Genomic Selection Using Environmental Covariates Within an Integrated Factor Analytic Linear Mixed Model

Daniel Tolhurst (✉ D.J.Tolhurst@sms.edu.ac.uk)

The University of Edinburgh The Roslin Institute <https://orcid.org/0000-0002-4787-080X>

R. Chris Gaynor

Bayer Crop Science United States: Bayer CropScience LP

Brian Gardunia

Bayer Crop Science United States: Bayer CropScience LP

John Hickey

Bayer CropScience LP

Gregor Gorjanc

The University of Edinburgh The Roslin Institute

Research Article

Keywords: Environmental covariates, Factor analytic model, Multi-environment trials, Genotype by environment interaction, Genomic selection, Linear mixed models

Posted Date: October 21st, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-995887/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Genomic selection using environmental covariates within an integrated factor analytic linear mixed model

Daniel J. Tolhurst · R. Chris Gaynor · Brian Gardunia ·
John M. Hickey · Gregor Gorjanc

Received: date / Accepted: date

Key message The integration of environmental covariates within a single-stage factor analytic approach provides breeders with an informative and practical framework to utilise genotype by environment interaction for genomic selection.

Abstract This paper introduces a single-stage genomic selection approach which directly integrates environmental covariates within a special factor analytic framework. The factor analytic approach of [Smith et al. \(2001\)](#) is an effective method of analysis for multi-environment trial (MET) datasets, but has limited biological interpretation since the underlying factors are latent so the modelled genotype by environment interaction (GEI) is observable, rather than predictable. The advantage of using known environmental covariates, such as soil moisture and daily temperature, is that the modelled GEI becomes directly interpretable, and thence predictable. This paper develops a model for both predictable and observable GEI in terms of a joint set of known and latent factors, as well as non-genetic sources of variation within trials and environments. This single-stage approach is referred to as the integrated factor analytic linear mixed model (IFA-LMM). The IFA-LMM is demonstrated on a late-stage cotton breeding MET dataset from Bayer Crop Science. The results show that the environmental covariates explain 34.6% of the genetic variance across environments (compared to only 23.3% for a conventional regression model). This represents 92.7% of the crossover GEI. The latent factors then explain 40.7% of the genetic variance, which represents 87.6% of the non-crossover GEI. This demonstrates the ability of the IFA-LMM to model crossover and non-crossover GEI in a manner that is both informative and practical to plant breeding.

Keywords Environmental covariates · Factor analytic model · Multi-environment trials · Genotype by environment interaction · Genomic selection · Linear mixed models

D.J. Tolhurst ✉

The Roslin Institute and Royal (Dick) School of Veterinary Studies, University of Edinburgh, Easter Bush, Midlothian EH25 9RG, United Kingdom
E-mail: D.J.Tolhurst@sms.ed.ac.uk

1 Introduction

This paper introduces a single-stage genomic selection (GS) approach which directly integrates environmental covariates within a special factor analytic framework. The factor analytic approach of [Smith et al. \(2001\)](#) is an effective method of analysis for multi-environment trial (MET) datasets, including a parsimonious model for genotype by environment interaction (GEI) as well as non-genetic sources of variation. The advantage of using regressions on environmental covariates, such as soil moisture and daily temperature, is that the modelled GEI becomes directly interpretable, and thence predictable. The GS approach developed in this paper exploits the desirable features of both classes of model.

Genomic selection is a form of marker-assisted selection that can improve the genetic gain in animal and plant breeding programmes ([Meuwissen et al., 2001](#)). In plant breeding, however, GS is often restricted by the presence of GEI, that is the change in genotype response to a change in environment. There are two appealing features of using environmental covariates for GS. Meaningful biological interpretation can be ascribed to GEI and predictions can be obtained for any tested or untested genotype into any observed or unobserved environment.

Environmental covariates were first used for the analysis of MET datasets by [Yates and Cochran \(1938\)](#). Their work was later popularised by [Finlay and Wilkinson \(1963\)](#), and involved a fixed regression for yield on a set of environment means (covariates) with a specific response (intercept and slope) for each genotype. [Knight \(1970\)](#) extended this approach to explore GEI with regards to a more complex set of covariates, instead of environment means. These approaches have distinct limitations when applied to the analysis of MET datasets, however. Many of which can be overcome by implementing linear mixed models ([Smith et al., 2005](#)).

Environmental covariates were first implemented within a linear mixed model by [Piepho et al. \(1998\)](#). [Heslot et al. \(2014\)](#) extended this approach to GS using a random regression on environmental stress covariates derived from daily weather observations. At a similar time, [Jarquín et al. \(2014\)](#) adapted a reaction norm model for a very large set of correlated stress covariates. They found that the genotype (marker) by covariate interactions explained only 23% of the genetic variance across environments. These examples highlight the current limitations of using environmental covariates for GS. That is, they are often highly correlated and only explain a small proportion of GEI ([Brancourt-Huettel et al., 2000](#)).

The factor analytic approach of [Smith et al. \(2001\)](#) includes a multiplicative model for GEI in terms of a small number of common environmental factors. The common factors are effective for modelling repeatable GEI common to all environments in the MET dataset. Any non-repeatable GEI specific to individual environ-

ments is then captured by a set of genetic residuals. The model also bears similarities to the ordinary regressions with one important difference. The environment loadings (covariates) are estimated from the data as well as the genotype scores (slopes).

The factor analytic approach has been widely and successfully adopted for the analysis of MET datasets ([Oakey et al., 2016](#); [Ukrainetz et al., 2018](#)). The two main variants involve pedigree data ([Oakey et al., 2007](#)) or, in terms of GS, molecular marker data. [Tolhurst et al. \(2019\)](#) demonstrated a factor analytic linear mixed model for GS within a major Australian wheat breeding programme. This single-stage approach incorporates an effective model for genotype (marker) by environment interaction as well as non-genetic sources of variation within trials and environments. There is one limitation of this approach, however. The common factors are latent so the modelled GEI is only observable, rather than predictable. This has led to ad hoc post processing of the latent factors with known environmental covariates ([Oliveira et al., 2020](#)).

Until now, the analysis of MET datasets has involved only one set of known covariates or latent environmental factors. The aim of this paper is to extend the factor analytic linear mixed model in [Tolhurst et al. \(2019\)](#) to directly integrate known environmental covariates. This new approach is hereafter referred to as the integrated factor analytic linear mixed model (IFA-LMM). There are three appealing features of the IFA-LMM.

1. The IFA-LMM includes a multiplicative model for GEI in terms of a small number of known and latent environmental factors. This simultaneously reduces the dimension of the environmental covariates as well as the dimension of the environments.
2. The multiplicative model captures predictable GEI in terms of known environmental covariates. This enables direct biological interpretation of the underlying factors.
3. The multiplicative model also captures observable GEI in terms of latent environmental factors, which are orthogonal to the known covariates. This captures any remaining GEI not modelled by the covariates.

The IFA-LMM is demonstrated on a late-stage cotton breeding MET dataset from Bayer Crop Science.

2 Statistical Models

2.1 Preliminaries

Assume a general MET dataset comprises t field trials conducted across p environments, where $t \geq p$ and $t = \sum_{j=1}^p t_j$, with t_j being the number of trials in environment j . Also assume that phenotypic and marker data are available on v genotypes. The n -vector of phenotypic data on v genotypes and p environments is given

by $\mathbf{y} = (\mathbf{y}_1^\top, \mathbf{y}_2^\top, \dots, \mathbf{y}_p^\top)^\top$, where $\mathbf{y}_j = (\mathbf{y}_{j;1}^\top, \mathbf{y}_{j;2}^\top, \dots, \mathbf{y}_{j;t_j}^\top)^\top$ is the n_j -vector of data for environment j and $\mathbf{y}_{j;k}$ is the n_{jk} -vector of data for trial k in environment j . The length of $\mathbf{y}$ is therefore given by:

$$n = \sum_{j=1}^p \sum_{k=1}^{t_j} n_{jk} = \sum_{j=1}^p n_j.$$

Lastly, assume that environmental covariate data are available on q covariates for p_d environments. To simplify the derivations of the IFA model, it is assumed that $p = p_d$ and $p \geq q$, where p is the number of environments with phenotypic data. The $p \times q$ matrix of environmental covariate data is therefore given by $\mathbf{S} = [\mathbf{s}_1 \ \mathbf{s}_2 \ \dots \ \mathbf{s}_q]$, with columns given by the centred and scaled covariates for each environment. Appendix A presents a computational approach to fitting the IFA-LMM when $p > p_d$. The application to scenarios where $p < p_d$ is the topic of current research.

Linear mixed model

The linear mixed model for $\mathbf{y}$ can be written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}\mathbf{u} + \mathbf{Z}_p\mathbf{u}_p + \mathbf{e}, \quad (1)$$

where $\boldsymbol{\tau}$ is a p -vector of fixed effects with $n \times p$ design matrix $\mathbf{X}$ (assumed to have full column rank), $\mathbf{u}$ is a vp -vector of random *genotype by environment* (GE) effects with $n \times vp$ design matrix $\mathbf{Z}$, $\mathbf{u}_p$ is a vector of random non-genetic (peripheral) effects with design matrix $\mathbf{Z}_p$ and $\mathbf{e}$ is the n -vector of residuals.

The vector of fixed effects, $\boldsymbol{\tau}$, includes the mean parameter for each environment. This vector is fitted as fixed, rather than random, following a classical quantitative genetics approach in which the GE effects in different environments are regarded as different correlated traits (Falconer and Mackay, 1996). The vector of random non-genetic effects, $\mathbf{u}_p$, accommodates the plot structures of trials and environments. Further effects in $\boldsymbol{\tau}$ or $\mathbf{u}_p$ may relate to genotypes without marker data (Tolhurst et al., 2019) or spatial modelling (Gilmour et al., 1997), respectively.

Model assumptions

The vectors of random effects and residuals in Equation 1 are assumed to have a joint normal distribution:

$$\begin{bmatrix} \mathbf{u} \\ \mathbf{u}_p \\ \mathbf{e} \end{bmatrix} \sim \text{N} \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{G} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{R} \end{bmatrix} \right), \quad (2)$$

Following Tolhurst et al. (2019), $\mathbf{G}_p$ is diagonal with a separate variance for each set of non-genetic effects and $\mathbf{R}$ is block diagonal with a two-dimensional (separable) autoregressive process of order one for each environment (Gilmour et al., 1997). The form of $\mathbf{G}$ is developed below, but note that all variance matrices in Equation 2 are fitted at the environment level, not trial level. This completely aligns the non-genetic and residual variance models with the genetic variance model.

2.2 Conventional variance models for the GE effects

The GE effects are modelled using molecular marker data, and therefore referred to as the *additive* GE effects. This formulation is an extension of the univariate genomic BLUP model (Stranden and Garrick, 2009), in which:

$$\mathbf{u} = (\mathbf{I}_p \otimes \mathbf{M})\mathbf{u}_m \quad \text{and} \quad \mathbf{G} = \mathbf{G}_e \otimes \mathbf{M}\mathbf{M}^\top/m, \quad (3)$$

where $\mathbf{M} = [\mathbf{m}_1 \ \mathbf{m}_2 \ \dots \ \mathbf{m}_r]$ is a $v \times r$ design matrix with columns given by the centred marker genotypes, $\mathbf{u}_m$ is a rp -vector of additive marker by environment effects, $\mathbf{G}_e$ is a $p \times p$ variance matrix between environments and $\mathbf{K} = \mathbf{M}\mathbf{M}^\top/m$ is the $v \times v$ genomic relationship matrix between genotypes (VanRaden, 2008).

The integrated factor analytic model is developed below for $\mathbf{u}$. The application to additive GE effects modelled using pedigree data or non-additive GE effects is straightforward.

Random regression model

A random effects analogue to the regression model in Piepho et al. (1998) is initially considered. This model enables a classical quantitative genetics approach to modelling GEI, which is considered as the lack of correlation between environments.

The additive GE effects modelled using a random regression on environmental covariates are given by:

$$\begin{aligned} \mathbf{u} &= (\mathbf{1}_p \otimes \mathbf{I}_v)\mathbf{g} + (\mathbf{s}_1 \otimes \mathbf{I}_v)\boldsymbol{\gamma}_1 + \dots + (\mathbf{s}_q \otimes \mathbf{I}_v)\boldsymbol{\gamma}_q + \boldsymbol{\delta} \\ &= (\mathbf{1}_p \otimes \mathbf{I}_v)\mathbf{g} + (\mathbf{S} \otimes \mathbf{I}_v)\boldsymbol{\gamma} + \boldsymbol{\delta}, \end{aligned} \quad (4)$$

where $\mathbf{g}$ is a v -vector of genotype main effects (intercepts), $\mathbf{S} = [\mathbf{s}_1 \ \mathbf{s}_2 \ \dots \ \mathbf{s}_q]$ is the $p \times q$ matrix of column centred and scaled environmental covariates, $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, \boldsymbol{\gamma}_2^\top, \dots, \boldsymbol{\gamma}_q^\top)^\top$ is the corresponding vp -vector of genotype-specific regression coefficients (slopes) in which $\boldsymbol{\gamma}_i$ reflects the response of genotypes to the i^{th} environmental covariate and $\boldsymbol{\delta} = (\boldsymbol{\delta}_1^\top, \boldsymbol{\delta}_2^\top, \dots, \boldsymbol{\delta}_p^\top)^\top$ is a vp -vector of *residual* GE effects (deviations) specific to individual environments.

It is assumed that:

$$\begin{bmatrix} \mathbf{g} \\ \boldsymbol{\gamma} \\ \boldsymbol{\delta} \end{bmatrix} \sim \text{N} \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma_g^2 \mathbf{K} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma} \otimes \mathbf{K} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \boldsymbol{\Psi} \otimes \mathbf{K} \end{bmatrix} \right),$$

where σ_g^2 is the main effect variance, $\boldsymbol{\Sigma} = \oplus_{i=1}^q \sigma_i^2$ and $\boldsymbol{\Psi} = \oplus_{j=1}^p \psi_j$ are $q \times q$ and $p \times p$ diagonal variance matrices for covariates and environments, respectively, in which σ_i^2 is the variance for the i^{th} covariate and ψ_j is the variance for the j^{th} environment. Other joint variance matrices can be used for $\mathbf{g}$ and $\boldsymbol{\gamma}$, including those which ensure the regression is invariant to linear transformations in $\mathbf{S}$.

The variance matrix for $\mathbf{u}$ is therefore given by:

$$\begin{aligned} \mathbf{G} &= (\sigma_g^2 \mathbf{J}_p + \mathbf{S}\boldsymbol{\Sigma}\mathbf{S}^\top + \boldsymbol{\Psi}) \otimes \mathbf{K} \\ &= \mathbf{G}_e \otimes \mathbf{K}, \end{aligned} \quad (5)$$

where $\mathbf{J}_p$ is a $p \times p$ unit matrix.

Factor analytic model

The factor analytic model (Smith et al., 2001) for GS is now considered. This model is effective for modelling the covariances between GE effects in terms of a small number of latent environmental factors (Kelly et al., 2007). The number, k , is called the order of the model and let FAK denote the factor analytic model with this order.

The FAK model for the additive GE effects is given by:

$$\mathbf{u} = (\boldsymbol{\lambda}_1 \otimes \mathbf{I}_v) \mathbf{f}_1 + \dots + (\boldsymbol{\lambda}_k \otimes \mathbf{I}_v) \mathbf{f}_k + \boldsymbol{\delta} \\ = (\boldsymbol{\Lambda} \otimes \mathbf{I}_v) \mathbf{f} + \boldsymbol{\delta}, \quad (6)$$

where $\boldsymbol{\Lambda} = [\boldsymbol{\lambda}_1 \boldsymbol{\lambda}_2 \dots \boldsymbol{\lambda}_k]$ is a $p \times k$ matrix of latent environmental covariates (loadings), $\mathbf{f} = (\mathbf{f}_1^\top, \mathbf{f}_2^\top, \dots, \mathbf{f}_k^\top)^\top$ is the corresponding vk -vector of genotype-specific coefficients (scores) in which $\mathbf{f}_l$ reflects the response of genotypes to the l^{th} latent factor and $\boldsymbol{\delta}$ is the vp -vector of residual GE effects.

The factor analytic model for $\mathbf{u}$ is postulated as a linear combination of latent environment loadings (covariates) and genotype-specific scores (slopes) for each factor, plus residual GE effects (deviations) specific to individual environments. This highlights the analogy to the random regression model in Equation 4.

It is assumed that:

$$\begin{bmatrix} \mathbf{f} \\ \boldsymbol{\delta} \end{bmatrix} \sim N \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{I}_k \otimes \mathbf{K} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Psi} \otimes \mathbf{K} \end{bmatrix} \right),$$

where $\boldsymbol{\Psi} = \oplus_{j=1}^p \psi_j$ is the $p \times p$ specific variance matrix across environments.

The variance matrix for $\mathbf{u}$ is therefore given by:

$$\mathbf{G} = (\boldsymbol{\Lambda} \boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}) \otimes \mathbf{K} \\ = \mathbf{G}_e \otimes \mathbf{K}. \quad (7)$$

2.3 Integrated factor analytic model

The integrated factor analytic (IFA) model was motivated by the desire to ascribe direct biological interpretation to the latent factors underlying the conventional factor analytic model, and thence utilise repeatable GEI for prediction.

The model for the additive GE effects, $\mathbf{u}$, in Equation 6 can be modified to directly integrate known environmental covariates. The factor loadings, $\boldsymbol{\Lambda}$, and residual GE effects, $\boldsymbol{\delta}$, become parametrised by two orthogonal sources of GEI, that is known covariates and latent environmental factors. The IFA model is therefore postulated as a special factor analytic model, in which:

$$\boldsymbol{\Lambda} = \mathbf{S} \boldsymbol{\Lambda}_s + \boldsymbol{\Gamma} \boldsymbol{\Lambda}_r \quad \text{and} \quad \boldsymbol{\delta} = ([\mathbf{S} \ \boldsymbol{\Gamma}] \otimes \mathbf{I}_v) \boldsymbol{\gamma} \\ = \mathbf{B} \begin{bmatrix} \boldsymbol{\Lambda}_s \\ \boldsymbol{\Lambda}_r \end{bmatrix} \quad = (\mathbf{B} \otimes \mathbf{I}_v) \boldsymbol{\gamma}, \quad (8)$$

where $\mathbf{B} = [\mathbf{S} \ \boldsymbol{\Gamma}]$ is a $p \times p$ matrix of basis functions, $\mathbf{S} = [\mathbf{s}_1 \ \mathbf{s}_2 \ \dots \ \mathbf{s}_q]$ is the $p \times q$ matrix of column centred

and scaled environmental covariates, $\boldsymbol{\Gamma}$ is a $p \times (p - q)$ projection matrix and $\boldsymbol{\gamma}$ is a vp -vector of regression coefficients which reflect the specific response of genotypes to the known covariates and latent environmental factors.

The $p \times k$ matrix of joint factor loadings is given by:

$$\begin{bmatrix} \boldsymbol{\Lambda}_s \\ \boldsymbol{\Lambda}_r \end{bmatrix} = \begin{bmatrix} \boldsymbol{\lambda}_{s_1} \ \boldsymbol{\lambda}_{s_2} \ \dots \ \boldsymbol{\lambda}_{s_k} \\ \boldsymbol{\lambda}_{r_1} \ \boldsymbol{\lambda}_{r_2} \ \dots \ \boldsymbol{\lambda}_{r_k} \end{bmatrix}, \quad (9)$$

where $\boldsymbol{\Lambda}_s$ is a $q \times k$ matrix of loadings for the known environmental covariates and $\boldsymbol{\Lambda}_r$ is a $(p - q) \times k$ matrix of latent environment loadings. The factors underlying $\boldsymbol{\Lambda}_s$ and $\boldsymbol{\Lambda}_r$ are therefore referred to as the *known* and *latent* factors, respectively.

The projection matrix, $\boldsymbol{\Gamma}$, in Equation 8 ensures the known and latent factors are orthogonal. This is achieved by projecting $\boldsymbol{\Lambda}_r$ into the orthogonal complement to the space spanned by $\mathbf{S}$. A convenient choice is:

$$[\mathbf{I}_p - \mathbf{S}(\mathbf{S}^\top \mathbf{S})^{-1} \mathbf{S}^\top], \quad (10)$$

where each column is a p -vector which forms a basis for the orthogonal complement to the range in $\mathbf{S}$. The projection matrix is then taken as the first $(p - q)$ columns in Equation 10:

$$\boldsymbol{\Gamma} = [\mathbf{r}_1 \ \mathbf{r}_2 \ \dots \ \mathbf{r}_{p-q}],$$

where $\mathbf{S}^\top \boldsymbol{\Gamma} = \mathbf{0}$. This specification ensures that the IFA model has the same number of variance parameters as the conventional factor analytic model in Equation 6. When $p \gg q$, however, it may be desirable to use fewer than $p - q$ columns, and thence fewer variance parameters. This enables the IFA model to be scalable to very large p . Other forms of $\boldsymbol{\Gamma}$ can be used, including a diagonal matrix.

The IFA model for the additive GE effects is obtained by substituting the components in Equation 8 into Equation 6, which gives:

$$\mathbf{u} = ([\mathbf{S} \boldsymbol{\Lambda}_s + \boldsymbol{\Gamma} \boldsymbol{\Lambda}_r] \otimes \mathbf{I}_v) \mathbf{f} + ([\mathbf{S} \ \boldsymbol{\Gamma}] \otimes \mathbf{I}_v) \boldsymbol{\gamma} \\ = \left(\mathbf{B} \begin{bmatrix} \boldsymbol{\Lambda}_s \\ \boldsymbol{\Lambda}_r \end{bmatrix} \otimes \mathbf{I}_v \right) \mathbf{f} + (\mathbf{B} \otimes \mathbf{I}_v) \boldsymbol{\gamma}, \quad (11)$$

where the genotype-specific scores, $\mathbf{f}$, reflect the response of genotypes to the joint set of known and latent factors, that is the *joint* factors. Other forms of $\mathbf{f}$ can be used, including the extension to separate sets of factors.

The residual GE effects, $\boldsymbol{\delta}$, then model any remaining GEI not explained by the joint factors. The parameterisation for $\boldsymbol{\delta}$ in Equation 8 may not be sensible, however, particularly for higher order IFA models in which the percentage of variance explained by these effects is typically very small. This leads to a simplified IFA model given by:

$$\mathbf{u} = \left(\mathbf{B} \begin{bmatrix} \boldsymbol{\Lambda}_s \\ \boldsymbol{\Lambda}_r \end{bmatrix} \otimes \mathbf{I}_v \right) \mathbf{f} + \boldsymbol{\delta}, \quad (12)$$

which is analogous to the conventional factor analytic model in Equation 6, but with one important difference.

The linear combination of loadings and scores now involve a joint set of known and latent factors which are parametrised by known covariates and latent environmental factors, respectively. This directly introduces a multiplicative model for interpretable, and thence predictable GEI, into the factor analytic model.

Model assumptions

The random effects in Equation 12 are assumed to have a joint normal distribution:

$$\begin{bmatrix} \mathbf{f} \\ \boldsymbol{\delta} \end{bmatrix} \sim N \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{I}_k \otimes \mathbf{K} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Psi} \otimes \mathbf{K} \end{bmatrix} \right),$$

where $\boldsymbol{\Psi} = \oplus_{j=1}^p \psi_j$ is the $p \times p$ specific variance matrix across environments.

The IFA variance matrix for $\mathbf{u}$ is therefore given by:

$$\begin{aligned} \mathbf{G} &= \left(\mathbf{B} \begin{bmatrix} \boldsymbol{\Lambda}_s \boldsymbol{\Lambda}_s^\top & \boldsymbol{\Lambda}_s \boldsymbol{\Lambda}_r^\top \\ \boldsymbol{\Lambda}_r \boldsymbol{\Lambda}_s^\top & \boldsymbol{\Lambda}_r \boldsymbol{\Lambda}_r^\top \end{bmatrix} \mathbf{B}^\top + \boldsymbol{\Psi} \right) \otimes \mathbf{K} \\ &= \mathbf{G}_e \otimes \mathbf{K}, \end{aligned} \quad (13)$$

where $\mathbf{G}_e = (\boldsymbol{\Lambda} \boldsymbol{\Lambda}^\top + \boldsymbol{\Psi})$ and $\boldsymbol{\Lambda} = \mathbf{B} \begin{bmatrix} \boldsymbol{\Lambda}_s \\ \boldsymbol{\Lambda}_r \end{bmatrix}$.

The known and latent variance matrices between environments are then given by:

$$\mathbf{G}_s = (\mathbf{S} \boldsymbol{\Lambda}_s \boldsymbol{\Lambda}_s^\top \mathbf{S}^\top + \boldsymbol{\Psi}) \text{ and } \mathbf{G}_r = (\boldsymbol{\Gamma} \boldsymbol{\Lambda}_r \boldsymbol{\Lambda}_r^\top \boldsymbol{\Gamma} + \boldsymbol{\Psi}), \quad (14)$$

where the covariance between any pair of environments is different for each source of GEI. Each variance matrix is converted to a correlation matrix for interpretability.

Model features

The IFA variance matrix in Equation 13 is postulated such that the underlying factors model both known and latent sources of GEI, and therefore have direct biological interpretation. The maximum amount of variance is explained by the known environmental covariates, that is $\mathbf{S} \boldsymbol{\Lambda}_s \boldsymbol{\Lambda}_s^\top \mathbf{S}^\top$. Any remaining variance in the underlying factors is then explained by the latent environmental factors, $\boldsymbol{\Gamma} \boldsymbol{\Lambda}_r \boldsymbol{\Lambda}_r^\top \boldsymbol{\Gamma}$. The projection matrix, $\boldsymbol{\Gamma}$, ensures the known and latent factors are orthogonal. The covariance matrices, $\mathbf{S} \boldsymbol{\Lambda}_s \boldsymbol{\Lambda}_r^\top \boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma} \boldsymbol{\Lambda}_r \boldsymbol{\Lambda}_s^\top \mathbf{S}$, reflect the joint regression across both factors.

There are two important results.

1. When $p = q$; $\boldsymbol{\Lambda} \equiv \mathbf{S} \boldsymbol{\Lambda}_s$ and the additive GE effects are modelled exclusively with environmental covariates:

$$\mathbf{u} = (\mathbf{S} \boldsymbol{\Lambda}_s \otimes \mathbf{I}_v) \mathbf{f} + \boldsymbol{\delta}. \quad (15)$$

This model is a factor analytic regression on environmental covariates. Genotype main effects (intercepts) can be included where appropriate.

2. When $q = 0$; $\boldsymbol{\Lambda} \equiv \boldsymbol{\Lambda}_r$ and the model for $\mathbf{u}$ is given by the conventional factor analytic model in Equation 6, in which the underlying factors are not directly interpretable.

Maximum number of factors

The IFA variance matrix in Equation 13 can be compared to an equivalent unstructured form to obtain the maximum number of underlying factors allowed for a particular number of covariates and environments. This comparison can be made using a more general IFA model with different numbers of known and latent factors. Let these numbers be denoted by k_s and k_r , respectively, and let IFA k_s - k_r denote the IFA model with this order. The total number of joint factors is then given by $k = \max(k_s, k_r)$.

The maximum number of known and latent factors ensures that $\mathbf{G}$ is not over-parametrised. There are two common scenarios.

1. When $k_s = k_r$, the total number of joint factors must satisfy:

$$k \leq \min \left(p, \frac{1}{2} \left[2p + 1 - \sqrt{8p + 1} \right] \right).$$

This is the same condition required for the conventional factor analytic model (Smith, 1999).

2. When $k_s > k_r$, the matrix of joint factor loadings in Equation 9 is reformulated as:

$$\begin{bmatrix} \boldsymbol{\Lambda}_s \\ \boldsymbol{\Lambda}_r \end{bmatrix} = \begin{bmatrix} \lambda_{s_1} & \dots & \lambda_{s_{k_r}} & \lambda_{s_{k_r+1}} & \dots & \lambda_{s_{k_s}} \\ \lambda_{r_1} & \dots & \lambda_{r_{k_r}} & \mathbf{0} & \dots & \mathbf{0} \end{bmatrix}, \quad (16)$$

and the number of known factors must satisfy:

$$k_s \leq \min \left(q, \frac{1}{2} \left[2p + 1 - \sqrt{8p + 1 - 8\xi(p - q)} \right] \right),$$

$$\text{where } \xi = k_s - k_r \leq \frac{8p+1}{8(p-q)}.$$

Similar results can be obtained when $k_s < k_r$, but note that additional column vector/s of zeros would be concatenated to $\boldsymbol{\Lambda}_s$, instead of $\boldsymbol{\Lambda}_r$.

2.4 Model estimation

The final model is obtained by substituting Equation 12 into Equation 1, which gives:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_\Lambda \mathbf{f} + \mathbf{Z}\boldsymbol{\delta} + \mathbf{Z}_p \mathbf{u}_p + \mathbf{e}, \quad (17)$$

$$\text{where } \mathbf{Z}_\Lambda = \mathbf{Z} \left(\mathbf{B} \begin{bmatrix} \boldsymbol{\Lambda}_s \\ \boldsymbol{\Lambda}_r \end{bmatrix} \otimes \mathbf{I}_v \right).$$

This model is referred to as the integrated factor analytic linear mixed model (IFA-LMM). The form of the IFA-LMM in this paper is an extension of the genomic BLUP model, and therefore referred to as the *additive* IFA-LMM. The equivalent extension of ridge-regression BLUP is referred to as the *marker-additive* IFA-LMM. Typically $r \gg v$, so the computational burden in fitting the additive IFA-LMM is substantially less, and hence its inclusion in this paper.

Two alternative variants of the IFA-LMM can also be derived. The only difference between these two variants is replacing the genomic relationship matrix in Equation 13 with either a pedigree relationship matrix or a diagonal matrix, respectively. The extension

to joint modelling of additive and non-additive GE effects is then straightforward.

All variants of the IFA-LMM were coded in *R* (R Core Team, 2021) using open source libraries. Appendix B presents a computational approach for fitting the additive IFA-LMM used in this paper. This approach obtains residual maximum likelihood (REML) estimates of the variance parameters using an extension of the sparse formulation of the average information algorithm (Thompson et al., 2003). The REML estimates of the key variance parameters are given by $\begin{bmatrix} \hat{\Lambda}_s \\ \hat{\Lambda}_r \end{bmatrix}$ and $\hat{\Psi}$. The empirical BLUPs of the associated random effects are then given by $\tilde{\mathbf{f}}$ and $\tilde{\delta}$.

Identifiability and rotation of loadings

When $k > 1$, the REML estimate of the joint factor loadings, $\begin{bmatrix} \hat{\Lambda}_s \\ \hat{\Lambda}_r \end{bmatrix}$, is not unique so that $k(k-1)/2$ constraints are required to ensure identifiability during estimation. There are two distinct scenarios for the IFA-LMM.

1. When $k_s \geq k_r$, the $k_s(k_s-1)/2$ elements are constrained in the upper triangle of $\hat{\Lambda}_s$.
2. When $k_s < k_r$, the $k_r(k_r-1)/2$ elements are constrained in the upper triangle of $\hat{\Lambda}_r$, instead of $\hat{\Lambda}_s$.

The factor loadings and genotype scores are then rotated to a principal component solution for interpretability. This rotation is given by:

$$\begin{aligned} \hat{\Lambda}^* &= \mathbf{S}\hat{\Lambda}_s^* + \mathbf{\Gamma}\hat{\Lambda}_r^* \quad \text{and} \quad \tilde{\mathbf{f}}^* = (c\mathbf{V}^T \otimes \mathbf{I}_v)\tilde{\mathbf{f}}, \\ &= \mathbf{B} \begin{bmatrix} \hat{\Lambda}_s^* \\ \hat{\Lambda}_r^* \end{bmatrix} \end{aligned} \quad (18)$$

where $\hat{\Lambda}_s^* = c\hat{\Lambda}_s\mathbf{V}$ and $\hat{\Lambda}_r^* = c\hat{\Lambda}_r\mathbf{V}$.

The constant c is chosen as either 1 or -1 to ensure the majority of the first rotated loadings in $\hat{\Lambda}^*$ are positive (Smith and Cullis, 2018). The $k \times k$ rotation matrix $\mathbf{V}$ is obtained via the singular value decomposition:

$$\hat{\Lambda} = \mathbf{U}\mathbf{D}\mathbf{V}^T. \quad (19)$$

The rotation applied in Equation 18 ensures the first rotated factor accounts for the maximum amount of variance, the second accounts for the next largest amount and is orthogonal to the first, and so on. Other orthogonal rotations can be used, including the quartimax (Carroll, 1953) and varimax criteria (Kaiser, 1958).

2.5 Model selection

Model selection is achieved using a combination of formal and informal criteria, with emphasis on the former. Formal assessment of model fit is achieved using the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). Informal assessment is achieved

using the percentage of variance explained by the underlying factors. The overall percentage of variance explained by the known factors is given by:

$$\bar{v}_s = 100 \operatorname{tr}(\mathbf{S}\hat{\Lambda}_s^*\hat{\Lambda}_s^{*\top}\mathbf{S}^T) / \operatorname{tr}(\hat{\Lambda}^*\hat{\Lambda}^{*\top} + \hat{\Psi}), \quad (20)$$

and for individual environments by:

$$\mathbf{v}_s = 100 \operatorname{diag}(\mathbf{S}\hat{\Lambda}_s^*\hat{\Lambda}_s^{*\top}\mathbf{S}^T) \oslash \operatorname{diag}(\hat{\Lambda}^*\hat{\Lambda}^{*\top} + \hat{\Psi}). \quad (21)$$

Similar measures are obtained for the joint factors, that is $\bar{v}$ and $\mathbf{v}$, using the variance explained by $\hat{\Lambda}^*\hat{\Lambda}^*$. The final model order is typically chosen such that $\bar{v}_s$ and $\bar{v}$ are sufficiently high and the number of environments with low variance explained in $\mathbf{v}_s$ and $\mathbf{v}$ is small. This may require different numbers of known and latent factors (see Section 2.3).

Specific examples of the measures above are presented in Section 3.

2.6 Model summaries and interpretation

The main limitation of the conventional factor analytic model is that the underlying factors are latent so they do not have direct biological interpretation. The IFA model overcomes this limitation since it integrates known covariates, as well as latent environmental factors.

The empirical BLUP of the additive GE effects is partitioned accordingly:

$$\begin{aligned} \tilde{\mathbf{u}} &= (\mathbf{S}\hat{\Lambda}_s^* \otimes \mathbf{I}_v)\tilde{\mathbf{f}}^* + (\mathbf{\Gamma}\hat{\Lambda}_r^* \otimes \mathbf{I}_v)\tilde{\mathbf{f}}^* + \tilde{\delta} \\ &= \tilde{\beta}_s + \tilde{\beta}_r + \tilde{\delta}, \end{aligned} \quad (22)$$

where $\tilde{\beta}_s$ and $\tilde{\beta}_r$ are vp -vectors based on the known and latent factors, so are therefore referred to as the *known* and *latent* GE effects, respectively. An important feature of the IFA-LMM is that the known and latent GE effects are orthogonal, that is $\tilde{\beta}_s^T\tilde{\beta}_r = 0$.

The empirical BLUP of the *joint* GE effects is then given by:

$$\begin{aligned} \tilde{\beta} &= \left([\mathbf{S}\hat{\Lambda}_s^* + \mathbf{\Gamma}\hat{\Lambda}_r^*] \otimes \mathbf{I}_v \right) \tilde{\mathbf{f}}^* \\ &= \tilde{\beta}_s + \tilde{\beta}_r. \end{aligned} \quad (23)$$

The joint GE effects are based on the joint factors and therefore capture repeatable GEI patterns common to all environments. The known and latent GE effects, $\tilde{\beta}_s$ and $\tilde{\beta}_r$, then capture predictable and observable forms of repeatable GEI in terms of known covariates and latent environmental factors, respectively. By comparison, the residual GE effects, $\tilde{\delta}$, capture non-repeatable GEI specific to individual environments (Smith et al., 2015).

The percentage of variance explained by the known GE effects, $\bar{v}_s$, is presented in Equation 20. This reflects the percentage of variance explained by all covariates, which is the percentage of GEI that is directly interpretable and thence predictable.

The percentage of variance explained by individual covariates is given by:

$$\mathbf{v}_c = 100 \operatorname{diag} \left(\hat{\mathbf{\Lambda}}_s^* \hat{\mathbf{\Lambda}}_s^{*\top} \mathbf{S}^\top \mathbf{S} \right) / \operatorname{tr} \left(\hat{\mathbf{\Lambda}}^* \hat{\mathbf{\Lambda}}^{*\top} + \hat{\mathbf{\Psi}} \right), \quad (24)$$

where $\bar{v}_c = \mathbf{1}_q^\top \mathbf{v}_c$, which is equivalent to $\bar{v}_s$.

Specific examples of this measure are presented in Section 3.

Regression plots for genotypes

Cullis et al. (2014) demonstrated latent regression plots to assess genotype performance and stability in response to the k rotated factors in the conventional factor analytic model. The regression plots developed in this section demonstrate performance and stability in response to the known covariates in the IFA model, as well as the latent environmental factors.

The *data points* on the first $l = 1$ plot are given by:

$$\mathbf{x}_{1i} = \mathbf{S} \hat{\boldsymbol{\lambda}}_{s_1}^* \quad \text{and} \quad \mathbf{y}_{1i} = \tilde{\boldsymbol{\beta}}_{s_i}, \quad (25)$$

where $\hat{\boldsymbol{\lambda}}_{s_1}^*$ is the q -vector of loadings for the first rotated factor and $\tilde{\boldsymbol{\beta}}_{s_i}$ is the p -vector of known GE effects for the i^{th} genotype, such that:

$$\hat{\mathbf{\Lambda}}_s^* = \begin{bmatrix} \hat{\boldsymbol{\lambda}}_{s_1}^* & \dots & \hat{\boldsymbol{\lambda}}_{s_{k_s}}^* \end{bmatrix} \quad \text{and} \quad \tilde{\boldsymbol{\beta}}_s = \operatorname{vec} \begin{bmatrix} \tilde{\boldsymbol{\beta}}_{s_1}^\top \\ \vdots \\ \tilde{\boldsymbol{\beta}}_{s_v}^\top \end{bmatrix}. \quad (26)$$

The *data points* on the remaining $l = 2, 3, \dots, k_s$ plots are then adjusted for the preceding rotated factor(s), such that:

$$\mathbf{x}_{li} = \mathbf{S} \hat{\boldsymbol{\lambda}}_{s_l}^* \quad \text{and} \quad \mathbf{y}_{li} = \tilde{\boldsymbol{\beta}}_{s_i} - \sum_{h=1}^{l-1} \mathbf{S} \hat{\boldsymbol{\lambda}}_{s_h}^* \tilde{f}_{hi}^*, \quad (27)$$

where $\hat{\boldsymbol{\lambda}}_{s_l}^*$ is the q -vector of loadings for the l^{th} rotated factor and $\tilde{f}_{li}^*$ is the corresponding score for the i^{th} genotype.

All $l = 1, 2, \dots, k_s$ plots are supplemented with a *regression line*.

- The *slope of the regression line* is given by the empirical BLUP of the genotype-specific score for that rotated factor, that is $\tilde{f}_{li}^*$.
- The *data points along the regression line* are given by the fitted GE effects for each environment for that rotated factor, that is $\mathbf{S} \hat{\boldsymbol{\lambda}}_{s_l}^* \tilde{f}_{li}^*$.
- The *deviations from the regression line* are given by the additive GE effects for the higher order factors, that is $\sum_{h=l+1}^k \mathbf{S} \hat{\boldsymbol{\lambda}}_{s_h}^* \tilde{f}_{hi}^*$.

Similar plots are constructed for the joint GE effects using $\hat{\mathbf{\Lambda}}^*$ and $\tilde{\boldsymbol{\beta}}_i$, which have similar definition to the known GE effects in Equation 26.

Interpretation of rotated factors

Interpretation of the k rotated factors is achieved using the percentage of variance explained by individual factors and covariates.

1. The percentage of variance explained by each rotated factor is given by:

$$\bar{v}_{s_l} = 100 \operatorname{tr} \left(\mathbf{S} \hat{\boldsymbol{\lambda}}_{s_l}^* \hat{\boldsymbol{\lambda}}_{s_l}^{*\top} \mathbf{S}^\top \right) / \operatorname{tr} \left(\hat{\mathbf{\Lambda}}^* \hat{\mathbf{\Lambda}}^{*\top} + \hat{\mathbf{\Psi}} \right), \quad (28)$$

where $\bar{v}_s = \sum_{l=1}^{k_s} \bar{v}_{s_l}$.

2. The percentage of variance in each rotated factor explained by individual covariates is given by:

$$\mathbf{v}_{cs_l} = 100 \operatorname{diag} \left(\hat{\boldsymbol{\lambda}}_{s_l}^* \hat{\boldsymbol{\lambda}}_{s_l}^{*\top} \mathbf{S}^\top \mathbf{S} \right) / \operatorname{tr} \left(\mathbf{S} \hat{\boldsymbol{\lambda}}_{s_l}^* \hat{\boldsymbol{\lambda}}_{s_l}^{*\top} \mathbf{S}^\top \right) \quad (29)$$

where $\bar{v}_{cs_l} = \mathbf{1}_q^\top \mathbf{v}_{cs_l}$ is the percentage of variance in the l^{th} rotated factor explained by all environmental covariates. Note that $\bar{v}_{cs_l} = 100\%$ since the known factors are fully explained by the covariates. Also note that the numerator in Equation 29 accounts for the variance of each covariate as well as half the covariance with the remaining covariates.

Similar measures are obtained for the joint factors using $\hat{\boldsymbol{\lambda}}_l^* \hat{\boldsymbol{\lambda}}_l^{*\top}$, that is $\bar{v}_l$ and $\mathbf{v}_{c_l}$, where $\bar{v} = \sum_{l=1}^k \bar{v}_l$ and $\bar{v}_{c_l} = \mathbf{1}_p^\top \mathbf{v}_{c_l}$.

Specific examples of these measures are presented in Section 3.

3 Application to example MET dataset

The Bayer Crop Science Cotton Breeding Programme evaluates the commercial merit of test genotypes by annually conducting multi-environment field trials. The cotton breeders use the analysis of MET datasets to select and promote superior genotypes through stages of the programme. There are three late-stages of field evaluation, referred to as preliminary commercial P1, P2 and P3. Seed cotton yield is the primary trait considered in this paper, and is measured in pounds per acre (lbs/a). The 2017 P1 MET dataset for cotton yield is detailed below.

3.1 Data description

Experimental design and phenotypic data

Table 1 presents a summary of the MET dataset. There were $t = 72$ field trials conducted in $p = 24$ environments across eight states in the northern, southern and western cotton growing regions of USA (Figure 1). A total of 208 genotypes were evaluated across all 24 environments. Each environment consisted of three trials. Each trial consisted of a unique set of 68 test genotypes plus four checks. This allocation ensured the MET dataset was balanced between environments, and between trials within environments, that is prior to removing missing plots.

Each trial was designed as a randomised complete block design with two replicate blocks of all genotypes. Yield data was recorded on the majority of plots with 6.54% missing. The number of non-missing plots per

test genotype ranged from 39 to 47 (mean of 45). The number of non-missing genotypes in common between environments ranged from 173 to 208 (mean of 204). This allocation produced excellent connectivity in the MET dataset.

Environmental covariate data

Table 2 presents a summary of the environmental covariates in the MET dataset. There were $q = 18$ covariates collated for all environments, including latitude and longitude coordinates. The remaining 16 covariates consisted of 11 weather and 5 soil observations that were collated using daily meteorological data collated by Bayer Crop Science. Table 2 summarises the observations for each covariate across the northern, southern and western cotton growing regions. Each covariate was then centred and scaled for use in the subsequent analyses.

Marker data

Molecular marker data were available for $v = 204$ (of the 208) genotypes, coded according to a high confidence set of 36,009 single-nucleotide polymorphisms (SNPs). For each marker, genotypes were coded as either 1 (homozygous for the major allele), 0 (heterozygous) or -1 (homozygous minor allele). The frequency of heterozygous marker scores was low given the level of selfing accumulated up to the P1 stage. Monomorphic markers were then removed and missing markers were imputed using the k -nearest neighbour approach of Troyanskaya et al. (2001) with $k = 10$.

The genomic relationship matrix $\mathbf{K}$ in Equation 3 was constructed using the *pedicure* package (Butler, 2021) in *R* (R Core Team, 2021). The default settings in *pedicure* were used as filters, with minor allele frequency $> 0.0002\%$ and missing marker frequency $< 0.9998\%$. A total of $r = 24,265$ markers were retained using this criteria. The diagonal elements in $\mathbf{K}$ ranged from 0.004 to 2.022 (mean of 1.234). The off-diagonals ranged from -0.388 to 1.322 (mean of -0.006).

3.2 Conventional linear mixed models

Table 3 presents a summary of the conventional linear mixed models fitted to the MET dataset. The important results from each model are detailed below.

The analyses commenced by fitting a diagonal variance structure to the additive GE effects in different environments. This process reflects the initial single-site analyses routinely performed on MET datasets, in which the GE effects in different environments are assumed to be uncorrelated. The single-site analyses are typically used to accommodate the experimental design, address spatial variations and identify potential outliers. This highlights three important processes.

1. Replicate blocks effects within trials and trial effects within environments were fitted to accommodate the plot structures of trials and environments (Bailey, 2008).
2. Random column effects were fitted for seven environments to address extraneous variation in the column direction (Gilmour et al., 1997). Random row effects were fitted for one environment only.
3. Few outliers were detected for six environments.

Table 1 presents the environment mean yield and generalised narrow-sense heritability (reliability) following Cullis et al. (2006). Both measures vary substantially between environments and growing regions.

A compound symmetric model was then fitted to the additive GE effects, which reflects many current applications of GS in plant breeding. The distinguishing feature of this model is that the GE effects in different environments are now assumed to be correlated. The form of the compound symmetric model is very restrictive, however, since it comprises a single variance component for genotype main effects and genotype by environment interaction effects. Table 3 shows the inferiority of this model in terms of the formal model selection criteria, AIC and BIC, as well as the informal measure, percentage of variance explained. Note that the latter measure was calculated as the percentage explained by the main effects alone, that is $\bar{v} = 36.2\%$.

Random regression model

The next linear mixed model involved a random regression on the set of 18 environmental covariates presented in Table 2. This model enables interpretation of the covariances between environments in terms of known environmental covariates. Table 3 shows the superiority of this model compared to the diagonal and compound symmetric models. In this case, the percentage of variance explained was calculated as the percentage explained by the genotype intercepts (main effects) as well as the genotype-specific slopes on each covariate, that is $\bar{v} = 58.4\%$. The percentage explained by the covariates alone is 23.3%.

Factor analytic model

The next series of linear mixed models involved fitting a conventional factor analytic model to the additive GE effects. This model provides a parsimonious alternative to the unstructured model for the covariances between environments. There were $k = 4$ factors required to reach an adequate fit and acceptable percentage of variance explained, that is $\bar{v} = 75.2\%$. Half of the 24 environments achieved at least 70% variance explained, however, two environments had $< 50\%$ (omitted for brevity). Table 3 shows the superiority of the FA4 model compared to all remaining conventional linear mixed models, including the higher order models which proved unnecessarily complicated.

3.3 Additive IFA-LMM

The analyses then proceeded by fitting a series of additive IFA-LMMs using the formal and informal model selection criteria. The full sequence of models is presented in Table 4.

There were $k_s = 4$ known and $k_r = 3$ latent factors (i.e. $k = 4$) required to reach an adequate fit and acceptable percentage of variance explained by these factors, that is $\bar{v}_s = 34.6\%$ and $\bar{v}_r = 40.7\%$ ($\bar{v} = 75.3\%$ in total). The former measure reflects the percentage of repeatable GEI common to all environments that is directly interpretable and thence predictable. The IFA4-3 model captures more predictable GEI than the conventional random regression, that is $\bar{v}_s = 34.6\%$ compared to only 23.3%, despite using the same set of environmental covariates. The remaining 24.7% not explained by the known and latent factors reflects any residual (non-repeatable) GEI specific to individual environments.

Table 4 shows the superiority of the IFA4-3 model compared to all remaining IFA-LMMs, as well as the conventional linear mixed models in Table 3. These tables also show the equivalence of the IFA-LMM and the conventional factor analytic linear mixed models, where the model selection criteria are the same for FA1 and IFA1-1, FA2 and IFA2-2, and so on. The distinguishing feature of the IFA-LMM is the ability to jointly model predictable and observable GEI using different numbers of known and latent factors. The total number of genetic parameters estimated in the IFA4-3 model is fewer than in the FA4 model, that is 108 compared to 114, despite very similar model selection criteria.

Summary of environments and covariates

Table 5 presents a summary of the cotton growing environments from the final additive IFA-LMM. The variance of each environment ranged from 7,548.0 to 47,574.2 (497,423.0 overall). These variances are obtained from the denominator in Equation 21. The percentage of variance explained by the known factors, $\mathbf{v}_s$, ranged from 12.7% to 86.3% ($\bar{v}_s = 34.6\%$). The percentage explained by the joint factors, $\mathbf{v}$, ranged from 44.9% to 100.0% ($\bar{v} = 75.3\%$). That is, the variance explained by the joint GE effects is substantially higher than the variance explained by the known GE effects, and thence by the environmental covariates alone.

Table 5 also presents the REML estimates of the environment loadings for each rotated factor, that is $\hat{\lambda}_r^*$. The first rotated factor comprises positive loadings and explains $\bar{v}_1 = 43.9\%$ of the variance. The higher order factors comprise both positive and negative loadings and explain $\bar{v}_{2:4} = 31.4\%$. The fitted GE effects for the first rotated factor therefore capture non-crossover GEI while the higher order factors primarily capture crossover GEI (Smith and Cullis, 2018). This is an important feature of the IFA-LMM, and will be discussed below.

Table 6 presents a similar summary for the environmental covariates. The covariance explained by each covariate ranged from -261,200.8 to 199,616.7 (172,040.9 overall). These covariances are obtained from the numerator in Equation 24. The percentage of variance explained by individual covariates, $\mathbf{v}_c$, ranged from -52.5% to 40.1% ($\bar{v}_s = 34.6\%$). The percentages can be positive or negative, since this measure involves (positive or negative) covariances between environmental covariates. Notable covariates here include maximum dew point temperature, maximum downward solar radiation and maximum temperature, as well as average cloud cover, minimum temperature and minimum humidity.

Table 6 also presents the REML estimates of the rotated factor loadings for each environmental covariate, that is $\hat{\Lambda}_s^*$. The interpretation of these loadings is similar to above, but with one important difference. All rotated factors comprise both positive and negative loadings so the additive GE effects for all rotated factors capture non-crossover GEI. The first rotated factor explains only $\bar{v}_{s1} = 5.4\%$ of the variance and the higher order factors explain $\bar{v}_{s2:4} = 29.2\%$. The practical implication of this will be discussed below.

3.4 Genetic correlations between environments

Figure 2 presents the REML estimates of the genetic correlations between environments in terms of the [a](#) known or [b](#) joint GE effects. Both sets of correlations are ordered based on a dendrogram (applied to [b](#)), where environments closer together are more correlated, and have more similar genotype rankings, than those further apart. The colourkey ranges from 1 (complete agreement in rankings) through zero (dissimilarity in rankings) to -1 (complete reversal and crossover of rankings).

Figure 2 suggests there is structure to the modelled forms of GEI. There are four notable features.

1. The correlations for the joint GE effects are predominately higher than the correlations for the known GE effects, that is due to the environmental covariates alone.
2. The highest correlations generally occur between environments in the same growing region. Environments along the east coast of the USA are also very similar as well as those along the Mississippi river, that is regardless of region (Figure 1).
3. The lowest correlations correspond to 17NC1. This environment is not similar to the remaining environments in the northern growing region, but is similar to those environments along the east coast.
4. The correlations in the northern and southern growing regions are well explained by maximum temperature. The correlations in the western growing region are well explained by maximum dew point temperature, maximum downward solar radiation and maximum temperature.

3.5 Regression plots for genotypes

Figure 3 presents a series of regression plots for check genotypes C1 and C2 in terms of the **a** known or **b** joint GE effects. The first of these plots are used to assess genotype performance and stability in response to the $k_s = 4$ rotated factors, that is in response to the environmental covariates alone. Since the first rotated factor comprises both positive and negative loadings (Table 6), the pair of regression lines intersect and the genotype rankings change across the sample of environments. For example, the fitted GE effects for C2 are higher than C1 for environments with negative loadings, but the converse is true for environments with positive loadings. This will be the case for any pair of genotypes in the MET dataset and indicates that changes in performance are attributed to both crossover and non-crossover GEI. Similar interpretation can be made for the higher order factors.

Figure 3a also demonstrates that the second rotated factor is correlated with longitude ($\rho = 0.80$), where the loadings on the left correspond to environments in the northern and southern growing regions while the loadings on the right correspond to the western growing region. This highlights an important limitation of the conventional factor analytic linear mixed model, where interpretation is often limited to post processing of the rotated loadings. The distinguishing feature of the IFA-LMM is the ability to ascribe direct biological interpretation to the rotated factors in terms of the known environmental covariates. This feature will be discussed below.

The regression plots in Figure 3b also summarise genotype performance and stability, but in response to the $k = 4$ rotated factors, that is in response to the known covariates as well as the latent environmental factors. In this case, the first rotated factor comprises only positive loadings (Table 5), so the pair of regression lines never intersect and the genotype rankings never change (crossover). This is an important difference to the plots in Figure 3a, and indicates that changes in performance are attributed to non-crossover GEI only. Furthermore, since the lines diverge, the difference between the fitted GE effects increases as the loadings increase. This demonstrates the ability of factor analytic models to capture heterogeneity of scale (non-crossover) variance between environments (Smith and Cullis, 2018). The higher order factors comprise both positive and negative loadings, so the additive GE effects primarily capture crossover interactions only.

Figure 3b suggests check C1 is predominately higher performing than C2 across the sample of environments since it has a large predicted score (*slope*) for the first rotated factor. Both checks are notably unstable, however, since they are highly responsive to the higher order factors and therefore have a large spread of additive GE effects (*deviations*) about the first factor regression.

Interpretation of rotated factors

Table 7 presents important supplementary information to the regression plots in Figure 3, and is used to ascribe biological interpretation to the rotated factors. This table presents the percentage of variance in each rotated factor explained by each environmental covariate. In terms of the known GE effects, the percentage of variance explained by individual covariates, $\mathbf{v}_{cs_l}$, ranged from -454.9 to 223.4 ($\bar{v}_{cs_l} = 100\%$; $l = 1, 2, 3$ or 4). The same measure for the joint GE effects, $\mathbf{v}_{c_l}$, ranged from only -162.8 to 197.1 ($\bar{v}_{c_l} = 12.4\%$ and $\bar{v}_{c_l} > 89\%$; $l = 2, 3$ or 4). This highlights three distinct features.

1. The percentages for the joint GE effects are predominately larger than the percentages for the known GE effects, that is due to the environmental covariates alone.
2. The percentages can be positive or negative, since this measure involves (positive or negative) covariances between environmental covariates.
3. Notable covariates include maximum dew point temperature, maximum downward solar radiation and maximum temperature, as well as average cloud cover, minimum temperature and minimum humidity.

Table 7 suggests the environmental covariates primarily model crossover GEI, with a substantial percentage of variance in the higher order joint factors explained by covariates alone, that is $\bar{v}_{c_{2:4}} = 92.7\%$. This feature can be visualised in Figure 3 where the regression plots for the higher order factors are very similar for the known and joint GE effects. By comparison, there is stark dissimilarity between the regression plots for the first rotated factor since the first joint factor is not well explained by covariates, that is $\bar{v}_{c_1} = 12.4\%$.

4 Discussion

This paper introduced an integrated factor analytic linear mixed model (IFA-LMM) for GS in plant breeding, or more generally a single-stage approach for the analysis of MET datasets. The IFA-LMM is a special factor analytic approach which directly integrates environmental covariates. This approach has the potential to become the preferred method of analysis in global plant breeding programmes.

There are two appealing features of using environmental covariates for GS. That is, meaningful biological interpretation can be ascribed to GEI and forward predictions can be obtained for any genotype into any environment. This represents a long-standing objective of plant breeding programmes. Many of the current approaches, however, fail to appropriately accommodate the general structure of the phenotypic and covariate data, which has limited their ability to adequately model genotype (and marker) by environment interaction (Heslot et al., 2014; Jarquín et al., 2014).

The factor analytic approach of [Smith et al. \(2001\)](#) provides an effective model for repeatable GEI in terms of a small number of common environmental factors. This approach, or some variant, has been widely and successfully adopted for the analysis of MET datasets ([Beeck et al., 2010](#); [Ukrainetz et al., 2018](#)). In terms of GS, [Tolhurst et al. \(2019\)](#) demonstrated a single-stage factor analytic linear mixed model which incorporates molecular marker data and appropriately accommodates the general structure of MET datasets. There is one limitation of this approach, however. The common factors are latent so the modelled GEI is only observable, rather than predictable.

The approach developed in this paper models predictable and observable GEI in terms of a joint set of known and latent factors, respectively. This reflects a special factor analytic linear mixed model which directly integrates known environmental covariates. There are three appealing features of the IFA-LMM.

1. The IFA-LMM includes a multiplicative model for GEI in terms of a small number of known and latent factors. This simultaneously reduces the dimension of the covariates as well as the environments.
2. The multiplicative model captures predictable GEI in terms of the covariances between environmental covariates. This introduces known environmental information into the underlying factors so they now have direct and meaningful interpretation.
3. The multiplicative model also captures observable GEI in terms of the latent covariances between environments. This captures any remaining GEI which is strictly orthogonal to the covariates.

These features address the long-standing objective of plant breeding programmes.

The IFA-LMM is embedded within an extension of the computationally efficient reduced-rank formulation of the average information algorithm (Appendix B). [Thompson et al. \(2003\)](#) developed the reduced rank formulation for conventional factor analytic models assuming independent relationships between genotypes, and thence exploited the maximum sparsity of MET datasets. [Kelly et al. \(2007\)](#) extended this formulation to incorporate a (sparse) numerator relationship matrix. The extension to a dense genomic relationship matrix involves complex matrix factorisations and finding more efficient formulations of the IFA-LMM is the topic current research.

The IFA-LMM was demonstrated on a late-stage cotton breeding MET dataset from Bayer Crop Science. The example MET dataset is typical of plant breeding data in terms of the large number of genotypes compared to the number of trials and environments, that is $v \gg t \geq p$. This dataset was used to conveniently illustrate the concepts and methods developed in Section 2. However, it only represents a small subset of trials in the **Bayer Cotton Breeding Programme** and is not intended to capture the full target population of

production environments. A larger MET dataset across multiple years and more locations is common in such applications, since transient and static interactions are substantial components of GEI within the full target population. This will ensure the scope of the covariates, loadings and scores are relevant to the forward prediction of new (untested) genotypes into new (unobserved) environments. Computational challenges are anticipated for these larger MET datasets and finding efficient ways to scale the IFA-LMM is the topic of current research.

There are three important results from Section 3.

1. The selected IFA-LMM involves fewer genetic parameters compared to the selected factor analytic linear mixed model, despite very similar model selection criteria (Tables 3 and 4). This highlights a distinct advantage of integrating known covariates within a special factor analytic framework.
2. More variance is explained by the environmental covariates in the selected IFA-LMM compared to the conventional random regression, that is $\bar{v}_s = 34.6$ compared to 23.3%. This represents 92.7% of the crossover GEI captured by the multiplicative model, and highlights the importance of appropriately modelling the covariates.
3. The latent factors then explain 40.7% of the variance across environments, which represents 87.6% of the non-crossover GEI. This feature can be visualised in Figure 2 where there is stark dissimilarity between the GEI patterns for the known and joint GE effects.

These results demonstrate the importance of categorising environments based on known covariates as well as latent factors.

Interpretation of the IFA-LMM was demonstrated using a series of regression plots (Figure 3). The regression plots are used to assess genotype performance and stability in response to the known and latent factors, that is in response to predictable and observable GEI. Previously, interpretation of genotype performance and stability has been limited to post processing of model terms, for example by correlating environmental covariates with latent factors ([Oliveira et al., 2020](#)) or by examining the response of reference and probe genotypes ([Mathews et al., 2011](#)). The distinguishing feature of the IFA-LMM is the ability to ascribe direct biological interpretation to the rotated factors in terms of the known environmental covariates, such as maximum temperature and minimum humidity.

The integration of environmental covariates within a single-stage IFA-LMM for GS provides breeders with an informative and practical framework to utilise repeatable GEI for selection.

Data Availability

The data that support the findings of this study are available from Bayer Crop Science. Restrictions apply to the availability of these data, which were used under license for this study.

Acknowledgements

The authors thank Bayer Crop Science for funding and use of their data. GG acknowledges funding from the BBSRC Strategic Programme to The Roslin Institute (BBS/E/D/30002275) and the University of Edinburgh's Data-Driven Innovation Chancellor's fellowship.

Author Contributions

DT developed the methodology, curated the data, ran the analyses and wrote the manuscript. CG and BG provided input on plant breeding perspectives. BG also provided the data. JH organised the research project and secured funding. GG reviewed the manuscript. All authors have read and approved the manuscript for submission.

Declarations

Conflict of Interest

The authors declare that they have no conflict of interest.

References

- Bailey RA (2008) Design of Comparative Experiments. Cambridge University Press, Cambridge, URL <http://doi.org/10.1017/CB09780511611483>
- Beeck CP, Cowling WA, Smith AB, Cullis BR (2010) Analysis of yield and oil from a series of canola breeding trials. Part I: fitting factor analytic models with pedigree information. *Genome* 53:992–1001, URL <http://doi.org/10.1139/G10-051>
- Brancourt-Hulmel M, Denis JB, Lecomte C (2000) Determining environmental covariates which explain genotype environment interaction in winter wheat through probe genotypes and biadditive factorial regression. *Crop and Pasture Science* 100:285–298, URL <http://doi.org/10.1007/BF02289233>
- Butler DG (2021) pedicure: pedigree tools. URL <http://mmade.org/pedicure/>, R package version 2.0.1
- Carroll JB (1953) An analytical solution for approximating simple structure in factor analysis. *Psychometrika* 18:23–38, URL <http://doi.org/10.1007/BF02289025>
- Cullis BR, Smith AB, Coombes NE (2006) On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological, and Environmental Statistics* 11:381–393, URL <http://doi.org/10.1198/108571106X154443>
- Cullis BR, Jefferson P, Thompson R, Smith AB (2014) Factor analytic and reduced animal models for the investigation of additive genotype by environment interaction in outcrossing plant species with application to a *pinus radiata* breeding program. *Theoretical and Applied Genetics* 127:2193–2210, URL <http://doi.org/10.1007/s00122-014-2373-0>
- Falconer DS, Mackay T (1996) Introduction to Quantitative Genetics, 4th Edition. Longman, Essex, England
- Finlay KW, Wilkinson GN (1963) The analysis of adaptation in a plant breeding programme. *Australian Journal of Agricultural Research* 14:742–754, URL <http://doi.org/10.1071/AR9630742>
- Gilmour AR, Smith AB, Borg LM, Gogel BJ, Cullis BR (1995) Estimation of factor analytic mixed models for the analysis of multi-treatment multi-environment trial data. *Biometrics* 51(4):1440–1450, URL <http://doi.org/10.2307/2533274>
- Gilmour AR, Cullis BR, Verbyla AP (1997) Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological, and Environmental Statistics* 2:269–293, URL <http://doi.org/10.2307/1400446>
- Heslot N, Akdemir D, Sorrells ME, Jannink JL (2014) Integrating environmental covariates and crop modeling into the genomic selection framework to predict genotype by environment interactions. *Theoretical and Applied Genetics* 127:463–480, URL <http://doi.org/10.1007/s00122-013-2231-5>
- Jarquín D, Crossa J, Lacaze X, Du Cheyron P, Daucourt J, Lorgeou J, Piraux F, Guerreiro L, Pérez P, Calus M, Burgueño J, de los Campos G (2014) A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics* 127:595–607, URL <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3931944{%7B%7Dtool=pmcentrez{%7B%7Drendertype=abstract>
- Kaiser HR (1958) The varimax criterion for analytic rotation in factor analysis. *Psychometrika* 23(3):187–200, URL <http://doi.org/10.1007/BF02289233>
- Kelly AM, Smith AB, Eccleston JA, Cullis BR (2007) The Accuracy of Varietal Selection Using Factor Analytic Models for Multi-Environment Plant Breeding Trials. *Crop Science* 47:1063–1070, URL <http://dx.doi.org/10.2135/cropsci2006.08.0540>
- Knight R (1970) The measurement and interpretation of genotype-environment interactions. *Euphytica* 19:225–235, URL <http://doi.org/10.1007/BF01902950>

- Mathews KL, Trethowan R, Milgate AW, Payne T, van Ginkel M, Crossa J, DeLacy I, Cooper M, Chapman SC (2011) Indirect selection using reference and probe genotype performance in multi-environment trials. *Crop and Pasture Science* 62:313–327, URL <http://doi.org/10.1071/CP10318>
- Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics* 157:1819–1829, URL <http://doi.org/10.1017/S0021859600043392>
- Oakey H, Verbyla AP, Cullis BR, Wei X, Pitchford WS (2007) Joint modelling of additive and non-additive (genetic line) effects in multi-environment trials. *Theoretical and Applied Genetics* 114:1319–1332, URL <http://doi.org/10.1007/s00122-007-0515-3>
- Oakey H, Cullis BR, Thompson R, Comadran J, Halpin C, Waugh R (2016) Genomic Selection in Multi-environment Crop Trials. *G3: Genes|Genomes|Genetics* 6:1313–1326, URL <http://doi.org/10.1534/g3.116.027524>
- Oliveira ICM, Guillen JHS, de Oliveira Ribeiro PC, Gezan SA, Schaffert RE, Simeone MLF, Damasceno CMB, de Souza Carneiro JE, Carneiro PCS, da Costa Parrella RA, Pastina MM (2020) Genotype-by-environment interaction and yield stability analysis of biomass sorghum hybrids using factor analytic models and environmental covariates. *Field Crops Research* 257:107929, URL <http://doi.org/10.1016/j.fcr.2020.107929>
- Piepho HP, Denis JB, Van Eeuwijk F (1998) Predicting cultivar differences using covariates. *Journal of Agricultural, Biological, and Environmental Statistics* 3(2):151–162, URL <http://doi.org/10.2307/1400648>
- R Core Team (2021) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL <http://www.R-project.org/>
- Smith AB (1999) Multiplicative mixed models for the analysis of multi-environment trial data. PhD thesis, University of Adelaide, URL <http://hdl.handle.net/2440/19539>
- Smith AB, Cullis BR (2018) Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data. *Euphytica* 214:143–161, URL <http://doi.org/10.1007/s10681-018-2220-5>
- Smith AB, Cullis BR, Thompson R (2001) Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics* 57:1138–1147, URL <http://doi.org/10.1111/j.0006-341X.2001.01138.x>
- Smith AB, Cullis BR, Thompson R (2005) The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. *Journal of Agricultural Science* 143:449–462, URL <http://doi.org/10.1017/S0021859605005587>
- Smith AB, Ganesalingam A, Kuchel H, Cullis BR (2015) Factor analytic mixed models for the provision of grower information from national crop variety testing programs. *Theoretical and Applied Genetics* 128:55–72, URL <http://doi.org/10.1007/s00122-014-2412-x>
- Smith AB, Borg LM, Gogel BJ, Cullis BR (2019) Estimation of factor analytic mixed models for the analysis of multi-treatment multi-environment trial data. *International Biometric Society* 24(4):573–588, URL <http://doi.org/10.1007/s13253-019-00362-6>
- Stranden I, Garrick A (2009) Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. *Journal of Dairy Science* 92:2971–2975, URL <http://doi.org/10.3168/jds.2008-1929>
- Thompson R, Cullis BR, Smith AB, Gilmour AR (2003) A sparse implementation of the average information algorithm for factor analytic and reduced rank variance models. *Australian and New Zealand Journal of Statistics* 45:445–459, URL <http://doi.org/10.1111/1467-842X.00297>
- Tolhurst DJ, Mathews KL, Smith AB, Cullis BR (2019) Genomic selection in multi-environment plant breeding trials using a factor analytic linear mixed model. *Journal of Animal Breeding and Genetics* 136:279–300, URL <http://doi.org/10.1111/jbg.12404>
- Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB (2001) Missing value estimation methods for DNA microarrays. *Bioinformatics* 17:520–525, URL <http://doi.org/10.1093/bioinformatics/17.6.520>
- Ukrainetz NK, Yanchuk AD, Mansfield S (2018) Climatic drivers of genotype-environment interactions in lodgepole pine based on multi-environment trial data and a factor analytic model of additive covariance. *Canadian Journal of Forest Research* 48:835–854, URL <http://doi.org/10.1139/cjfr-2017-0367>
- VanRaden PM (2008) Efficient Methods to Compute Genomic Predictions. *Journal of Dairy Science* 91:4414–4423, URL <http://doi.org/10.3168/jds.2007-0980>
- Yates F, Cochran WG (1938) The analysis of groups of experiments. *Journal of Agricultural Science, Cambridge* 28:556–580, URL <http://doi.org/10.1017/S0021859600050978>

Table 1 Summary of the phenotypic data in the MET dataset. Presented for each environment is the number of genotypes (with one, two or more replicates) and number of plots (total and missing), as well as the mean yield (lbs/a) and reliability. The data are presented such that each *white* or *grey block* represents a state and the *black lines* distinguish the $\triangle$ northern, $\circ$ southern and $\times$ western growing regions.

State	Env	Trials	Genotypes*				Plots		Yield	
			Total	1rep	2rep	>2rep	Total	NAs	Mean	Rel
$\triangle$ Arkansas	17AR1	3	207	18	185	4	432	20	881.7	0.32
	17AR2	3	205	2	199	4	432	9	1,455.0	0.90
$\triangle$ Missouri	17MO1	3	207	69	134	4	432	76	1,738.4	0.74
$\triangle$ North Carolina	17NC1	3	208	15	189	4	432	16	1,278.1	0.61
$\triangle$ Mississippi	17MS1	3	207	6	197	4	432	10	1,718.9	0.74
$\circ$ Mississippi	17MS2	3	204	9	191	4	432	19	1,079.2	0.70
	17MS3	3	207	140	63	4	432	150	815.9	0.67
$\circ$ Georgia	17GA1	3	208	2	202	4	432	3	1,531.1	0.71
	17GA2	3	208	2	202	4	432	2	1,711.9	0.74
	17GA3	3	208	2	202	4	432	2	1,552.0	0.62
	17GA4	3	208	2	202	4	432	2	1,447.2	0.61
$\circ$ Louisiana	17LA1	3	208	4	200	4	432	6	1,174.7	0.82
	17LA2	3	208	11	193	4	432	12	1,038.0	0.72
$\circ$ South Carolina	17SC1	3	206	0	202	4	432	5	1,457.3	0.71
	17SC2	3	183	52	127	4	432	107	1,726.8	0.52
	17SC3	3	208	5	199	4	432	5	2,070.1	0.62
$\times$ Texas	17TX1	3	208	1	203	4	432	1	1,887.4	0.73
	17TX2	3	208	2	202	4	432	2	1,599.1	0.72
	17TX3	3	207	4	199	4	432	7	1,829.6	0.82
	17TX4	3	208	4	200	4	432	4	1,662.5	0.50
	17TX5	3	198	132	62	4	432	161	1,231.9	0.66
	17TX6	3	206	29	173	4	432	33	1,742.8	0.55
	17TX7	3	208	7	197	4	432	7	1,583.3	0.68
	17TX8	3	208	18	186	4	432	19	2,289.9	0.52
Total	24	72	208	-	-	-	10,368	678	-	-

* Total number after missing plots removed.

Table 2 Summary of the environmental covariates in the MET dataset. Presented for each covariate is the minimum, mean and maximum for the $\triangle$ northern, $\circ$ southern and $\times$ western growing regions, that is prior to centring and scaling.

Covariate	Description	$\triangle$ North			$\circ$ South			$\times$ West		
		Min	Mean	Max	Min	Mean	Max	Min	Mean	Max
LAT	latitude	33.6	34.9	36.4	31.0	32.7	34.5	31.4	33.2	34.9
LONG	longitude	-91.9	-88.3	-78.0	-100.8	-86.6	-80.0	-102.3	-101.1	-99.5
avgCCR	average cloud cover	0.47	0.50	0.58	0.33	0.52	0.59	0.32	0.35	0.37
maxDPT	max dew point temperature	18.9	20.0	20.7	16.4	20.9	22.1	13.5	15.7	17.6
maxDSR	max downward solar radiation	2.68	2.75	2.78	2.72	2.77	3.06	2.97	3.05	3.11
minHUM	min humidity	52.0	52.8	53.7	34.1	48.7	55.9	30.1	34.0	40.4
maxNSR	max net solar radiation	2.22	2.27	2.31	2.26	2.33	2.52	2.44	2.46	2.51
maxPRP	max precipitation	0.029	0.037	0.053	0.028	0.047	0.060	0.018	0.024	0.030
totPRP	total precipitation	3.01	3.32	3.45	1.74	3.54	4.94	1.34	1.61	2.11
maxTMP	max temperature	27.6	28.4	29.0	29.1	30.1	31.5	28.7	30.3	32.1
minTMP	min temperature	17.9	19.0	20.0	18.3	20.0	21.0	15.4	17.4	19.5
minWSP	min wind speed	4.74	4.97	5.31	4.61	5.35	8.71	7.39	8.13	9.41
avgWDR	average wind direction	159	168	181	151	168	181	145	153	163
maxST1	max soil temperature 1	27.0	27.9	28.9	27.7	29.7	32.5	29.5	32.2	34.5
minST1	min soil temperature 1	19.3	20.1	21.1	20.6	21.7	23.2	19.0	20.6	22.8
minST4	min soil temperature 4	20.0	21.1	22.4	21.6	22.7	24.2	21.0	22.9	25.2
avgSM3	soil moisture 3	0.28	0.33	0.42	0.07	0.24	0.31	0.11	0.19	0.26
avgSM4	soil moisture 4	0.33	0.36	0.45	0.10	0.28	0.38	0.08	0.16	0.22

Table 3 Summary of the conventional linear mixed models fitted to the MET dataset. Presented for each model is the number of estimated genetic variance parameters, residual log-likelihood, AIC, BIC and percentage of variance explained ($\bar{v}$).

Model	Params	Loglik	AIC	BIC	$\bar{v}$
<i>diag</i>	24	-54,766.1	109,836.1	110,925.4	-
<i>comp</i>	2	-54,511.2	109,282.4	110,214.0	36.2
<i>rreg</i>	43	-54,270.6	108,883.2	110,108.6	58.4
FA1	48	-54,348.3	109,048.6	110,309.9	43.2
FA2	71	-54,188.0	108,773.9	110,200.1	60.4
FA3	93	-54,075.2	108,592.3	110,176.1	70.7
FA4	114	-54,037.1	108,558.2	110,292.5	75.2
FA5	134	-54,005.3	108,534.6	110,412.2	79.0

Note: 128 non-genetic and residual variance parameters estimated in all models. The selected factor analytic linear mixed model is distinguished with *bold font*.

Table 4 Summary of the additive IFA-LMMs fitted to the the MET dataset. Presented for each model is the number of estimated known and latent variance parameters, residual log-likelihood, AIC, BIC and the percentage of variance explained by the known ($\bar{v}_s$) or joint ($\bar{v}$) factors.

Model	Params	Loglik	AIC	BIC	$\bar{v}_s$	$\bar{v}$
IFA1-1	18-30	-54,348.3	109,048.6	110,309.9	7.0	43.2
IFA2-1	35-30	54,196.3	108,778.6	110,161.7	20.1	60.0
IFA2-2	35-36	-54,188.0	108,773.9	110,200.1	20.1	60.4
IFA3-2	51-36	-54,086.6	108,603.3	110,144.0	29.9	70.3
IFA3-3	51-42	-54,075.2	108,592.3	110,176.1	30.1	70.7
IFA4-3	66-42	-54,043.5	108,559.0	110,250.3	34.6	75.3
IFA4-4	66-48	-54,037.1	108,558.2	110,292.5	34.3	75.2
IFA5-4	80-48	-54,009.0	108,530.0	110,364.6	37.2	79.3
IFA5-5	80-54	-54,005.3	108,534.6	110,412.2	36.9	79.0

Note: 128 non-genetic and residual variance parameters estimated in all models. The selected IFA-LMM is distinguished with *bold font*.

Table 5 The selected additive IFA-LMM, Part 1: Summary of growing environments. Presented for each environment is the mean yield (lbs/a), REML estimate of the variance and percentage of variance explained by the known ($\mathbf{v}_s$) or joint ($\mathbf{v}$) factors, as well as the estimates of the rotated factor loadings ($\hat{\lambda}_l^*$; $l = 1, 2, 3$ or 4).

State	Env	Mean	Var	$\mathbf{v}_s$	$\mathbf{v}$	$\hat{\lambda}_1^*$	$\hat{\lambda}_2^*$	$\hat{\lambda}_3^*$	$\hat{\lambda}_4^*$
	17AR1	881.7	10,035.7	49.0	100.0	65.6	-0.9	74.8	-11.8
△ Arkansas	17AR2	1,455.0	47,574.2	32.2	89.4	180.3	-44.9	78.9	-42.5
△ Missouri	17MO1	1,738.4	47,542.2	26.7	82.5	180.8	-47.9	34.7	-55.2
△ North Carolina	17NC1	1,278.1	9,756.5	86.3	70.0	28.5	-11.6	-76.3	-7.7
△ Mississippi	17MS1	1,718.9	20,696.1	47.8	77.9	112.4	-35.3	35.1	-32.0
	17MS2	1,079.2	22,864.9	46.5	82.2	106.3	0.7	60.8	61.5
○ Mississippi	17MS3	815.9	22,260.3	37.3	100.0	123.0	-24.6	53.2	60.7
	17GA1	1,531.1	21,839.4	23.3	64.2	94.2	-24.0	-67.4	3.3
	17GA2	1,711.9	25,116.4	19.2	54.3	86.6	-27.3	-73.4	-2.0
	17GA3	1,552.0	16,175.1	30.0	82.9	99.5	-35.1	-46.1	12.3
○ Georgia	17GA4	1,447.2	18,592.9	27.1	68.0	85.0	-27.7	-65.3	-19.5
	17LA1	1,174.7	21,422.2	19.9	72.0	105.3	-48.2	1.7	44.8
○ Louisiana	17LA2	1,038.0	13,438.3	22.8	77.1	93.6	-29.4	-24.8	10.4
	17SC1	1,457.3	16,216.8	12.7	56.9	83.1	-16.2	-39.8	21.8
	17SC2	1,726.8	11,273.0	41.0	49.0	34.9	-7.5	-63.8	13.1
○ South Carolina	17SC3	2,070.1	18,788.0	23.9	90.9	108.2	-39.5	-61.6	4.3
	17TX1	1,887.4	19,013.2	61.8	92.2	69.3	111.8	-8.6	-12.5
	17TX2	1,599.1	16,340.6	36.9	62.3	54.1	78.5	-22.3	-24.3
	17TX3	1,829.6	36,710.4	41.6	74.2	99.6	130.3	15.4	8.8
	17TX4	1,662.5	10,757.3	32.9	65.2	48.7	61.9	-16.3	23.2
	17TX5	1,231.9	30,353.6	30.1	62.2	93.3	97.7	-1.9	25.0
	17TX6	1,742.8	7,548.0	81.8	44.9	27.2	48.4	-10.7	-14.0
	17TX7	1,583.3	17,732.9	44.6	67.0	55.2	94.0	2.3	-2.8
× Texas	17TX8	2,289.9	15,374.9	24.4	72.7	58.8	79.2	-27.4	-26.6
Overall	-	-	497,423.0	34.6%	75.3%	43.9%	16.3%	11.0%	4.0%

Note: The overall percentage of variance explained by the known ($\bar{v}_s$) or joint ($\bar{v}$) factors, and by each rotated factor ($\bar{v}_l$) is presented in the final row.

Table 6 The selected additive IFA-LMM, Part 2: Summary of environmental covariates. Presented is the REML estimate of the covariance explained by each covariate and percentage of variance explained by each covariate ($\mathbf{v}_c$), as well as the estimates of the rotated factor loadings ($\hat{\lambda}_{s_l}^*$; $l = 1, 2, 3$ or 4).

Cov	Covar	$\mathbf{v}_c$	$\hat{\lambda}_{s_1}^*$	$\hat{\lambda}_{s_2}^*$	$\hat{\lambda}_{s_3}^*$	$\hat{\lambda}_{s_4}^*$
LAT	12,704.6	2.6	7.8	29.4	48.0	27.7
LONG	40,187.0	8.1	-84.6	10.1	-130.5	-46.0
avgCCR	-141,370.1	-28.4	-173.1	87.1	3.9	-39.9
maxDPT	199,270.9	40.1	221.1	-130.2	159.5	29.4
maxDSR	199,616.7	40.1	-138.9	117.5	24.1	-22.8
minHUM	-261,200.8	-52.5	-289.1	66.7	-240.2	-154.4
maxNSR	33,681.1	6.8	23.4	31.6	43.5	41.5
maxPRP	-4,991.6	-1.0	16.4	13.9	42.8	77.7
totPRP	24,002.6	4.8	50.6	-1.7	-11.9	21.1
maxTMP	139,194.4	28.0	-143.8	24.1	-136.7	-44.4
minTMP	-145,014.3	-29.2	-21.7	124.2	156.7	141.3
minWSP	4,957.8	1.0	-61.0	-25.7	-72.4	-23.0
avgWDR	-26,248.1	-5.3	16.2	38.7	3.0	46.6
maxST1	-28,924.7	-5.8	40.2	15.9	63.1	34.8
minST1	30,209.0	6.1	171.7	-135.4	-35.6	-135.3
minST4	73,000.5	14.7	-138.5	91.5	-24.4	67.9
avgSM3	-14,914.7	-3.0	46.0	33.0	-22.2	-26.8
avgSM4	37,880.5	7.6	-46.3	-42.7	58.9	57.6
Overall	172,040.9	34.6%	5.4%	15.4%	9.9%	3.9%

Note: The overall percentage of variance explained by the known factors ($\bar{v}_s$), and by each rotated factor ($\bar{v}_{s_l}$) is presented in the final row.

Table 7 The selected additive IFA-LMM, Part 3: Interpretation of rotated factors. Presented is the percentage of variance in each rotated factor explained by each environmental covariate in terms of the known ($\mathbf{v}_{\mathbf{cs}_l}$) or joint ($\mathbf{v}_{\mathbf{c}_l}$) GE effects ($l = 1, 2, 3$ or 4).

Cov	Known GE effects				Joint GE effects			
	$\mathbf{v}_{\mathbf{cs}_1}$	$\mathbf{v}_{\mathbf{cs}_2}$	$\mathbf{v}_{\mathbf{cs}_3}$	$\mathbf{v}_{\mathbf{cs}_4}$	$\mathbf{v}_{\mathbf{c}_1}$	$\mathbf{v}_{\mathbf{c}_2}$	$\mathbf{v}_{\mathbf{c}_3}$	$\mathbf{v}_{\mathbf{c}_4}$
LAT	5.6	3.3	31.4	-35.0	0.7	3.1	28.0	-34.2
LONG	-34.4	-14.0	128.6	-14.9	-4.3	-13.2	114.8	-14.6
avgCCR	-130.3	-132.5	-3.2	-15.7	-16.1	-125.1	-2.8	-15.4
maxDPT	220.1	208.8	-51.5	25.6	27.3	197.1	-45.9	25.1
maxDSR	171.1	193.1	6.7	10.8	21.2	182.3	6.0	10.6
minHUM	-454.9	-100.4	-91.1	-85.5	-56.3	-94.7	-81.3	-83.7
maxNSR	-26.7	47.0	7.4	6.7	-3.3	44.3	6.6	6.6
maxPRP	6.4	-18.2	-30.0	113.4	0.8	-17.2	-26.7	110.9
totPRP	64.1	2.6	2.0	18.9	7.9	2.5	1.8	18.4
maxTMP	223.4	7.9	159.8	-29.3	27.7	7.5	142.6	-28.7
minTMP	-10.8	-172.5	-78.5	147.8	-1.3	-162.8	-70.1	144.6
minWSP	104.6	-39.6	11.7	6.6	12.9	-37.4	10.5	6.5
avgWDR	9.6	-37.3	-2.6	5.4	1.2	-35.2	-2.4	5.3
maxST1	-75.2	17.7	-42.5	-6.8	-9.3	16.7	-37.9	-6.6
minST1	-102.7	82.7	36.9	-121.1	-12.7	78.0	32.9	-118.5
minST4	140.7	23.2	11.6	59.0	17.4	21.9	10.4	57.7
avgSM3	42.4	-24.8	-13.0	-5.3	5.3	-23.4	-11.6	-5.2
avgSM4	-53.1	52.9	16.2	19.6	-6.6	49.9	14.4	19.1
Total	100%	100%	100%	100%	12.4%	94.4%	89.2%	97.8%

Note: The percentage of variance in each factor explained by all covariates ($\bar{v}_{\mathbf{cs}_l}$ or $\bar{v}_{\mathbf{c}_l}$) is presented in the final row.

Fig. 1 Map of the cotton growing environments in the MET dataset. States are distinguished by *colour* and the growing regions are distinguished by *shape*.

Fig. 2 Heatmap of the REML estimates of correlations between growing environments in terms of the (a) known or (b) joint GE effects. Both sets of correlations are ordered based on the dendrogram applied to (b), with *black lines* used to distinguish the $\triangle$ northern, $\circ$ southern and $\times$ western growing regions.

Fig. 3 Regression plots for check genotypes C1 and C2 in terms of the (a) known or (b) joint GE effects. The additive GE effects for each growing region are distinguished by *shape*, and are adjusted for the preceding factor(s) where appropriate.

Appendix A Inclusion of environments without covariate data

The IFA-LMM can be extended to scenarios where there are p environments with phenotypic data, but there are only $p_d < p$ environments with covariate data. That is, there are p_d environments with covariate data and $p - p_d$ environments without covariate data.

The matrix of environmental covariates and projection matrix are respecified as:

$$\mathbf{S} = \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{0} \end{bmatrix} \quad \text{and} \quad \mathbf{\Gamma} = \begin{bmatrix} \mathbf{\Gamma}_1 & \mathbf{0} \\ \mathbf{0}^\top & \mathbf{I}_{p-p_d} \end{bmatrix}, \quad (30)$$

with $\mathbf{S}^\top \mathbf{\Gamma} = \mathbf{0}$, where $\mathbf{S}$ is a $p \times q$ matrix and $\mathbf{S}_1$ is a $p_d \times q$ matrix of centred and scaled covariates, $\mathbf{\Gamma}$ is a $p \times (p-q)$ matrix and $\mathbf{\Gamma}_1$ is a $p_d \times (p_d-q)$ matrix which is taken as the first $(p_d - q)$ columns in:

$$[\mathbf{I}_{p_d} - \mathbf{S}_1^\top (\mathbf{S}_1^\top \mathbf{S}_1)^{-1} \mathbf{S}_1^\top], \quad (31)$$

where each column forms a basis for the orthogonal complement to the range in $\mathbf{S}_1$.

The IFA-LMM for $\mathbf{y}$, the n -vector of phenotypic data on v genotypes and p environments, can be written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad (32)$$

where $\boldsymbol{\varepsilon} = \mathbf{Z}\boldsymbol{\delta} + \mathbf{Z}_p \mathbf{u}_p + \mathbf{e}$.

The form of the joint GE effects is obtained by substituting the components in Equation 30 into Equation 32, which gives:

$$\begin{aligned} \boldsymbol{\beta} &= \left(\begin{bmatrix} \mathbf{S}_1 \\ \mathbf{0} \end{bmatrix} \mathbf{\Lambda}_s^* \otimes \mathbf{I}_v \right) \mathbf{f}^* + \left(\begin{bmatrix} \mathbf{\Gamma}_1 & \mathbf{0} \\ \mathbf{0}^\top & \mathbf{I}_{p-p_d} \end{bmatrix} \mathbf{\Lambda}_r^* \otimes \mathbf{I}_v \right) \mathbf{f}^* \\ &= \boldsymbol{\beta}_s + \boldsymbol{\beta}_r, \end{aligned} \quad (33)$$

where $\boldsymbol{\beta}_s^\top \boldsymbol{\beta}_r = 0$. The vp -vector of latent GE effects can be written as $\boldsymbol{\beta}_r = (\boldsymbol{\beta}_{r_1}^\top, \boldsymbol{\beta}_{r_2}^\top)^\top$, where $\boldsymbol{\beta}_{r_1}$ is a vp_d -vector and $\boldsymbol{\beta}_{r_2}$ is a $v(p - p_d)$ -vector which correspond to environments with or without covariate data, respectively. The joint GE effects are partitioned accordingly:

$$\begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{\beta}_{s_1} \\ \mathbf{0} \end{pmatrix} + \begin{pmatrix} \boldsymbol{\beta}_{r_1} \\ \boldsymbol{\beta}_{r_2} \end{pmatrix}, \quad (34)$$

where $\boldsymbol{\beta}_{r_1} = (\mathbf{\Gamma}_1 \mathbf{\Lambda}_{r_1}^* \otimes \mathbf{I}_v) \mathbf{f}^*$ and $\boldsymbol{\beta}_{r_2} = (\mathbf{\Lambda}_{r_2}^* \otimes \mathbf{I}_v) \mathbf{f}^*$, with $\mathbf{\Lambda}_r^* = \begin{bmatrix} \mathbf{\Lambda}_{r_1}^* \\ \mathbf{\Lambda}_{r_2}^* \end{bmatrix}$ partitioned conformably with $\boldsymbol{\beta}$.

The IFA-LMM for $\mathbf{y}$ is then obtained by substituting Equation 34 into Equation 32, which gives:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_1(\boldsymbol{\beta}_{s_1} + \boldsymbol{\beta}_{r_1}) + \mathbf{Z}_2\boldsymbol{\beta}_{r_2} + \boldsymbol{\varepsilon}, \quad (35)$$

where $\mathbf{Z} = [\mathbf{Z}_1 \mathbf{Z}_2]$ is also partitioned conformably with $\boldsymbol{\beta}$. The extension to scenarios where environments have some, but not all, covariates available is straight forward.

The form of the IFA-LMM in Equation 35 is used to illustrate the inclusion of environments with phenotypic data but without covariate data, in a manner that is consistent with the concepts presented in this

paper. Note, however, it is more computationally efficient to use the form of the IFA-LMM commensurate with Equation 17, since this utilises the reduced rank form of the IFA-LMM presented in Appendix B.

Appendix B Reduced rank formulation of the IFA-LMM

B.1 Preliminaries

The computationally efficient (fully) reduced-rank form of the IFA-LMM for $\mathbf{y}$, the n -vector of phenotypic data on v genotypes and p environments, is given by:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z}_\Lambda \mathbf{f} + \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\varepsilon}, \quad (36)$$

where $\mathbf{Z}_\Lambda = \mathbf{Z} \left(\mathbf{B} \begin{bmatrix} \mathbf{\Lambda}_s \\ \mathbf{\Lambda}_r \end{bmatrix} \otimes \mathbf{I}_v \right)$ is the $n \times vk$ integrated factor analytic design matrix, which is updated during REML estimation, $\mathbf{B} = [\mathbf{S} \mathbf{\Gamma}]$ is the $p \times p$ matrix of basis functions and $\boldsymbol{\varepsilon} = \mathbf{Z}_p \mathbf{u}_p + \mathbf{e}$ with $\text{var}(\boldsymbol{\varepsilon}) = \mathbf{R}_\varepsilon$.

The mixed model equations for the reduced rank form of the IFA-LMM are then given by:

$$\begin{bmatrix} \mathbf{X}^\top \mathbf{R}_\varepsilon^{-1} \mathbf{X} & \mathbf{X}^\top \mathbf{R}_\varepsilon^{-1} \mathbf{Z}_\Lambda & \mathbf{X}^\top \mathbf{R}_\varepsilon^{-1} \mathbf{Z} \\ \mathbf{Z}_\Lambda^\top \mathbf{R}_\varepsilon^{-1} \mathbf{X} & \mathbf{Z}_\Lambda^\top \mathbf{R}_\varepsilon^{-1} \mathbf{Z}_\Lambda + \mathbf{I}_k \otimes \mathbf{K}^{-1} & \mathbf{Z}_\Lambda^\top \mathbf{R}_\varepsilon^{-1} \mathbf{Z} \\ \mathbf{Z}^\top \mathbf{R}_\varepsilon^{-1} \mathbf{X} & \mathbf{Z}^\top \mathbf{R}_\varepsilon^{-1} \mathbf{Z}_\Lambda & \mathbf{Z}^\top \mathbf{R}_\varepsilon^{-1} \mathbf{Z} + \mathbf{\Psi}^{-1} \otimes \mathbf{K}^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\tau}} \\ \hat{\mathbf{f}} \\ \hat{\boldsymbol{\delta}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^\top \mathbf{y} \\ \mathbf{Z}_\Lambda^\top \mathbf{y} \\ \mathbf{Z}^\top \mathbf{y} \end{bmatrix}$$

Alternatively, Equations 36 and 37 can be written as (Smith et al., 2019):

$$\mathbf{y} = \mathbf{W}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \quad \text{and} \quad \mathbf{C}\tilde{\boldsymbol{\beta}} = \mathbf{W}^\top \mathbf{R}_\varepsilon^{-1} \mathbf{y}, \quad (38)$$

where $\mathbf{W} = [\mathbf{X} \mathbf{Z}_\Lambda \mathbf{Z}]$ and $\mathbf{C} = \mathbf{W}^\top \mathbf{R}_\varepsilon^{-1} \mathbf{W} + \mathbf{G}^{-1}$, with:

$$\tilde{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\tau}} \\ \hat{\mathbf{f}} \\ \hat{\boldsymbol{\delta}} \end{pmatrix} \quad \text{and} \quad \mathbf{G}^{-1} = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_k \otimes \mathbf{K}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{\Psi}^{-1} \otimes \mathbf{K}^{-1} \end{bmatrix} \quad (39)$$

The empirical BLUPs of the key random effects, $\hat{\mathbf{f}}$ and $\hat{\boldsymbol{\delta}}$, are obtained via absorption of the coefficient matrix, $\mathbf{C}$, onto $\mathbf{y}^\top \mathbf{R}_\varepsilon^{-1} \mathbf{y}$ (Smith, 1999). Absorption of $\mathbf{C}$ also produces the prediction error variances of $\hat{\mathbf{f}}$ and $\hat{\boldsymbol{\delta}}$.

B.2 Estimation of factor loadings and specific variances

The reduced rank form of the IFA-LMM is embedded within an extension of the sparse formulation of the average information (AI) algorithm (Thompson et al., 2003). The AI algorithm is an iterative scheme which computes updates of variance parameters as:

$$\boldsymbol{\kappa}^{(m+1)} = \boldsymbol{\kappa}^{(m)} + \left[\mathcal{I}_a^{(m)} \right]^{-1} \mathbf{s}(\boldsymbol{\kappa}^{(m)}), \quad (40)$$

where $\boldsymbol{\kappa}^{(m)}$ is a b -vector of variance parameters, $\mathcal{I}_a^{(m)}$ is the average information matrix and $\mathbf{s}(\boldsymbol{\kappa}^{(m)})$ is the score function for the m^{th} iteration.

The score function for κ_i is given by (Gilmour et al., 1995):

$$s(\kappa_i) = -\frac{1}{2} \left[\text{tr}(\mathbf{P}\dot{\mathbf{H}}_i) - \mathbf{y}^\top \mathbf{P}\mathbf{q}_i \right], \quad (41)$$

where $\mathbf{P} = \mathbf{R}_\epsilon^{-1} - \mathbf{R}_\epsilon^{-1}\mathbf{W}\mathbf{C}^{-1}\mathbf{W}^\top\mathbf{R}_\epsilon^{-1}$ is the $n \times n$ residual sum of squares matrix, $\dot{\mathbf{H}}_i = \frac{\partial \mathbf{H}}{\partial \kappa_i}$ with $\text{var}(\mathbf{y}) = \mathbf{H}$ and $\mathbf{q}_i$ is the working variate for the i^{th} variance parameter. The components of the inverse coefficient matrix, $\mathbf{C}^{-1}$, are obtained during the absorption of $\mathbf{C}$ (Smith, 1999).

The working variate for κ_i is given by (Gilmour et al., 1995):

$$\mathbf{q}_i = \dot{\mathbf{H}}_i \mathbf{P} \mathbf{y}. \quad (42)$$

Let the vectors of key variance parameters in the IFALMM be denoted by:

$$\boldsymbol{\lambda}_{\text{sr}} = \text{vect}(\boldsymbol{\Lambda}_{\text{sr}}) \quad \text{and} \quad \boldsymbol{\psi} = \text{diag}(\boldsymbol{\Psi}), \quad (43)$$

$$\text{where } \boldsymbol{\Lambda}_{\text{sr}} = \begin{bmatrix} \boldsymbol{\Lambda}_{\text{s}} \\ \boldsymbol{\Lambda}_{\text{r}} \end{bmatrix}.$$

The working variates for the key variance parameters are given by:

$$\mathbf{q}_i = \begin{cases} \mathbf{Z} \left[\mathbf{B}(\boldsymbol{\Lambda}_{\text{sr}} \dot{\boldsymbol{\Lambda}}_{\text{sr}i}^\top + \dot{\boldsymbol{\Lambda}}_{\text{sr}i} \boldsymbol{\Lambda}_{\text{sr}}^\top) \mathbf{B}^\top \otimes \mathbf{K} \right] \mathbf{Z}^\top \mathbf{P} \mathbf{y} & \kappa_i \in \boldsymbol{\lambda}_{\text{sr}} \\ \mathbf{Z} \left[\dot{\boldsymbol{\Psi}}_i \otimes \mathbf{K} \right] \mathbf{Z}^\top \mathbf{P} \mathbf{y} & \kappa_i \in \boldsymbol{\psi} \end{cases} \quad (44)$$

where $\dot{\boldsymbol{\Lambda}}_{\text{sr}i} = \frac{\partial \boldsymbol{\Lambda}_{\text{sr}}}{\partial \lambda_{\text{sr}i}}$ and $\dot{\boldsymbol{\Psi}}_i = \frac{\partial \boldsymbol{\Psi}}{\partial \psi_i}$.

The working variates for the specific variances can be simplified by utilising the assumption that the residual GE effects, $\boldsymbol{\delta}$, are specific to individual environments. This working variate becomes:

$$\mathbf{q}_i = \begin{cases} \mathbf{Z}_i \mathbf{K} \mathbf{Z}_i^\top \mathbf{P}_i \mathbf{y} & \kappa_i \in \boldsymbol{\psi} \end{cases} \quad (45)$$

where $\mathbf{Z}_i$ is the $n_i \times v$ design matrix for the i^{th} environment, $\mathbf{P}_i$ are the corresponding n_i rows in the residual sum of squares matrix and $\mathbf{Z}_i$ are the corresponding v columns in the full design matrix.

The trace term in Equation 41 is then given by:

$$\text{tr}(\mathbf{P} \dot{\mathbf{H}}_i) = \begin{cases} 2 \text{tr} \left[(\dot{\boldsymbol{\Lambda}}_{\text{sr}i}^\top \mathbf{B}^\top \otimes \mathbf{I}_v) \mathbf{Z}^\top \mathbf{R}_\epsilon^{-1} \mathbf{W} \mathbf{C} \cdot \tilde{\mathbf{f}} \right] & \kappa_i \in \boldsymbol{\lambda}_{\text{sr}} \\ \text{tr} \left[(\dot{\boldsymbol{\Psi}}_i \boldsymbol{\Psi}^{-1} \otimes \mathbf{I}_v) \mathbf{Z}^\top \mathbf{R}_\epsilon^{-1} \mathbf{W} \mathbf{C} \cdot \tilde{\boldsymbol{\delta}} \right] & \kappa_i \in \boldsymbol{\psi} \end{cases} \quad (46)$$

where $\mathbf{C} \cdot \tilde{\mathbf{f}}$ are the vk columns in the inverse coefficient matrix corresponding to $\tilde{\mathbf{f}}$ and $\mathbf{C} \cdot \tilde{\boldsymbol{\delta}}$ are the vp columns corresponding to $\tilde{\boldsymbol{\delta}}$.

The trace terms in Equation 46 avoid working with the (dense) genomic relationship matrix. When $\mathbf{K}$ is not prohibitively large, the trace term for the specific variances can be computed as:

$$\text{tr}(\mathbf{P} \dot{\mathbf{H}}_i) = \begin{cases} \frac{v}{\psi_i} - \text{tr}[\mathbf{K} \mathbf{C}^{\boldsymbol{\eta}_i \boldsymbol{\eta}_i}] & \kappa_i \in \boldsymbol{\psi} \end{cases} \quad (47)$$

where $\mathbf{C}^{\boldsymbol{\eta}_i \boldsymbol{\eta}_i}$ is the prediction error variance of $\boldsymbol{\eta}_i = \mathbf{Z}_i^\top \mathbf{P}_i \mathbf{y}$.

Lastly, the AI matrix in Equation 40 is given by:

$$\mathcal{I}_{\mathbf{a}} = \frac{1}{2} \mathbf{Q}^\top \mathbf{P} \mathbf{Q}, \quad (48)$$

where $\mathbf{Q} = [\mathbf{q}_1 \ \mathbf{q}_2 \ \dots \ \mathbf{q}_b]$ is the $n \times b$ matrix of working variates across all b variance parameters. The AI matrix is obtained via absorption of $\mathbf{C}$ onto $\mathbf{Q}^\top \mathbf{R}_\epsilon^{-1} \mathbf{Q}$ (Smith, 1999).