $\mathbf{K}$ via EVD (eq. 35). Marker effects are recovered with

$$\beta = \mathbf{Z'R}\alpha, \tag{42}$$

where the rotation matrix ($\mathbf{R}$, eq. 37) originates from the decomposition of $\mathbf{K}$. This approach reduces the number of parameters from p to n. When n is also large, one can create the rotation $\tilde{\mathbf{R}}$ from the subset ($\tilde{n} < n$) as described in equation 37. The full design matrix is created as

$$\begin{aligned} \mathbf{Q_{n|\tilde{n}}} &= \mathbf{K_{n,\tilde{n}}}\tilde{\mathbf{R}} \\ &= \mathbf{Z}\tilde{\mathbf{Z}}'\tilde{\mathbf{R}} \end{aligned} \tag{43}$$

where $\tilde{\mathbf{Z}}$ and $\tilde{\mathbf{R}}$ herein represent the genotypes and rotation matrix generated from a subset of individuals. Under this framework, the dimensionality of the matrix is reduced to less than the original number of parameters and observations.

2.10 Rotations using SVD

The kernel rotation parameterizations aforementioned, $\mathbf{Q}$ and $\mathbf{Q_{n|\tilde{n}}}$ can be also derived directly from the single-value decomposition $\mathbf{Z}$ as

$$\mathbf{Z} = \mathbf{UDV'}, \tag{44}$$

which yields the same $\mathbf{U}$ and $\mathbf{D}$ from equation 35. Principal components are obtained either with $\mathbf{Q} = \mathbf{UD}$ or $\mathbf{Q} = \mathbf{ZV}$. Subset rotation (eq. 43) are obtained with $\mathbf{Q_{n|\tilde{n}}} = \mathbf{Z_nV_{\tilde{n}}}$, where $\mathbf{V_{\tilde{n}}}$ comes from the SVD of a population subset. Models parameterized by $\mathbf{Q}\alpha$ (eq. 36) using SVD recover the marker effects (eq. 3) with $\beta = \mathbf{V'}\alpha$.

2.11 Diagonalization

Diagonalization refers to converting a dense structure into a diagonal matrix, thus suiting relationship models. Under $\mathbf{g} \sim N(0, \mathbf{K}\sigma^2_g)$, the EVD of the relationship matrix allows to parameterized the genetic term as either the regression of eigenvectors as ($g = \mathbf{U}\theta$), resulting on the following linear model

$$\begin{aligned} \mathbf{y} &= \mu + \mathbf{U}\theta + \mathbf{e}, \\ \theta &\sim N(0, \mathbf{D}^2\sigma^2_g), \\ \mathbf{e} &\sim N(0, \mathbf{I}\sigma^2_e), \end{aligned} \tag{45}$$

which differs from the equation 36 because $Var(\alpha) = \mathbf{I}\sigma^2_g$ whereas $Var(\theta) = \mathbf{D}^2\sigma^2_g$. Equation 45 is further adapted with

$$\begin{aligned} \mathbf{y} &= \mu + \mathbf{U}\theta + \mathbf{e} \\ \mathbf{U'y} &= \mathbf{U'}\mu + \mathbf{U'U}\theta + \mathbf{U'e} \\ \mathbf{U'y} &= \mathbf{U'}\mu + \theta + \mathbf{e}, \end{aligned} \tag{46}$$

since $\mathbf{U'U} = \mathbf{I}$. Residuals are unaffected as $Var(\mathbf{e}) = \mathbf{U'U}\sigma^2_e = \mathbf{I}\sigma^2_e$. Computational advantages provided by equation 46 include the sparsity design matrices and kernels, and it allows for dimensionality reduction by not using all eigenvectors. For the multivariate case, diagonalization makes it feasible to solve variance components using REML using commercial software (Gilmour et al., 2017) when data is balanced (*i.e.,* all genotypes are observed in all locations), and this approach has become particularly useful for genome-wide association methods, both univariate and multivariate (Zhou and Stephens, 2012, 2014).

2.12 Sparse inversion of kernels

Kernel-based models (eq. 34) are commonly solved through the inversion of $\mathbf{K}$. When the number of genotyped individuals is large, the algorithm for proven and young (APY, Pocrnic et al., 2016, Bermann et al., 2022) provides a sparse representation of the inverse genomic relationship matrices. Under APY, dense inversion is performed only on a subset of the relationship matrix containing a representative sample of individuals referred to as the core set. Thus,

$$\mathbf{K}^{-1}_{\mathbf{APY}} = \begin{bmatrix} \mathbf{K}^{-1}_{\mathbf{cc}} + \mathbf{P_{cn}}\mathbf{M}^{-1}_{\mathbf{nn}}\mathbf{P_{nc}} & -\mathbf{P_{cn}}\mathbf{M}^{-1}_{\mathbf{nn}} \\ -\mathbf{M}^{-1}_{\mathbf{nn}}\mathbf{P_{nc}} & \mathbf{M}^{-1}_{\mathbf{nn}} \end{bmatrix}, \tag{47}$$

where $\mathbf{c}$ and $\mathbf{n}$ describe the core and non-core set of genotypes, respectively, $\mathbf{P} = \mathbf{K_{nc}}\mathbf{K}^{-1}_{\mathbf{cc}}$, and $\mathbf{M}$ is a diagonal matrix with the element-wise Schur complement, as $\mathbf{m}_{ii} = \{\mathbf{k_{ii}} - \mathbf{k}_{ic}\mathbf{K}^{-1}_{\mathbf{cc}}\mathbf{k}_{ci}\}$.

3 Scalability of different parameterizations

A general summary of the scalability of the different parameterization is provided in Table 1. The table showcases that no method is completely scalable in all scenarios. For example, as the number of markers increases, SVD ($\mathbf{Q}\alpha$) and relationship-based methods ($\mathbf{ZZ}'$) are preferred over SNP-regressions ($\mathbf{Z}\beta$ and $\mathbf{Z}'\mathbf{Z}$). That is the case when genotype-by-sequencing (GBS) data is deployed. Conversely, datasets with more genotypes than markers are common when SNP arrays are utilized in large studies (Song et al., 2017), for multiple breeding programs, over multiple years (Allen et al., 2017), benefits from SNP regression as the dimensionality of models relies less on the number of observations. When the dataset displays a large number of genotypes and markers, dimensionality reduction (*e.g.*, $\mathbf{Q}_{n|\tilde{n}}\alpha$ and $\mathbf{K}^{-1}_{\mathbf{APY}}$) provides computational feasibility without loss in accuracy, as long as there are enough principal components to capture the genetic diversity (Pocrnic et al., 2019).

Table 1: Scalability

	Parameterization	Solver	No. of genotypes	No. of markers	No. of traits
1	$\mathbf{Z}'\mathbf{Z}$ equation	REML/BGS	****	*	*
2	$\mathbf{ZZ}'$ equation	REML/BGS	**	****	*
3	$\mathbf{K}^{-1}_{\mathbf{APY}}$	REML/BGS	***	****	*
4	$\mathbf{U}\theta$	REML/BGS	**	****	**
5	$\mathbf{\Psi Y}$	BGS	**	**	**
6	$\mathbf{Q}\alpha$	PEGS	**	***	***
7	$\mathbf{Q}_{n\|\tilde{n}}\alpha$	PEGS	****	***	***
8	$\mathbf{Z}\beta$	PEGS	**	**	***
9	$\mathbf{Z}\beta$ (BayesABC)	BGS	**	**	*
10	$\mathbf{F\Lambda}$ (MegaLMM)	BGS	*	****	****
11	$\mathbf{F\Lambda}$ (MegaSEM)	PEGS	***	**	****

Technical guidelines for modeling large datasets with multiple traits have been provided by Misztal, 2008. In that study, the author recommends starting from running univariate analysis and subsequently progressing to multivariate models. At the time, the modeling of hundreds of traits has not been considered possible up to recent developments, because the computational cost of REML methods increases n^2 and k^3 (Misztal, 2008) with the number of genotypes (n) and traits (k), and more efficient methods, such as canonical transformation, would require balanced data.

Bayesian Gibbs sampling (BGS) is the mainstream alternative to REML (Sorensen et al., 2002), as it provides a more computationally stable framework to estimate variance components and regression coefficients at low memory cost. However, BGS may take a long time to run as it requires a large number of Markov Chain Monte Carlo (MCMC) samples to provide satisfying convergence.

Efficient alternatives to MCMC are available when the variance components are known as *priori* and only coefficients need to be inferred, those include PCG and Gauss-Seidel Legarra and Misztal, 2008, Misztal and Legarra, 2017. Only a few software implement Gauss-Seidel as the main approach to estimate marker effects (Legarra et al., 2011). Herein, we assessed the PEGS solver from Xavier and Habier, 2022 as an approach to use Gauss-Seidel while estimating variance components. While computationally efficient, the PEGS solver is unsuitable for the computation of accuracies and confidence intervals as it neither computes the left-hand side of the mixed model equation nor samples the coefficients in a stochastic manner.

4 Sparse testing benchmark

One hundred environments were simulated to test the predictive performance of efficient approaches, with varying amounts of missing values per environment (75% and 90%), different levels of heritability ($h^2 = \{0.2, 0.4, 0.6\}$), and different levels of correlation among environments ($\rho_{GxE} = \{0.2, 0.4, 0.6\}$). The accuracy was measured as the correlation between simulated and true breeding values within environments. The simulated dataset used real genomic information from six soybean families from the SoyNAM panel (Xavier et al., 2018, Diers et al., 2018, Song et al., 2017), where each population consists of approximately 140 individuals genotyped with 4240 markers, where all individuals are either full- or half-siblings. Genomic data is available in the R package `mas`.

MegaLMM is implemented in the R package MegaLMM (Runcie et al., 2021), and other prediction methods and simulation functions are implemented in the R package bWGR (Xavier et al., 2019). We used functions `MRR3F` (UV-, MV- and XFA-PEGS) and `XSEMF` (MegaSEM). XFA model reconstructed the covariance matrix with 3 eigenpairs. Coefficients and variance components are estimated via PEGS for all methods except for MegaLMM, which uses Bayesian Gibbs Sampling.

Results are presented in Figure 1. Benefits from methods that capture GxE increase with the percentage of missing values and genetic correlations. Among the multivariate methods, MegaLMM and XFA-PEGS performed best, closely followed by the multivariate PEGS, and then by MegaSEM and univariate. The simulated results indicate that methods that rely on simplified covariance structures (*i.e.*, MegaLMM and XFA-PEGS) outperformed the multivariate model with more complex covariance under sparse testing. That may be attributed to the precision of covariance estimates and the numerical stability of $\mathbf{\Sigma}_\beta$ as the number of environments increases, indicating that lowering the dimensionality of large GxE models may benefit predictions.

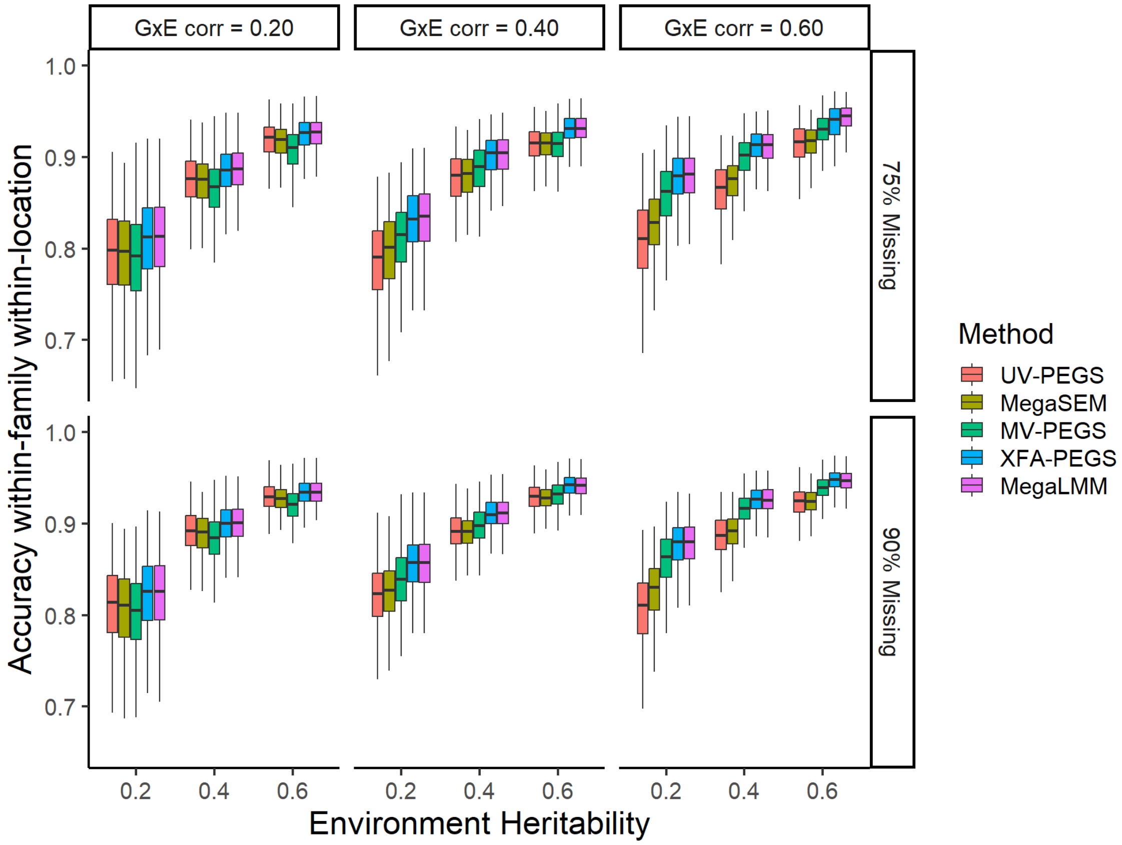

Figure 1: Prediction accuracy within-population within-environment on 100 simulated environments, varying heritability, percentage of missing values, genotype-by-environment correlation. Genomic information was sourced from six SoyNAM families.

Univariate prediction outperformed MegaSEM and MV-PEGS in the scenario with low heritability and low GxE correlation. This trend was also observed by Xavier and Habier, 2022, which is possibly associated with a low signal-to-noise ratio where the GxE information does not contribute to the predictions.

5 Computational benchmark under dense testing

Runtime was assessed on five simulated scenarios varying the number of environments and individuals. A simulated bi-parental F2 population with varying numbers of individuals ($n = \{500, 2000, 20000\}$) and environments ($k = \{10, 50, 200\}$), and heritability within-environment set to 0.2. The genomic information was based on 10 chromosomes of 500 cM with one marker per cM. Runtime and accuracy were recorded. Runtime was recorded in terms of minutes to fit the model. Accuracy was recorded to contextualize runtime as faster methods may implicate loss of prediction capability, and it was measured as the mean correlation between predicted and true breeding values within environments.

The complete dataset allows for testing approaches that can only be tested under balanced conditions, including diagonalization (eq. 46) and canonical transformation (eq. 12, 13, 14). The following methods were evaluated: REML-based GBLUP model (GREML) implemented in asreml-r 4.2 and its diagonalized counterpart (Diag. GREML), MegaLMM, MegaSEM, univariate by environment (UV-PEGS), multivariate PEGS (MV-PEGS), and canonical transformation (CT-PEGS). The REML-based approaches were not tested under scenarios with more than 50 environments due to the runtime and lack of algorithmic stability. MegaLMM was not evaluated in the scenario with 20,000 individuals due to computational requirements, with two test runs surpassing 30 hours each.

Table 2: Average runtime in minutes (s.e.) for the balanced experimental design based on 10 simulated replicates. Five scenarios vary in terms of the number of environments and individuals (No. environments / No. individuals).

Scenario (Env/Ind)	10 / 500	10 / 2,000	50 / 2,000	200 / 2,000	200 / 20,000
GREML	46.75 (0.37)	172.61 (17.93)	-	-	-
Diag. GREML	0.06 (<0.1)	0.19 (<0.1)	8.32 (3.51)	-	-
MegaLMM	0.31 (0.01)	4.38 (0.06)	7.23 (1.19)	17.71 (4.02)	-
MegaSEM	<0.01 (<0.01)	0.01 (<0.01)	0.04 (<0.01)	0.14 (<0.01)	5.26 (0.07)
MV-PEGS	<0.01 (<0.01)	<0.1 (<0.01)	0.02 (<0.01)	9.12 (1.62)	82.22 (5.71)
XFA-PEGS	<0.01 (<0.01)	<0.1 (<0.01)	0.03 (<0.01)	0.49 (0.09)	rerunning
CT-PEGS	<0.01 (<0.01)	0.01 (<0.01)	0.04 (<0.01)	0.15 (0.01)	5.25 (0.05)
UV-PEGS	<0.01 (<0.01)	0.01 (<0.01)	0.04 (<0.01)	0.14 (<0.1)	5.20 (0.06)

Table 3: Within environment accuracy for the balanced experimental design based on 10 simulated replicates. Five scenarios vary in terms of the number of environments and individuals (No. environments / No. individuals).

Scenario (Env/Ind)	10 / 500	10 / 2,000	50 / 2,000	200 / 2,000	200 / 20,000
GREML	0.81 (0.03)	0.89 (<0.01)	-	-	-
MegaLMM	0.78 (0.04)	0.87 (<0.01)	0.87 (<0.01)	0.89 (<0.01)	-
MegaSEM	0.79 (0.04)	0.88 (<0.01)	0.89 (<0.01)	0.89 (<0.01)	0.96 (<0.01)
MV-PEGS	0.80 (0.03)	0.89 (<0.01)	0.89 (<0.01)	0.90 (<0.01)	0.96 (<0.01)
XFA-PEGS	0.80 (0.04)	0.89 (<0.01)	0.89 (<0.01)	0.89 (<0.01)	rerunning
CT-PEGS	0.81 (0.03)	0.89 (<0.01)	0.88 (<0.01)	0.87 (<0.01)	0.95 (<0.01)
UV-PEGS	0.78 (0.04)	0.87 (<0.01)	0.87 (<0.01)	0.87 (<0.01)	0.95 (<0.01)

Results are displayed in Tables 2 and 3. All methods were substantially faster than GREML, while only CT-PEGS and MV-PEGS provided the same accuracy as GREML in the scenarios where all methods were evaluated. Diagonalization considerably decreased the runtime of GREML. The accuracy of the diagonalized REML was not included in Table 3 due to odd results attributed to singularities in the AI matrix. The most computationally efficient methods were UV-PEGS, CT-PEGS, and MegaSEM. The runtime of MegaLMM was sensitive to the number of individuals and MV-PEGS to the number of environments. The accuracy of univariate predictions was insensitive to the number of environments as it does not capture any GxE information. All methods that capture GxE information were as predictive or better than univariate, although the difference in predictive performance declined as the number of individuals increased. Under a balanced scenario XFA- and MV-PEGS provided slightly higher accuracy than MegaLMM.

CT-PEGS provided approximately the same runtime as UV-PEGS as it consists of running univariate models in the transformed spaces. CT was advantageous over univariate for scenarios with up to 50 environments, and we attribute this decline in predictive ability to the numeric precision, as the scale of canonical traits becomes smaller and smaller as the number of traits increases. MegaSEM provided a runtime similar to univariate and CT, with accuracy that was lower than CT in scenarios with under 50 environments, but displaying accuracy comparable to MV-PEGS, XFA-PEGS, and MegaLMM for scenarios with more traits. Among the methods parameterized to capture GxE, MegaSEM displayed the lowest runtime. However, the accuracy of MegaSEM was sensitive to the sparseness of the data, as it did not perform well in the sparse simulations (Fig. 1).

When taking both runtime and accuracy into account, our results suggest that the best method depends on the dimensionality of the data. CT and diagonalization should be considered when data is completely balanced data and the number of environments is modest. MegaLMM suits datasets with more than one hundred traits but with a moderate number of individuals, five thousand or less. MV-PEGS and XFA-PEGS are suitable for datasets with up to 200 environments, due to the non-linear increase in runtime. MegaSEM suits scenarios with a large number of individuals and traits, as well as for analyses are time-sensitive.

6 Prediction of unobserved environments

A schematic evolution of methods integrating genomics and environmental information is provided by Crossa et al. [2022], including crop models and reaction norms. The covariance is inferred by the environmental parameters and, consequently, confined to that parameter space. Alternatively, the associations between environmental factors and marker effects can be inferred in subsequent analysis. For instance, Della Coletta et al., 2023 generated genotype-by-environment interaction networks from the correlation of the principal components of marker effects and the principal components of environmental covariates.

Post-hoc modeling of the factors responsible for genotype-by-environment interactions can be built from the output of unstructured models. Unlike crop models and reaction norms, it does not assume that all interactions can be explained by the environmental parameters available for modeling. The approach described herein works by modeling covariances and, subsequently, generates predictions using conditional expectations.

Consider a scenario where a set of individuals A, observed in a set of environments X, is used to predict a new set of individuals B observed in a set of environments Z. The estimated marker effects from prediction models are based on the observed data AX. The prediction of B individuals in observed environments is given by

$$\hat{\mathbf{G}}_{BX} = \mathbf{Z}_B \hat{\mathbf{B}}_{AX}, \tag{48}$$

where $\hat{\mathbf{G}}_{BX}$ is the matrix of GEBV of B individuals on X environments, $\mathbf{Z}_B$ is the marker information for B individuals, and $\hat{\mathbf{B}}_{AX}$ is the matrix of marker effects for X environments. The next step consists of projecting B individuals into Z environments. That is attained with the conditional expectation, where Z environments are predicted from a linear combination of X environments. Thus,

$$\hat{\mathbf{G}}_{BZ|BX} = \mathbf{\Sigma}_{ZX} \hat{\mathbf{\Sigma}}_X^{-1} \hat{\mathbf{G}}_{BX}, \tag{49}$$

where $\mathbf{\Sigma}_X$ is the genetic variance-covariance matrix of X environments, and $\mathbf{\Sigma}_{ZX}$ is the covariance matrix between X and Z environments. Note that $\mathbf{\Sigma}_X$ is estimated from the multivariate model (eq. 25) that estimated the marker effects, since $\mathbf{B}_{AX} \sim N(0, \mathbf{\Sigma}_X \otimes \mathbf{I})$. Estimating $\mathbf{\Sigma}_{ZX}$ requires prediction if Z has not been observed.

The prediction of $\mathbf{\Sigma}_{ZX}$ can be inferred from parameters that drive genotype-by-environment interaction. Let

$$\begin{aligned} \mathbf{\Sigma}_X &= \mathbf{U}_X \mathbf{D}_X^2 \mathbf{U}_X' \\ \mathbf{\Sigma}_X &= \mathbf{Q}_X \mathbf{Q}_X' \end{aligned} \tag{50}$$

where $\mathbf{Q}_X = \mathbf{U}_X \mathbf{D}_X$. Now, assuming that the principal components $\mathbf{Q}_X$ can be modeled as a linear function of parameters that drive genotype-by-environment interaction ($\mathbf{W}_X$), we obtain

$$\mathbf{Q}_X = \mathbf{W}_X \mathbf{\Omega}_X + \mathbf{E}_X \tag{51}$$

where $\mathbf{W}_X$ is the design matrix of explanatory variables, $\mathbf{\Omega}_X$ is the matrix of regression coefficients, $\mathbf{E}_X$ is the matrix of residuals. Note that equation 51 acts as a *post hoc* modeling of environmental reaction norms, such that if the same set of parameters is known for environments Z, principal components can be predicted using $\mathbf{W}_Z$. Thus,

$$\hat{\mathbf{Q}}_{ZX} = \mathbf{W}_Z \hat{\mathbf{\Omega}}_X \tag{52}$$

and the covariance between environments X and Z can be inferred as

$$\hat{\mathbf{\Sigma}}_{ZX} = \hat{\mathbf{Q}}_{ZX} \mathbf{Q}_X'. \tag{53}$$

Note that equation 51 utilizes a linear model to fit the eigenstructure of the genotype-by-environment covariance, however, the evaluation of non-parametric models (*e.g.*, random forest) is encouraged when the interaction patterns are complex beyond additivity (Alves et al., 2020, Waters et al., 2023).

7 Conclusion

The conclusion paragraph goes here. Reiterate the use of the word "megavariate". Set the next steps, including the goal to catch up with the accuracy of XFA-PEGS and MegaLMM but at the computational cost of MegaSEM.

References

Ani A Elias, Kelly R Robbins, RW Doerge, and Mitchell R Tuinstra. Half a century of studying genotype× environment interactions in plant breeding experiments. *Crop Science*, 56(5):2090–2105, 2016.

T H E Meuwissen, B J Hayes, and M E Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–1829, 2001.

D Habier, RL Fernando, and Jack CM Dekkers. The impact of genetic relationship information on genome-assisted breeding values. *Genetics*, 177(4):2389–2397, 2007.

Jose Crossa, Roberto Fritsche-Neto, Osval A Montesinos-Lopez, Germano Costa-Neto, Susanne Dreisigacker, Abelardo Montesinos-Lopez, and Alison R Bentley. The modern plant breeding triangle: optimizing the use of genomics, phenomics, and enviromics data. *Frontiers in plant science*, 12:651480, 2021.

Alencar Xavier and David Habier. A new approach fits multivariate genomic prediction models efficiently. *Genetics Selection Evolution*, 54(1):1–15, 2022.

Nicolas Heslot, Deniz Akdemir, Mark E Sorrells, and Jean-Luc Jannink. Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theoretical and applied genetics*, 127:463–480, 2014.

Craig Hardner. Exploring opportunities for reducing complexity of genotype-by-environment interaction models. *Euphytica*, 213(11):248, 2017.

Johannes WR Martini, Jose Crossa, Fernando H Toledo, and Jaime Cuevas. On hadamard and kronecker products in covariance structures for genotype× environment interaction. *The Plant Genome*, 13(3):e20033, 2020.

Jaime Cuevas, José Crossa, Víctor Soberanis, Sergio Pérez-Elizalde, Paulino Pérez-Rodríguez, Gustavo de los Campos, OA Montesinos-López, and Juan Burgueño. Genomic prediction of genotype× environment interaction kernel regression models. *The plant genome*, 9(3):plantgenome2016–03, 2016.

Diego Jarquín, José Crossa, Xavier Lacaze, Philippe Du Cheyron, Joëlle Daucourt, Josiane Lorgeou, François Piraux, Laurent Guerreiro, Paulino Pérez, Mario Calus, et al. A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and applied genetics*, 127:595–607, 2014.

Jose Crossa, Osval Antonio Montesinos-Lopez, Paulino Pérez-Rodríguez, Germano Costa-Neto, Roberto Fritsche-Neto, Rodomiro Ortiz, Johannes WR Martini, Morten Lillemo, Abelardo Montesinos-Lopez, Diego Jarquin, et al. Genome and environment based prediction models and methods of complex traits incorporating genotype× environment interaction. *Genomic Prediction of Complex Traits: Methods and Protocols*, pages 245–283, 2022.

Daniela Bustos-Korts, Marcos Malosetti, Scott Chapman, and Fred van Eeuwijk. Modelling of genotype by environment interaction and prediction of complex traits across multiple environments as a synthesis of crop growth modelling, genetics and statistics. *Crop Systems Biology: Narrowing the gaps between crop modelling and genetics*, pages 55–82, 2016.

Marcos Malosetti, Jean-Marcel Ribaut, and Fred A van Eeuwijk. The statistical analysis of multi-environment data: modeling genotype-by-environment interaction and its genetic basis. *Frontiers in physiology*, 4:37433, 2013.

Arthur R Gilmour. Average information residual maximum likelihood in practice. *Journal of animal breeding and genetics*, 136(4):262–272, 2019.

Daniel E Runcie, Jiayi Qu, Hao Cheng, and Lorin Crawford. Megalmm: mega-scale linear mixed models for genomic predictions with thousands of traits. *Genome biology*, 22:1–25, 2021.

Ivan Pocrnic, Daniela AL Lourenco, Yutaka Masuda, and Ignacy Misztal. Dimensionality of genomic information and performance of the algorithm for proven and young for different livestock species. *Genetics Selection Evolution*, 48:1–9, 2016.

J Möhring and H-P Piepho. Comparison of weighting in two-stage analysis of plant breeding trials. *Crop Science*, 49 (6):1977–1988, 2009.

Hans-Peter Piepho, Jens Möhring, Torben Schulz-Streeck, and Joseph O Ogutu. A stage-wise approach for the analysis of multi-environment trials. *Biometrical Journal*, 54(6):844–860, 2012.

Shizhong Xu. Theoretical basis of the beavis effect. *Genetics*, 165(4):2259–2268, 2003.

David Habier, Rohan L Fernando, and Dorian J Garrick. Genomic blup decoded: a look into the black box of genomic prediction. *Genetics*, 194(3):597–607, 2013.

LR Schaeffer. Pseudo expectation approach to variance component estimation. *Journal of Dairy Science*, 69(11): 2884–2889, 1986.

Andres Legarra and I Misztal. Computing strategies in genome-wide selection. *Journal of dairy science*, 91(1): 360–366, 2008.

Anna Ma, Deanna Needell, and Aaditya Ramdas. Convergence properties of the randomized extended gauss–seidel and kaczmarz methods. *SIAM Journal on Matrix Analysis and Applications*, 36(4):1590–1604, 2015.

PM VanRaden and YC Jung. A general purpose approximation to restricted maximum likelihood: the tilde-hat approach. *Journal of Dairy Science*, 71(1):187–194, 1988.

Karin Meyer. Maximum likelihood estimation of variance components for a multivariate mixed model with equal design matrices. *Biometrics*, pages 153–165, 1985.

KV Konstantinov and GJ Erasmus. Using transformation algorithms to estimate (co) variance components by reml in models with equal design matrices. *South African Journal of Animal Science*, 23(5-6):187–191, 1993.

Daniel Gianola and Daniel Sorensen. Quantitative genetic models for describing simultaneous and recursive relationships between phenotypes. *Genetics*, 167(3):1407–1424, 2004.

Bruno D Valente, Guilherme JM Rosa, Daniel Gianola, Xiao-Lin Wu, and Kent Weigel. Is structural equation modeling advantageous for the genetic improvement of multiple traits? *Genetics*, 194(3):561–572, 2013.

JF Hayes and WG Hill. Modification of estimates of parameters in the construction of genetic selection indices ('bending'). *Biometrics*, pages 483–493, 1981.

Karin Meyer. "bending" and beyond: Better estimates of quantitative genetic parameters? *Journal of animal breeding and genetics*, 136(4):243–251, 2019.

Ismo Strandén and DJ Garrick. Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *Journal of dairy science*, 92(6):2971–2975, 2009.

Abelardo Montesinos-López, Osval Antonio Montesinos-López, José Cricelio Montesinos-López, Carlos Alberto Flores-Cortes, Roberto de la Rosa, and José Crossa. A guide for kernel generalized regression methods for genomic-enabled prediction. *Heredity*, 126(4):577–596, 2021.

Gustavo de los Campos, John M Hickey, Ricardo Pong-Wong, Hans D Daetwyler, and Mario PL Calus. Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics*, 193(2):327–345, 2013.

Gustavo de Los Campos, Daniel Gianola, Guilherme JM Rosa, Kent A Weigel, and José Crossa. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel hilbert spaces methods. *Genetics Research*, 92(4):295–308, 2010.

Jørgen Ødegård, Ulf Indahl, Ismo Strandén, and Theo HE Meuwissen. Large-scale genomic prediction using singular value decomposition of the genotype matrix. *Genetics Selection Evolution*, 50(1):1–12, 2018.

AR Gilmour, DG Butler, BR Cullis, BJ Gogel, and R Thompson. Asreml-r reference manual version 4. *VSN International Ltd, Hemel Hempstead, HP1 1ES, UK*, 2017.

Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics*, 44(7):821–824, 2012.

Xiang Zhou and Matthew Stephens. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods*, 11(4):407–409, 2014.

Matias Bermann, Daniela Lourenco, Natalia S Forneris, Andres Legarra, and Ignacy Misztal. On the equivalence between marker effect models and breeding value models and direct genomic values with the algorithm for proven and young. *Genetics Selection Evolution*, 54(1):52, 2022.

Qijian Song, Long Yan, Charles Quigley, Brandon D Jordan, Edward Fickus, Steve Schroeder, Bao-Hua Song, Yong-Qiang Charles An, David Hyten, Randall Nelson, et al. Genetic characterization of the soybean nested association mapping population. *The Plant Genome*, 10(2):plantgenome2016–10, 2017.

Alexandra M Allen, Mark O Winfield, Amanda J Burridge, Rowena C Downie, Harriet R Benbow, Gary LA Barker, Paul A Wilkinson, Jane Coghill, Christy Waterfall, Alessandro Davassi, et al. Characterization of a wheat breeders' array suitable for high-throughput snp genotyping of global accessions of hexaploid bread wheat (triticum aestivum). *Plant biotechnology journal*, 15(3):390–401, 2017.

Ivan Pocrnic, Daniela AL Lourenco, Yutaka Masuda, and Ignacy Misztal. Accuracy of genomic blup when considering a genomic relationship matrix based on the number of the largest eigenvalues: a simulation study. *Genetics Selection Evolution*, 51:1–10, 2019.

IJJOaB Misztal. Reliable computing in estimation of variance components. *Journal of animal breeding and genetics*, 125(6):363–370, 2008.

Daniel Sorensen, Daniel Gianola, and Daniel Gianola. *Likelihood, Bayesian and MCMC methods in quantitative genetics*. Springer, 2002.

I Misztal and Andres Legarra. Invited review: efficient computation strategies in genomic selection. *Animal*, 11(5): 731–736, 2017.

A Legarra, A Ricard, and O Filangi. Gs3: Genomic selection, gibbs sampling, gauss-seidel (and bayescπ). *Paris, France: INRA*, 2011.

Alencar Xavier, Diego Jarquin, Reka Howard, Vishnu Ramasubramanian, James E Specht, George L Graef, William D Beavis, Brian W Diers, Qijian Song, Perry B Cregan, 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.

Brian W Diers, Jim Specht, Katy Martin Rainey, Perry Cregan, Qijian Song, Vishnu Ramasubramanian, George Graef, Randall Nelson, William Schapaugh, Dechun Wang, et al. Genetic architecture of soybean yield and agronomic traits. *G3: Genes, Genomes, Genetics*, 8(10):3367–3375, 2018.

Alencar Xavier, William Muir, and Katy Rainey. bwgr: Bayesian whole-genome regression. *Bioinformatics*, 36(6): 1957–1959, 2019. doi:10.1093/bioinformatics/btz794.

Rafael Della Coletta, Sharon E Liese, Samuel B Fernandes, Mark A Mikel, Martin O Bohn, Alexander E Lipka, and Candice N Hirsch. Linking genetic and environmental factors through marker effect networks to understand trait plasticity. *Genetics*, 224(4):iyad103, 2023.

Rodrigo Silva Alves, Marcos Deon Vilela de Resende, Camila Ferreira Azevedo, Fabyano Fonseca e Silva, João Romero do Amaral Santos de Carvalho Rocha, Andrei Caíque Pires Nunes, Antônio Policarpo Souza Carneiro, and Gleison Augusto dos Santos. Optimization of eucalyptus breeding through random regression models allowing for reaction norms in response to environmental gradients. *Tree Genetics & Genomes*, 16:1–8, 2020.

Dominic L Waters, Julius HJ van der Werf, Hannah Robinson, Lee T Hickey, and Sam A Clark. Partitioning the forms of genotype-by-environment interaction in the reaction norm analysis of stability. *Theoretical and Applied Genetics*, 136(5):99, 2023.