NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong, Australia

Working Paper

06-18

Plant Breeding Selection Tools Built on Factor Analytic Mixed Models for Multi-Environment Trial Data.

Alison B. Smith and Brian R. Cullis

National Institute for Applied Statistics Research Australia, University of Wollongong,
Wollongong NSW 2522, Australia Phone +61 2 4221 5076, Fax +61 2 4221 4998.
Email: karink@uow.edu.au

Plant breeding selection tools built on factor analytic mixed models for multi-environment trial data.

Alison B. Smith · Brian R. Cullis

Received: date / Accepted: date

Abstract An advanced and widely used method of analysis for multi-environment trial data involves a linear mixed model with factor analytic (FA) variance structures for the variety by environment effects. This model can accommodate unbalanced data, that is, not all varieties in all environments, it allows the use of pedigree information and appropriate modelling of individual trial plot structures, and most importantly the FA structure for the variety by environment effects is parsimonious and regularly results in a good fit to the data. The model provides accurate predictions of the variety effects for every environment in the data-set but this constitutes a large and unwieldly amount of information to process for the purpose of variety selection. We address this issue in the current paper by proposing factor analytic selection tools (FAST) to summarise the predictions in a concise yet informative manner. The tools, which are natural derivatives of the FA structure, result in measures of overall performance and stability across the environments in the data-set. All measures are expressed on the same scale as the trait under consideration and can easily be combined to form an index for selection.

Keywords factor analytic model · linear mixed model · multi-environment trial · selection · variety by environment interaction

1 Introduction

Plant breeders use information from the analysis of multi-environment trial (MET) data to select superior varieties. A commonly occurring impediment to selection is the presence of variety by environment interaction (VEI), which is characterised

A. B. Smith
Centre for Bioinformatics and Biometrics, National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong, Australia E-mail: alismith@uow.edu.au

B. R. Cullis
Centre for Bioinformatics and Biometrics, National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong, Australia

by the differential response of varieties to a change in environment. A statistical analysis of MET data must therefore aim to accurately encapsulate VEI. To highlight some of the difficulties in analysing MET data, and the short-comings of current approaches we consider an example from a tree breeding programme.

The Radiata Pine Breeding Company (RPBC) conducts genetic trials in which progeny trees obtained from selected parental trees are grown in test plantations. The MET data-set under study in this paper comprises a series of 92 genetic trials planted in different years and in various geographic locations in New Zealand and Australia. This is an extension of the data-set used in Cullis et al (2014) and includes 15 more recently planted trials. The trait of interest is tree stem diameter (cm) measured at breast height (DBH). Each trial comprised a number of plots (areas of land measuring approximately 3m x 3m) to which progeny trees were allocated. In the majority of trials there was no true replication of progeny trees so that each plot was allocated a genetically distinct tree. In eight trials the progeny trees were cloned so that there were multiple plots (true replicates) of each tree. The experimental designs varied between trials and the possible types of plot structures are described in Cullis et al (2014). The full MET data-set comprised 348806 plots (and thence data records) corresponding to 336528 trees. Pedigree information was available on 339589 trees which comprised the 336528 progeny trees that were grown in the trials and 3061 parental trees. Note that none of the parental trees were grown in the trials.

The aim of the analysis was to obtain predicted additive genetic effects (also known as estimated breeding values, EBVs) for use not only within the breeding programme but also the New Zealand and Australian forest industry as a whole. The breeding programme uses EBVs to select parents for crossing, with the aim of producing genetically improved germplasm. At an industry level, forest owners purchase seed from nurseries and pay a premium for seed from superior parents. The quality of the parents is reflected in a rating (GF Plus, 2006) that is derived from their EBVs. In the data-set under study, EBVs were required for the majority (3057) of the existing parents (so-called backward selections) and also the 1551 progeny trees in the clonal trials since they are potential new parents (forward selections). Here-after, these 4608 trees will be called "varieties" in order to align with the standard terminology of METs, in particular "variety by environment interaction". We stress that this is for pedagogical reasons only since *Pinus radiata* is an outcrossing species so these trees are not varieties in the sense of inbred crops. In terms of the trait of DBH, both the breeding programme and industry are primarily interested in varieties that will produce progeny that are likely to grow well across a wide range of environments as represented by the geographic locations and planting dates of the trials in the data-set.

This example illustrates some important requirements and difficulties that arise in practice. First is that pedigree information must be included in the analysis in order to investigate additive VEI and obtain EBVs. Second is the potential computational burden associated with modelling VEI given the large numbers of trees and environments. Cullis et al (2014) overcame this by proposing an approximate reduced animal (ARA) model which meant that additive VEI was modelled using only the 4608 varieties rather than the full set of 339589 trees in the pedigree. This reduced the size of the problem but we note that there remained a substantial degree of imbalance in the data since not all of these varieties were represented in all trials. In fact, only 4% of all possible variety by environment combinations

were represented. The inclusion of both clonal and non-clonal trials and the use of a range of experimental designs added to the complexity. The analysis must therefore be sophisticated enough to allow all of these issues to be adequately accommodated. Finally, it is noted that the breeding programme and industry require informative summaries of EBVs across environments in order to make selections and produce ratings.

It is instructive to consider any statistical analysis as comprising several distinct, but linked, components (Nelder, 1994) which can be condensed into (1) model fitting and checking and (2) inference and prediction of effects of interest. In the context of MET data, component (1) requires the use of a model with appropriate genetic and non-genetic effects. The former must encapsulate VEI and the latter must reflect sources of variation and correlation associated with individual trial plot structures. Component (2) requires summaries that provide breeders with concise and accurate information on which to base selection. In the presence of VEI this typically includes some measure of overall performance and "stability" across environments for each variety.

Many current approaches for the analysis of MET data focus on component (2) at the expense of component (1). So they aim to provide simple summaries by fitting simple models. We provide a brief history of these models here but the reader is referred to Smith et al (2005) for a comprehensive review. Early methods of analysis for MET data involved an Analysis of Variance (ANOVA) of the two-way table of variety by environment means. The total variation in the data was partitioned into sources due to varieties, environments (trials) and residual variation which is a composite of VEI and within-trial error. Overall performance for a variety was obtained as the estimate of the variety main effect. Various stability measures have been derived for this model (see Lin et al, 1986, for a review) and are typically some function of the residuals. A commonly used measure is the stability variance of Shukla (1972) which is the sample variance of the residuals for individual varieties.

A greater emphasis on interpreting VEI lead to the use of more complex models than ANOVA, in particular models involving regressions onto an independent environmental variable or onto the marginal (environment) means of the two-way table (Yates and Cochran, 1938; Finlay and Wilkinson, 1963). In these models the stability measure is the slope of the response for each variety. The motivation to examine more general patterns in VEI lead to the use of principal component analysis (PCA). Kempton (1984) used PCA on the residual effects from the two-way ANOVA model and displayed the results using bi-plots. This method of analysis was subsequently badged as AMMI (Additive Main effects and Multiplicative Interaction Gauch, 1992) and is still one of the most widely used methods for MET data. Note that the aim of this method is to be able to visualise relationships between varieties and environments. It does not provide simple numerical summaries that could be used for variety selection.

The models discussed thus far have numerous deficiencies. Some key issues are that they involve piecemeal approaches (typically first requiring analyses of individual trials to obtain variety by evironment means for use as data in a subsequent analysis) so are inherently inefficient (Welham et al, 2010; Gogel et al, 2018); most require balanced data, that is, all varieties in all environments; they assume variety effects to be fixed rather than random (see Smith et al, 2005, for a discussion),

which has particular limitations for our example since it is not possible to include pedigree information and finally, they rarely provide a good fit to the data.

These methods therefore do not satisfactorily address the model fitting component of the MET analysis. In contrast, we consider the linear mixed model approach of Smith et al (2001) and its extensions, in particular Oakey et al (2007) and Beeck et al (2010) who include pedigree information in order to partition genetic effects into additive and non-additive effects and Cullis et al (2014) who use a modification for estimating additive genetic effects in out-crossing plant species. Underpinning these approaches is a one-stage analysis of individual plot data combined across trials, and factor analytic (FA) structures for the variety effects in individual environments. Additionally separate non-genetic models are used for individual trials. The approach has been used in Australia for over 15 years and is now the preferred method of analysis in all major plant breeding programmes. The FA structure has been found to perform extremely well in terms of providing a good fit to the data and a parsimonious model for VEI (Kelly et al, 2007). The success of the approach for plant breeding programmes has also led to its adoption within the Australian National Variety Trials (NVT) system (Smith et al, 2015; Gogel et al, 2018).

In terms of component (2) of the analysis, various summaries from the FA model have been used to aid in examining VEI. These include heatmaps for visualising estimated genetic correlations between environments, and so-called latent regression plots which provide visual representations of the multiple regression implicit in the FA model (see Cullis et al, 2010, 2014, for example). The FA model can also be regarded as a random effects analogue of the AMMI model so that bi-plots as per Kempton (1984) or Gauch (1992) can be constructed. Although these graphics are informative they do not directly address the fundamental issue of variety selection. The FA mixed model provides predictions of genetic effects for each variety in each environment in the data-set. These predictions provide a complete inventory of the two-way table of variety by environment effects which need to be summarised to facilitate selection. Until now there has been no statistically or biologically satisfactory way to achieve this without sacrificing information on VEI.

To summarise, the Smith et al (2001) approach out-performs others in terms of the model fitting component of a MET analysis but it has failed to deliver on the prediction component, in the sense of providing concise information to aid with variety selection. We rectify this in the current paper by presenting factor analytic selection tools (FAST) which exploit the underlying form of the FA structure, in particular its analogy with multiple linear regression. FAST include natural measures of overall performance, stability and sensitivity for each variety which can then be used to form a relatively simple index for selection. We note that the measures accommodate selection for both broad adaptation and also for specific adaptation as guided by patterns of VEI revealed in the analysis itself. FAST can be routinely implemented in a plant breeding programme to identify both superior varieties and superior parents.

The paper is arranged as follows. In Section 2 the general linear mixed model for MET data with the inclusion of pedigree information is described, with particular attention given to the FA structures for variety by environment effects. The specific model for the motivating example is given. Methods for the investigation of VEI following the fitting of an FA structure are given in Section 3. This includes a

detailed development of the new tools (FAST). The application of FAST to the motivating example and some remarks on the general applicability of the tools are given in Section 4.

2 Statistical models

It is assumed that the MET data-set comprises t trials that have been conducted in p environments. Often, trials are synonymous with environments, so that $p = t$. However there are situations in which there are multiple trials in an environment, so that $p < t$. Let $\mathbf{y}_j$ denote the n_j-vector of data for the j^{th} trial. We then let $\mathbf{y}$ denote the $n-$vector of data combined across all trials in the MET, so write $\mathbf{y} = (\mathbf{y}_1^\top, \mathbf{y}_2^\top, \ldots, \mathbf{y}_t^\top)^\top$. Note that $n = \sum_{j=1}^{t} n_j$. The linear mixed model for $\mathbf{y}$ can be written as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\tau} + \mathbf{Z_g}\mathbf{u_g} + \mathbf{Z_p}\mathbf{u_p} + \mathbf{e}$$

where $\boldsymbol{\tau}$ is a vector of fixed effects with associated design matrix $\mathbf{X}$ (assumed to have full column rank); $\mathbf{u_g}$ is the vector of random genetic effects with associated design matrix $\mathbf{Z_g}$; $\mathbf{u_p}$ is a vector of random non-genetic (or peripheral) effects with associated design matrix $\mathbf{Z_p}$ and $\mathbf{e} = (\mathbf{e}_1^\top, \mathbf{e}_2^\top, \ldots, \mathbf{e}_t^\top)^\top$ is the combined vector of residuals from all trials. The vector of fixed effects includes mean parameters for individual environments. The vector of random peripheral effects includes effects associated with the plot structures of individual trials.

The random genetic effects comprise the variety effects nested within environments, and will be referred to as the variety by environment (VE) effects. If we let m denote the total number of unique varieties across all environments, then the vector $\mathbf{u_g}$ has length mp. Typically, not all varieties are grown in all environments so that the design matrix $\mathbf{Z_g}$ will contain columns in which all the elements are zero. We assume the VE effects to be ordered as varieties within environments so they can be written as $\mathbf{u_g} = (\mathbf{u_g}_1^\top, \mathbf{u_g}_2^\top, \ldots, \mathbf{u_g}_p^\top)^\top$ where $\mathbf{u_g}_j$ is the $m-$vector of VE effects for environment j.

The effects $\mathbf{u_g}, \mathbf{u_p}$ and $\mathbf{e}$ are assumed to be mutually independent, and distributed as multivariate Gaussian, with zero means. The variance matrix for $\mathbf{u_g}$ will be described in detail below. The variance matrix for $\mathbf{u_p}$ is typically given by $\mathbf{G_p} = \oplus_{i=1}^{b} \sigma_{p_i}^2 \mathbf{I}_{q_i}$ where b is the number of components in $\mathbf{u_p}$ and q_i is the number of effects in (length of) $\mathbf{u_p}_i$. The variance matrix for the residuals is assumed to be block diagonal, so that $\mathbf{R} = \oplus_{j=1}^{t} \mathbf{R}_j$ where $\mathbf{R}_j = \mathrm{var}(\mathbf{e}_j)$ is the variance matrix for the residuals for the j^{th} trial.

In this paper we allow for the inclusion of pedigree information in the analysis, so partition the VE effects into additive and non-additive (residual VE) effects as follows:

$$\mathbf{u_g} = \mathbf{u_a} + \mathbf{u_e}$$

It is assumed that $\mathrm{var}(\mathbf{u_a}) = \mathbf{G_a} \otimes \mathbf{A}$ where $\mathbf{A}$ is the numerator relationship matrix (see Oakey et al, 2007; Beeck et al, 2010, for example), and $\mathbf{G_a}$ is a $p \times p$ symmetric positive (semi)-definite matrix that will be referred to as the between environment additive genetic variance matrix. The diagonal elements are the variances of the additive VE effects for individual environments and the off-diagonal elements are the covariances between additive VE effects in different environments. In terms of the non-additive effects, it is assumed that $\mathrm{var}(\mathbf{u_e}) = \mathbf{G_e} \otimes \mathbf{I}_m$ where $\mathbf{G_e}$ is

a $p \times p$ symmetric positive (semi)-definite matrix that will be referred to as the between environment non-additive genetic variance matrix. The variance matrix of the total VE effects (that is, additive plus non-additive) is therefore given by

$$\text{var}\left(\mathbf{u_g}\right) = \mathbf{G_a} \otimes \mathbf{A} + \mathbf{G_e} \otimes \mathbf{I}_m \tag{1}$$

Note that if pedigree information is not included in the analysis, the VE effects simplify to $\mathbf{u_g} = \mathbf{u_e}$ with $\text{var}\left(\mathbf{u_g}\right) = \mathbf{G_e} \otimes \mathbf{I}_m$.

2.1 Variance models for genetic effects

Following Smith et al (2001) we propose a factor analytic model for the between environment additive genetic variance matrix. The aim is to account for the covariances of the additive VE effects between environments in terms of a small number, k_a, of (unknown) common factors. The number, k_a, is called the order of the model and we let FAk_a denote the FA model with this order. The model is postulated in terms of the additive VE effects as linear combinations of the common factors, plus an error term. Thus the additive VE effect for variety i and environment j is written as

$$u_{a_{ij}} = \lambda_{a_{1j}} f_{a_{1i}} + \lambda_{a_{2j}} f_{a_{2i}} + \ldots + \lambda_{a_{k_a j}} f_{a_{k_a i}} + \delta_{a_{ij}} \tag{2}$$

Each of the first k_a terms is the product of a variety effect ($f_{a_{ri}}$), which is known as a score, and an environment effect ($\lambda_{a_{rj}}$), which is known as a loading. The final term, $\delta_{a_{ij}}$, represents the error or lack of fit in the model. This model can be written in vector notation as

$$\begin{aligned} \mathbf{u_a} &= \left(\boldsymbol{\lambda}_{\mathbf{a}_1} \otimes \mathbf{I}_m\right) \mathbf{f}_{\mathbf{a}_1} + \left(\boldsymbol{\lambda}_{\mathbf{a}_2} \otimes \mathbf{I}_m\right) \mathbf{f}_{\mathbf{a}_2} + \ldots + \left(\boldsymbol{\lambda}_{\mathbf{a}_{k_a}} \otimes \mathbf{I}_m\right) \mathbf{f}_{\mathbf{a}_{k_a}} + \boldsymbol{\delta}_{\mathbf{a}} \\ &= \left(\boldsymbol{\Lambda}_{\mathbf{a}} \otimes \mathbf{I}_m\right) \mathbf{f}_{\mathbf{a}} + \boldsymbol{\delta}_{\mathbf{a}} \end{aligned} \tag{3}$$

where $\boldsymbol{\lambda}_{\mathbf{a}_r}$ is the $p-$vector of environment loadings for the r^{th} common factor and $\mathbf{f}_{\mathbf{a}_r}$ is the associated $m-$vector of variety scores; $\boldsymbol{\Lambda}_{\mathbf{a}} = \left[\boldsymbol{\lambda}_{\mathbf{a}_1}\ \boldsymbol{\lambda}_{\mathbf{a}_2}\ \ldots\ \boldsymbol{\lambda}_{\mathbf{a}_{k_a}}\right]$ is the $p \times k_a$ matrix of loadings; $\mathbf{f_a} = \left(\mathbf{f}_{\mathbf{a}_1}^\top, \mathbf{f}_{\mathbf{a}_2}^\top, \ldots, \mathbf{f}_{\mathbf{a}_{k_a}}^\top\right)^\top$ is the mk_a-vector of variety scores and $\boldsymbol{\delta}_{\mathbf{a}} = (\boldsymbol{\delta}_{\mathbf{a}_1}^\top, \boldsymbol{\delta}_{\mathbf{a}_2}^\top, \ldots, \boldsymbol{\delta}_{\mathbf{a}_p}^\top)^\top$ is the $mp-$vector of lack of fit effects where $\boldsymbol{\delta}_{\mathbf{a}j}$ is the $m-$vector for the j^{th} environment.

It is assumed that $\mathbf{f_a}$ and $\boldsymbol{\delta}_{\mathbf{a}}$ are independent and are distributed as multivariate Gaussian with zero means and variance matrices given by

$$\text{var}\left(\mathbf{f_a}\right) = \mathbf{I}_{k_a} \otimes \mathbf{A} \quad \text{and} \quad \text{var}\left(\boldsymbol{\delta}_{\mathbf{a}}\right) = \boldsymbol{\Psi}_{\mathbf{a}} \otimes \mathbf{A}$$

where $\boldsymbol{\Psi}_{\mathbf{a}}$ is a $p \times p$ diagonal matrix with elements ψ_{aj}, which are the so-called additive specific variances for individual environments. These assumptions lead to a factor analytic form for the between environment additive genetic variance matrix, namely $\mathbf{G_a} = (\boldsymbol{\Lambda}_{\mathbf{a}}\boldsymbol{\Lambda}_{\mathbf{a}}^\top + \boldsymbol{\Psi}_{\mathbf{a}})$. Note that the variance matrix for the additive VE effects for variety i is given by $a_{ii}\mathbf{G_a}$, where a_{ii} is the i^{th} diagonal element of $\mathbf{A}$ and is defined to be $1 + F_i$, where F_i is the inbreeding coefficient for variety i.

The FA model in equation (3) can be naturally separated in two parts. We let $\boldsymbol{\beta}_{\mathbf{a}} = \left(\boldsymbol{\Lambda}_{\mathbf{a}} \otimes \mathbf{I}_m\right) \mathbf{f_a}$ so that

$$\mathbf{u_a} = \boldsymbol{\beta}_{\mathbf{a}} + \boldsymbol{\delta}_{\mathbf{a}}$$

The effects in $\boldsymbol{\beta}_{\mathbf{a}}$ will be called the *common* additive VE effects because they are associated with the common factors that explain the additive genetic covariance

between environments. To show this we partition $\boldsymbol{\beta}_{\mathbf{a}}$ conformably with $\mathbf{u}_{\mathbf{a}}$ so that $\boldsymbol{\beta}_{\mathbf{a}j}$ denotes the $m-$vector for the j^{th} environment. Table 1 shows that the covariance between $\mathbf{u}_{\mathbf{a}j}$ and $\mathbf{u}_{\mathbf{a}h}$ is identical to, and therefore entirely defined by, the covariance between $\boldsymbol{\beta}_{\mathbf{a}j}$ and $\boldsymbol{\beta}_{\mathbf{a}h}$. The common VE effects may also be thought of as the *correlated* VE effects since the effects for one environment are correlated with those in at least one other environment. In contrast, the lack of fit effects, $\boldsymbol{\delta}_{\mathbf{a}j}$ are uncorrelated between environments (see Table 1). In other words these effects are *specific to* the environment so that $\boldsymbol{\delta}_{\mathbf{a}}$ will be called the vector of specific additive VE effects.

Table 1 Variances of additive VE effects in environment j and covariances between environments j and h when a factor analytic model of order k_a is fitted. Variances and covariances are given for VE, common VE and specific VE effects.

VE effects for environment j	Variance of VE effects for environment j	Covariance of VE effects in environments j and h
VE ($\mathbf{u}_{\mathbf{a}j}$)	$\sum_{r=1}^{k_a} \lambda_{a_{rj}}^2 \mathbf{A} + \psi_{a_j} \mathbf{A}$	$\sum_{r=1}^{k_a} \lambda_{a_{rj}} \lambda_{a_{rh}} \mathbf{A}$
Common VE ($\boldsymbol{\beta}_{\mathbf{a}j}$)	$\sum_{r=1}^{k_a} \lambda_{a_{rj}}^2 \mathbf{A}$	$\sum_{r=1}^{k_a} \lambda_{a_{rj}} \lambda_{a_{rh}} \mathbf{A}$
Specific VE ($\boldsymbol{\delta}_{\mathbf{a}j}$)	$\psi_{a_j} \mathbf{A}$	$0\mathbf{A}$

In a similar manner to the additive VE effects, a factor analytic model may also be assumed for the non-additive effects. The order of the model is denoted by k_e and this may differ from k_a. The FAk_e model for the non-additive effects is then given by

$$\begin{aligned} \mathbf{u}_{\mathbf{e}} &= (\boldsymbol{\lambda}_{\mathbf{e}_1} \otimes \mathbf{I}_m)\, \mathbf{f}_{\mathbf{e}_1} + (\boldsymbol{\lambda}_{\mathbf{e}_2} \otimes \mathbf{I}_m)\, \mathbf{f}_{\mathbf{e}_2} + \ldots + \left(\boldsymbol{\lambda}_{\mathbf{e}_{k_e}} \otimes \mathbf{I}_m\right) \mathbf{f}_{\mathbf{e}_{k_e}} + \boldsymbol{\delta}_{\mathbf{e}} \\ &= (\boldsymbol{\Lambda}_{\mathbf{e}} \otimes \mathbf{I}_m)\, \mathbf{f}_{\mathbf{e}} + \boldsymbol{\delta}_{\mathbf{e}} \\ &= \boldsymbol{\beta}_{\mathbf{e}} + \boldsymbol{\delta}_{\mathbf{e}} \end{aligned} \tag{4}$$

where the terms are defined in an analogous manner to those for the additive VE effects. It is assumed that $\mathbf{f}_{\mathbf{e}}$ and $\boldsymbol{\delta}_{\mathbf{e}}$ are independent and are distributed as multivariate Gaussian with zero means and variance matrices given by

$$\text{var}\,(\mathbf{f}_{\mathbf{e}}) = \mathbf{I}_{k_e} \otimes \mathbf{I}_m \quad \text{and} \quad \text{var}\,(\boldsymbol{\delta}_{\mathbf{e}}) = \boldsymbol{\Psi}_{\mathbf{e}} \otimes \mathbf{I}_m$$

and thence $\mathbf{G}_{\mathbf{e}} = (\boldsymbol{\Lambda}_{\mathbf{e}} \boldsymbol{\Lambda}_{\mathbf{e}}^{\top} + \boldsymbol{\Psi}_{\mathbf{e}})$.

2.2 Model fitting, estimation and prediction

The model fitting process commences with the estimation of the variance parameters using residual maximum likelihood (REML). Then given these estimates, empirical best linear unbiased estimates (EBLUEs) and empirical best linear unbiased predictions (EBLUPs) of the fixed and random effects, respectively, can be obtained.

All models in this paper have been fitted using ASReml-R (Butler et al, 2009). In terms of the FA models, a sparse implementation of the average information algorithm (Thompson et al, 2003) is used so that reduced rank (RR) variance models are accommodated. In this paper we have exploited this by splitting the FA models into their component parts and explicitly fitting the common VE effects (with RR variance structure) separately from the specific VE effects. The associated variance parameters are the loadings and specific variances. The REML estimates of these parameters will be denoted by $\hat{\lambda}_{s_{rj}}$ and $\hat{\psi}_{s_j}$ (for $s = a, e$). It is important to note that when $k_s > 1$, the matrix of loadings is not unique so that constraints are required to ensure identifiability. The constraints used in ASReml-R (Butler et al, 2009) are to fix the elements in the upper triangle of $\hat{\mathbf{\Lambda}}_\mathbf{s}$ to zero.

EBLUEs and EBLUPs of the fixed and random effects are obtained as solutions to the mixed model equations (MME) (Henderson, 1950). Given the sparse RR formulation of the FA model (Thompson et al, 2003), this provides EBLUPs of the common and specific VE effects and the variety scores, namely $\tilde{\boldsymbol{\beta}}_\mathbf{s}$, $\tilde{\boldsymbol{\delta}}_\mathbf{s}$ and $\tilde{\mathbf{f}}_\mathbf{s}$. Note that EBLUPs of the VE effects can then be obtained as $\tilde{\mathbf{u}}_\mathbf{s} = \tilde{\boldsymbol{\beta}}_\mathbf{s} + \tilde{\boldsymbol{\delta}}_\mathbf{s}$.

Note also, that if $\mathbf{C}$ is used to denote the coefficient matrix of the MME, then $\mathbf{C}^{-1}$ provides prediction error variances of effects. These can then be used to calculate a measure of accuracy for predictions of genetic effects. We let $\tilde{\mathbf{u}}_\mathbf{c} = (\tilde{\mathbf{u}}_\mathbf{c_a}^\top, \tilde{\mathbf{u}}_\mathbf{c_e}^\top)^\top$ be the vector of EBLUPs of all genetic effects in the MME and where $\tilde{\mathbf{u}}_\mathbf{c_s} = (\tilde{\boldsymbol{\beta}}_\mathbf{s}^\top, \tilde{\boldsymbol{\delta}}_\mathbf{s}^\top, \tilde{\mathbf{f}}_\mathbf{s}^\top)^\top$. We consider scalar predictions of the form $\tilde{\pi} = \mathbf{d}^\top \tilde{\mathbf{u}}_\mathbf{c}$. The accuracy of this prediction is given by

$$\text{cor}\,(\tilde{\pi},\ \pi) = \sqrt{1 - \text{pev}\,(\tilde{\pi})\,/\text{var}\,(\pi)} \tag{5}$$

where $\text{pev}\,(\tilde{\pi})$ is the prediction error variance which is given by $\mathbf{d}^\top \mathbf{C}^{cc} \mathbf{d}$ where $\mathbf{C}^{cc}$ is the partition of $\mathbf{C}^{-1}$ that relates to $\tilde{\mathbf{u}}_\mathbf{c}$.

2.3 Statistical model for motivating example

The data from the motivating example were analysed using the approximate reduced animal (ARA) model of Cullis et al (2014). In the current paper the analysis itself is not of primary interest, but rather the post-processing of the results to facilitate selection. The interested reader is therefore referred to Cullis et al (2014) for full details of the linear mixed model. The key features of the linear mixed model for the motivating example were the inclusion of a fixed main effect for each trial and random effects associated with experimental design terms. In accordance with the ARA model, additive VE effects were included for parents and clones (across all trials) and non-additive VE effects were included for clones (across clonal trials only). The former were modelled using a factor analytic model of order 3 (FA3). The variance matrix for the latter was assumed to have a diagonal form since the estimated non-additive genetic variances were found to be relatively small. A separate residual variance was fitted for each trial.

3 Post-processing to investigate variety by environment interaction

Varietal selection in METs is made more complex by the presence of VEI. VEI is characterised by the differential response of varieties to environments and can

be broadly categorised as either crossover or non-crossover. The former is often regarded as the most important for varietal selection since it is associated with changes in the rank of varieties between environments. In this section tools for exploring VEI will be developed in the context of the additive genetic effects. The methods are directly applicable to the non-additive effects but the total VE effects require special attention and will be discussed in section 3.4.

It is useful to write the additive VE effects as a two-way structure, namely the $m \times p$ matrix $\mathbf{U_a} = [\mathbf{u_{a_1}}\ \mathbf{u_{a_2}}\ \ldots\ \mathbf{u_{a_p}}]$. By definition, VEI involves interrelationships between the columns (environments) and rows (varieties) of this matrix and may be considered from either perspective. In terms of variety selection, it is crucial to focus on VEI from the variety perspective. This can be done by exploring various aspects of the FA model, in particular the analogy with multiple regression. Given the FAk_a model, the EBLUPs of the additive VE effects can be written as

$$\begin{aligned}\tilde{\mathbf{u}}_\mathbf{a} &= \left(\hat{\boldsymbol{\lambda}}_{\mathbf{a}_1} \otimes \mathbf{I}_m\right)\tilde{\mathbf{f}}_{\mathbf{a}_1} + \left(\hat{\boldsymbol{\lambda}}_{\mathbf{a}_2} \otimes \mathbf{I}_m\right)\tilde{\mathbf{f}}_{\mathbf{a}_2} + \ldots + \left(\hat{\boldsymbol{\lambda}}_{\mathbf{a}_{k_a}} \otimes \mathbf{I}_m\right)\tilde{\mathbf{f}}_{\mathbf{a}_{k_a}} + \tilde{\boldsymbol{\delta}}_\mathbf{a} \qquad (6)\\ &= \tilde{\boldsymbol{\beta}}_\mathbf{a} + \tilde{\boldsymbol{\delta}}_\mathbf{a}\end{aligned}$$

Equation (6) can then be viewed as a series of multiple regressions in which the independent variables are the estimated environment loadings, $\hat{\boldsymbol{\lambda}}_{\mathbf{a}_1}\ \ldots\ \hat{\boldsymbol{\lambda}}_{\mathbf{a}_{k_a}}$. There is a separate regression for each variety and the predicted regression coefficients are given by the predicted variety scores, $\tilde{\mathbf{f}}_{\mathbf{a}_1}\ \ldots\ \tilde{\mathbf{f}}_{\mathbf{a}_{k_a}}$.

3.1 Rotation of REML estimates of loadings

In later sections, individual terms in the regression will be explored. When $k_a > 1$, this first requires the independent variables, namely the estimated loadings, to be rotated to a meaningful solution. Recall from section 2.2 that when $k_a > 1$, the loadings are estimated subject to constraints that are imposed for computational convenience. In order for the estimated loadings to have a meaningful interpretation we choose to rotate to a principal component solution so that the first rotated estimated loading accounts for the maximum amount of covariance in the VE effects, the second accounts for the next largest amount and is orthogonal to the first, and so on. The properties of orthogonality and decreasing contributions to covariance will be shown to be important for the selection tools developed in Section 3.3.

Thus, after model fitting, we obtain the singular value decomposition of $\hat{\boldsymbol{\Lambda}}_\mathbf{a}$:

$$\hat{\boldsymbol{\Lambda}}_\mathbf{a} = \mathbf{B_a}\mathbf{L_a}\mathbf{V_a}$$

where $\mathbf{L_a}$ is a diagonal matrix with elements given by the square roots of the eigenvalues of $\hat{\boldsymbol{\Lambda}}_\mathbf{a}\hat{\boldsymbol{\Lambda}}_\mathbf{a}^\top$ (arranged in decreasing order) and $\mathbf{B_a}$ and $\mathbf{V_a}$ are orthogonal matrices with columns given by the eigenvectors of $\hat{\boldsymbol{\Lambda}}_\mathbf{a}\hat{\boldsymbol{\Lambda}}_\mathbf{a}^\top$ and $\hat{\boldsymbol{\Lambda}}_\mathbf{a}^\top\hat{\boldsymbol{\Lambda}}_\mathbf{a}$, respectively. Then we obtain the rotated estimated loadings as

$$\hat{\boldsymbol{\Lambda}}_\mathbf{a}^* = c\hat{\boldsymbol{\Lambda}}_\mathbf{a}\mathbf{V_a}$$

where c is a constant that is either 1 or -1. The sign is chosen to ensure the majority of first rotated loadings are positive rather than negative. This aids with

interpretation, particularly in the context of the selection tools developed in Section 3.3.

After rotation it is meaningful to consider the percentage of additive genetic variance that is explained by individual factors. This can be computed for factor r and environment j as

$$v_{a_{rj}} = 100(\hat{\lambda}^*_{a_{rj}})^2 / \left(\sum_{s=1}^{k_a} (\hat{\lambda}^*_{a_{sj}})^2 + \hat{\psi}_{a_j} \right)$$

In this way the mean contribution of factor r to additive genetic variance can be computed across environments as $\sum_{j=1}^{p} v_{a_{rj}}/p$.

Given the rotation of the estimated loadings, the predicted variety scores must also be rotated to:

$$\tilde{\mathbf{f}}^*_{\mathbf{a}} = \left(c\mathbf{V}_{\mathbf{a}}{}^{\top} \otimes \mathbf{I}_m \right) \tilde{\mathbf{f}}_{\mathbf{a}}$$

Note that the EBLUPs of both the VE effects and the common VE effects are invariant to the rotation.

3.1.1 Rotation of REML estimates of loadings for motivating example

In the motivating example $k_a = 3$ factors were fitted. The estimated loadings after rotation are given in the Appendix. In summary, the first vector of loadings ranges from 1.0 to 19.1 with a mean of 10.0; the second vector ranges from -18.4 to 13.6 with a mean of 0.9; the third vector ranges from -11.8 to 14.7 with a mean of -0.4. The mean percentage variance accounted for by individual factors was 50.7%, 14.8% and 13.7% for factors 1, 2 and 3, respectively, and 79.3% for the FA3 model as a whole (that is, the mean across all factors and environments). Note that the order of $k_a = 3$ was chosen using a pragmatic approach that aimed to balance parsimony and goodness of fit. The latter was assessed in terms of both the overall percentage variance accounted for and the distribution of individual environment values (also see Cullis et al, 2014). With respect to the latter, a large number (53) of the environments had a variance accounted for greater than 80%, whilst only a small number (14) had a variance accounted for less than 50%.

3.2 VE effects for individual environments

It would seem intuitive to use the EBLUPs, $\tilde{\mathbf{u}}_{\mathbf{a}}$, of the additive VE effects to examine variety performance in individual environments. However, the EBLUPs, $\tilde{\boldsymbol{\beta}}_{\mathbf{a}}$, of the common additive VE effects provide an alternative that have several appealing features. First, by definition, they represent VEI that is driven by influences that are common to several environments and therefore exclude isolated VEI that is specific to a single environment. They also have a natural interpretation as a set of "smoothed" VE effects. This is clear using the regression analogy of an FA model. In a similar manner to a standard multiple regression problem, predictions of the dependent variable in equation (6) may be obtained as fitted values along the regression surface. These are given by the sum of the first k_a terms which is equivalent to the EBLUPs, $\tilde{\boldsymbol{\beta}}_{\mathbf{a}}$, of the common VE effects.

Finally we note that the EBLUPs of the common VE effects provides a set of predictions that are compatible across environments, irrespective of whether there

is data on the variety in the environment. This is because they are fitted values on the regression surface. In contrast, the EBLUPs of the VE effects also include the EBLUPs of the specific VE effects which are the residuals in the regression. Predictions of the specific VE effects are intrinsically different for varieties with and without data in an environment. In the case of non-additive effects, the EBLUP of the specific VE effect will be zero when there is no data on the variety in that environment. Cullis et al (2010) provide a more theoretical discussion of this issue and recommend the use of predictions that are "marginal" to the specific VE effects. Thus in the remainder of this paper it will be assumed that the EBLUPs of the common VE effects will be used for selections. Generalisations for the use of VE effects are straight-forward.

Note that all predictions can be accompanied by a measure of accuracy as detailed in Section 2.2 so that, in particular, the accuracy of VE predictions with and without data can be assessed.

3.3 Factor Analytic Selection tools (FAST)

From a breeding perspective, varietal selection using the full matrix of predicted common VE effects may be a formidable task unless the number of environments is very small. Broadly speaking, there is a need to obtain measures of overall performance and stability across environments for each variety.

Cullis et al (2014) suggested the use of latent regression plots to examine variety stability. These are similar to added variable plots but exploit the orthogonality property of the rotated factors. They comprise a series of k_a plots for each variety in which the $y-$ and $x-$ axes for variety i are defined by

Plot 1: $y_j = \tilde{\beta}_{a_{ij}}$ and $x_j = \hat{\lambda}^*_{a_{1j}}$
Plot 2: $y_j = \tilde{\beta}_{a_{ij}} - \hat{\lambda}^*_{a_{1j}} \tilde{f}^*_{a_{1i}}$ and $x_j = \hat{\lambda}^*_{a_{2j}}$
$\vdots$
Plot k_a: $y_j = \tilde{\beta}_{a_{ij}} - \sum_{r=1}^{k_a-1} \hat{\lambda}^*_{a_{rj}} \tilde{f}^*_{a_{ri}}$ and $x_j = \hat{\lambda}^*_{a_{k_a j}}$

The points on plot r $(= 1 \dots k_a)$ can be supplemented with a line which has slope given by the EBLUP of the variety score for that factor, that is, $\tilde{f}^*_{a_{ri}}$.

In the examination of the latent regression plots from their analysis, Cullis et al (2014) state that "Since all the estimated loadings for [the first] factor are positive this then means that large positive regression coefficients [scores] for this factor are desirable for DBH." This is an oblique reference to overall performance. In terms of stability, Cullis et al (2014) comment on the "sensitivity" of varieties to individual factors as depicted in the plots.

Latent regression plots are very informative, but are not an ideal tool for selection since they only provide an informal examination of overall performance and stability. Additionally, it can be a laborious task to examine and compare the plots for all varieties under consideration. Typically, breeders require one or two relevant measures to incorporate into a selection index. In the following we summarise information in the latent regression plots to obtain formal measures of overall performance and stability that can be used for this purpose.

In order to develop these measures we separate the EBLUPs of the common VE effects in equation (6) into the effects associated with the first factor and the

remainder. Thus for variety i and environment j we have

$$\tilde{\beta}_{a_{ij}} = \hat{\lambda}^*_{a_{1j}} \tilde{f}^*_{a_{1i}} + \tilde{\epsilon}^*_{a_{ij}} \tag{7}$$

If represented graphically for a single variety, this corresponds to the first latent regression plot. If all varieties are graphed together, this represents a series of m straight lines, with the slope for variety i being given by $\tilde{f}^*_{a_{1i}}$. The points along the regression lines are the fitted values for the first factor, which are given by $\hat{\lambda}^*_{a_{1j}} \tilde{f}^*_{a_{1i}}$, and the deviations from the regression lines are given by $\tilde{\epsilon}^*_{a_{ij}}$. There are no explicit intercepts included in the model so that all lines pass through the origin, that is the point $(0, 0)$. Thus if all the (rotated) estimated loadings for the first factor are positive, the lines for any pair of varieties do not intersect within the range of the data. This means that the fitted values associated with the first factor represent non-crossover interaction and this characteristic can be exploited to obtain measures of both overall performance and stability.

The concepts are illustrated in Figure 1 for two varieties, labelled V1 and V6, from the example. In terms of the fitted values for the first factor, that is, $\hat{\lambda}^*_{a_{1j}} \tilde{f}^*_{a_{1i}}$, it is clear that the rankings of the varieties do not change between environments because the regression lines do not intersect. This is due to the fact that the rotated loadings are all positive (see Appendix). Furthermore, the lines diverge which indicates that the difference between the fitted values for the two varieties increases as the loadings increase. This is the classic representation of non-crossover VEI and is typically linked to changes in scale or the so-called "discriminating ability" of environments. It is therefore natural to use the fitted values for the first factor to form a measure of overall performance for each variety. The predicted scores $\tilde{f}^*_{a_{1i}}$ could be used for this purpose, but because they are standardised, it may be preferable to obtain a measure that is on the same scale as the trait being analysed. If the analogy with regression is continued, an obvious choice is the fitted value at the mean value of the "regressor" which is also the mean of the fitted values. If we let $\bar{\lambda}_1$ denote the mean of the rotated estimated loadings for the first factor, then the overall performance (OP) measure for variety i is computed as

$$\bar{\lambda}_1 \tilde{f}^*_{a_{1i}} = \frac{1}{p} \sum_{j=1}^{p} \hat{\lambda}^*_{a_{1j}} \tilde{f}^*_{a_{1i}} \tag{8}$$

In the example, $\bar{\lambda}_1 = 10.0$ and the OP of the two varieties is given by 34.9cm for V1 and 18.8cm for V6 (also see Figure 1).

If the fitted values from the first factor regression represent non-crossover VEI for pairs of varieties, it is also natural to base measures of variety stability on the remaining factors. In this way, changes in variety performance due primarily to changes in scale are eliminated from the examination of stability. A single global measure of stability for each variety can be obtained as the root mean square deviation (RMSD) from the regression line associated with the first factor. This is given for variety i by

$$\sqrt{\frac{1}{p} \sum_{j=1}^{p} \tilde{\epsilon}^{*2}_{a_{ij}}} \tag{9}$$

As with OP this measure is on the scale of the data. This measure of stability is easily visualised using the first latent regression plot. For example, Figure 1 shows

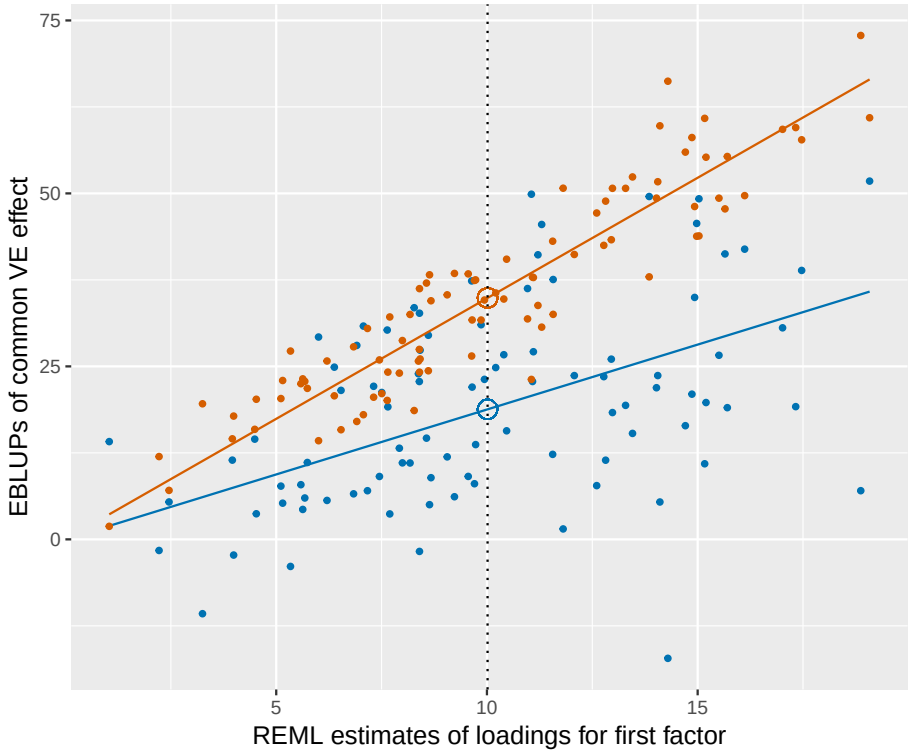

Fig. 1 Superimposed first latent regression plots for two varieties, V6 (coloured blue) and V1 (coloured orange). Slopes of the solid lines are given by the EBLUPs of the (rotated) variety scores for the first factor. The open circles are the overall performance measure for each variety, namely the value on the regression line at the mean value of the estimated loadings for the first factor (vertical dotted line).

a much larger spread about the line for variety V6 compared with V1 and this is formally quantified by the RMSD of 5.6cm for V1 and 12.7cm for V6.

In addition to the global measure of stability, responses of varieties to individual factors (excluding the first) may be of interest. These can be assessed visually using the latent regression plots $2 \dots k_a$ (see Cullis et al, 2014, for example). The essential information on each plot is the slope of the line which is given by $\tilde{f}^*_{a_{ri}}$ for variety i and factor r. These slopes could be used to rank varieties in terms of their responsiveness to the factors but they do not give any indication of the magnitude of the response with reference to the data. Given that the (rotated) estimated loadings for each of the factors $2 \dots k_a$ typically include both positive and negative values, (see Appendix, for example), they reflect contrasts between environments. In this case, the magnitude of the response of a variety to the factor may be quantified as the average contrast in terms of the fitted values, namely the mean of the fitted values for positive loadings minus the mean of the fitted values for negative loadings. If we let $\bar{\lambda}_{r+}$ and $\bar{\lambda}_{r-}$ denote the mean of the positive and negative estimated loadings for factor r, then the responsiveness of variety i

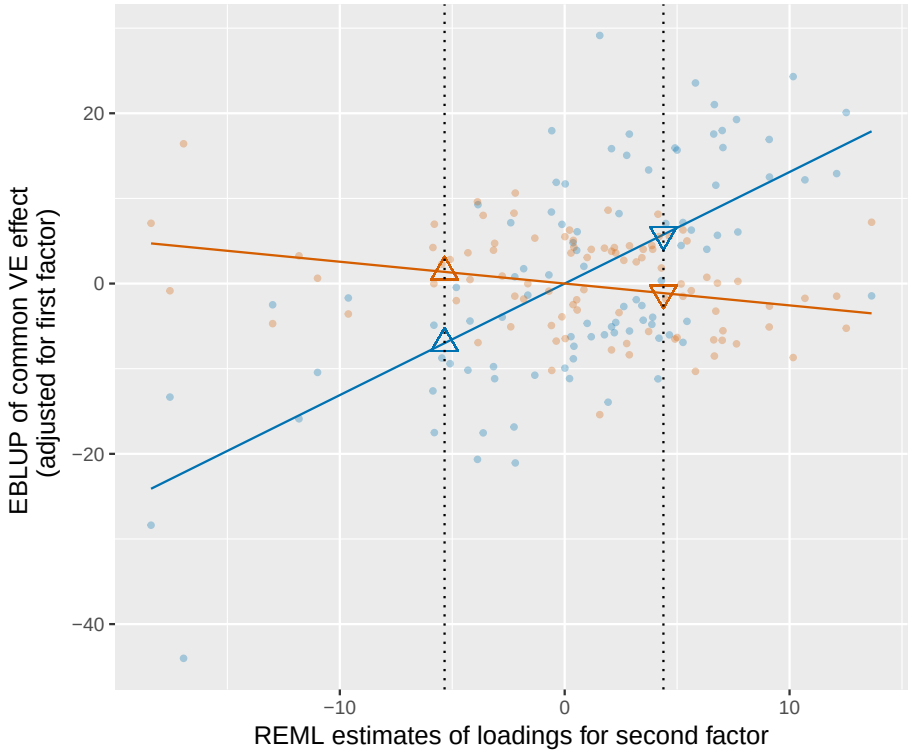

Fig. 2 Superimposed second latent regression plots for two varieties, V6 (coloured blue) and V1 (coloured orange). Slopes of the solid lines are given by the EBLUPs of the (rotated) variety scores for the second factor. The open triangles for each variety are the values on the regression line at the mean of the positive (downward pointing triangles) and negative (upward pointing triangles) estimated loadings for the second factor. The associated mean loadings are shown as vertical dotted lines.

to factor r is computed as

$$\left(\bar{\lambda}_{r+} - \bar{\lambda}_{r-}\right) \tilde{f}^{*}_{a_{ri}} \tag{10}$$

This is illustrated graphically for varieties V1 and V6 for the second factor in Figure 2. The means of the positive and negative estimated loadings are given by $\bar{\lambda}_{r+} = 4.4$ and $\bar{\lambda}_{r-} =$ -5.3. The responsiveness to the second factor for each variety is represented by the vertical distance from the downward pointing open triangle to the upward pointing open triangle. This is given by -1.1 - 1.4 = -2.5cm for V1 and 5.8 - (-7) = 12.8cm for V6. Thus, on average, the predicted common VE effects for variety V6 increase by 12.8cm in response to the covariate implicit in the second factor, whereas they decrease by 2.5cm for V1.

3.4 FAST for total VE effects

Given that factor analytic variance structures may be used for both the additive and non-additive VE effects, it is possible to extend the concepts developed in

section 3.3 for total (additive plus non-additive) VE effects. This is achieved by noting that it is possible to derive a special factor analytic form for the variance matrix of the total VE effects. We commence by deriving the between environment total genetic variance matrix, $\mathbf{G_g}$, say, that has an analogous interpretation to the additive and non-additive matrices. A complication arises due to the fact that the variance matrix for the total VE effects (see equation (1)) is a weighted sum of the additive and non-additive variance matrices. The variance matrix of the total VE effects for variety i is given by $a_{ii}\mathbf{G_a}+\mathbf{G_e}$ so that the pattern of heterogeneity of correlation between environments differs between varieties and depends on the inbreeding coefficient. We therefore consider either a range of inbreeding values, or in simple cases a single inbreeding coefficient that is representative of the varieties in the data-set. Either way, we proceed with a pre-specified inbreeding coefficient, $\bar{F}=\bar{a}-1$, say, and define

$$\mathbf{G_g}=\bar{a}\mathbf{G_a}+\mathbf{G_e} \tag{11}$$

to be the between environment total genetic variance matrix (for a given level of inbreeding).

Using the FA forms for $\mathbf{G_a}$ and $\mathbf{G_e}$ as presented in section 2.1, the between environment total genetic variance matrix of equation (11) can be written as

$$\begin{aligned}\mathbf{G_g}&=\bar{a}\left(\mathbf{\Lambda_a}\mathbf{\Lambda_a}^{\top}+\mathbf{\Psi_a}\right)+\left(\mathbf{\Lambda_e}\mathbf{\Lambda_e}^{\top}+\mathbf{\Psi_e}\right)\\&=\mathbf{\Lambda_g}\mathbf{\Lambda_g}^{\top}+\mathbf{\Psi_g}\end{aligned} \tag{12}$$

where $\mathbf{\Lambda_g}=\left[\sqrt{\bar{a}}\mathbf{\Lambda_a}\ \mathbf{\Lambda_e}\right]$ and $\mathbf{\Psi_g}=\bar{a}\mathbf{\Psi_a}+\mathbf{\Psi_e}$. This has the form of an FA variance structure of order $k_g=k_a+k_e$. Finally, we note that this variance structure arises from a model for the total VE effects, which is obtained as the sum of the models for the additive and non-additive VE effects in equations (3) and (4), but with a simple re-scaling of the loadings and scores for the additive effects:

$$\begin{aligned}\mathbf{u_g}&=\left(\sqrt{\bar{a}}\boldsymbol{\lambda}_{\mathbf{a}_1}\otimes\mathbf{I}_m\right)\mathbf{f}_{\mathbf{a}_1}/\sqrt{\bar{a}}+\ldots+\left(\sqrt{\bar{a}}\boldsymbol{\lambda}_{\mathbf{a}_{k_a}}\otimes\mathbf{I}_m\right)\mathbf{f}_{\mathbf{a}_{k_a}}/\sqrt{\bar{a}}+\boldsymbol{\delta}_{\mathbf{a}}\\&\quad+\left(\boldsymbol{\lambda}_{\mathbf{e}_1}\otimes\mathbf{I}_m\right)\mathbf{f}_{\mathbf{e}_1}+\ldots+\left(\boldsymbol{\lambda}_{\mathbf{e}_{k_e}}\otimes\mathbf{I}_m\right)\mathbf{f}_{\mathbf{e}_{k_e}}+\boldsymbol{\delta}_{\mathbf{e}}\\&=\left(\boldsymbol{\lambda}_{\mathbf{g}_1}\otimes\mathbf{I}_m\right)\mathbf{f}_{\mathbf{g}_1}+\left(\boldsymbol{\lambda}_{\mathbf{g}_2}\otimes\mathbf{I}_m\right)\mathbf{f}_{\mathbf{g}_2}+\ldots+\left(\boldsymbol{\lambda}_{\mathbf{g}_{k_g}}\otimes\mathbf{I}_m\right)\mathbf{f}_{\mathbf{g}_{k_g}}+\boldsymbol{\delta}_{\mathbf{g}}\\&=\left(\mathbf{\Lambda_g}\otimes\mathbf{I}_m\right)\mathbf{f_g}+\boldsymbol{\delta}_{\mathbf{g}}\end{aligned} \tag{13}$$

where $\boldsymbol{\lambda}_{\mathbf{g}_1}=\sqrt{\bar{a}}\boldsymbol{\lambda}_{\mathbf{a}_1},\ldots,\boldsymbol{\lambda}_{\mathbf{g}_{k_a}}=\sqrt{\bar{a}}\boldsymbol{\lambda}_{\mathbf{a}_{k_a}}$, $\boldsymbol{\lambda}_{\mathbf{g}_{(k_a+1)}}=\boldsymbol{\lambda}_{\mathbf{e}_1},\ldots,\boldsymbol{\lambda}_{\mathbf{g}_{k_g}}=\boldsymbol{\lambda}_{\mathbf{e}_{k_e}}$ and $\mathbf{f}_{\mathbf{g}_1}=\mathbf{f}_{\mathbf{a}_1}/\sqrt{\bar{a}},\ldots,\mathbf{f}_{\mathbf{g}_{k_a}}=\mathbf{f}_{\mathbf{a}_{k_a}}\sqrt{\bar{a}}$, $\mathbf{f}_{\mathbf{g}_{(k_a+1)}}=\mathbf{f}_{\mathbf{e}_1},\ldots,\mathbf{f}_{\mathbf{g}_{k_g}}=\mathbf{f}_{\mathbf{e}_{k_e}}$ and $\boldsymbol{\delta}_{\mathbf{g}}=\boldsymbol{\delta}_{\mathbf{a}}+\boldsymbol{\delta}_{\mathbf{e}}$. The variance matrix of the VE effects from equation (13) for a variety with inbreeding coefficient of $\bar{F}=\bar{a}-1$ is equivalent to that given in equation (12).

In order to apply selection tools, we first obtain EBLUPs of the common total VE effects as $\tilde{\beta}_{g_{ij}}=\tilde{\beta}_{a_{ij}}+\tilde{\beta}_{e_{ij}}$. The REML estimate of $\mathbf{\Lambda_g}$ in equation (13) is obtained by replacing $\boldsymbol{\lambda}_{\mathbf{a}_r}$ and $\boldsymbol{\lambda}_{\mathbf{e}_r}$ with their (unrotated) REML estimates. Similarly the EBLUP of $\mathbf{f_g}$ is obtained by replacing $\mathbf{f}_{\mathbf{a}_r}$ and $\mathbf{f}_{\mathbf{e}_r}$ with their EBLUPs. The matrix of estimated loadings is then rotated to a principal component solution as described in section 3.1. This provides $\hat{\mathbf{\Lambda}}^{*}_{\mathbf{g}}$ and also $\tilde{\mathbf{f}}^{*}_{\mathbf{g}}$. The tools of section 3.3 are then directly applicable. The implementation of this approach is the subject of future work.

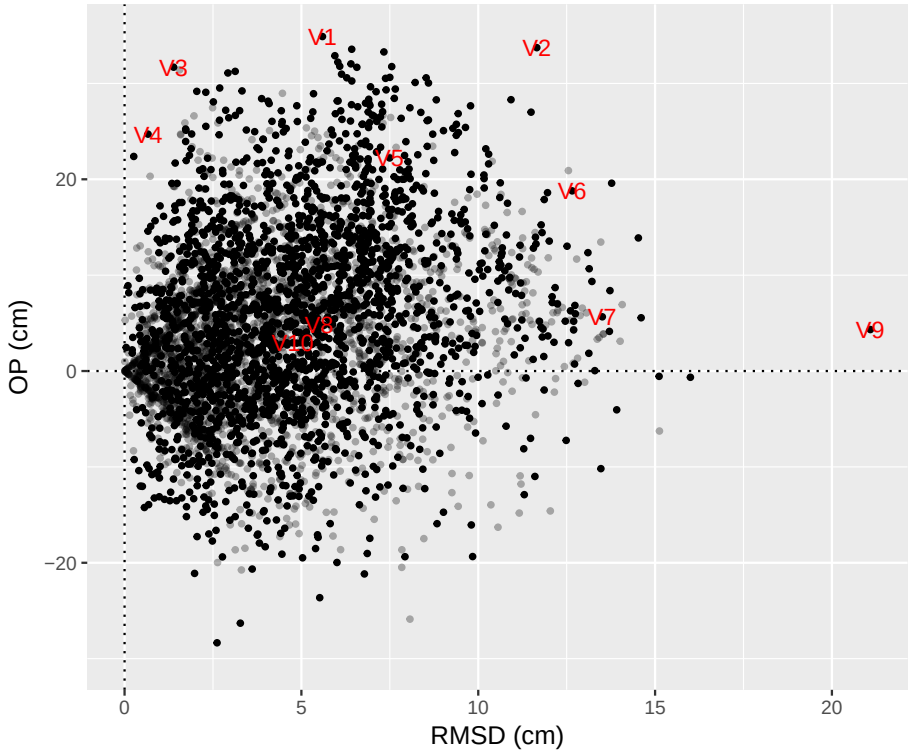

Fig. 3 Motivating example: overall performance (OP) vs stability measure (RMSD) for DBH for all 4608 varieties under consideration for selection as parents. Varieties V1-V10 labelled. Lighter coloured points correspond to varieties with an accuracy for OP of less than 0.8.

4 Results and discussion

4.1 Application of FAST to motivating example

Recall from Section 1 that the aim of the analysis of the motivating example is to obtain EBVs to enable the breeding programme to select varieties to use as parents and to provide ratings for use by industry. In the RPBC breeding programme the emphasis on selection for the trait of DBH is on high EBVs across a wide range of environments. Thus OP and RMSD are two main drivers in the selection process. They are easily jointly assessed using an $x - y$ plot for all the varieties under consideration for selection (see Figure 3). Note that the selection process can be further enhanced by considering the accuracy of individual OP values. To this end, the points in Figure 3 have been shaded lighter if the OP accuracy is less than 0.8.

An examination of Figure 3 allows easy and quick identification of varieties of interest. For example, variety V1 has the highest OP (34.9cm) and an average level of stability (RMSD of 5.6cm). We note that this is an existing parent and although it has the highest OP for DBH it is not widely deployed as it has an issue with spiral grain. Varieties V2 and V3 are two potential new parents (forward selections). Variety V2 has the second highest OP (33.7cm) but is relatively unstable (RMSD

of 11.7cm), whereas V3 has a slightly lower OP (31.7cm) and is very stable (RMSD of 1.4cm). It is noteworthy that V2 and V3 share a common parent which is V1.

Once varieties of interest have been identified on Figure 3, we then recommend examining the EBVs (EBLUPs of common VE effects) for these varieties in greater detail using first latent regression plots. These plots show the full set of EBVs for a variety and lend visual support to the OP and RMSD values. Varietal comparisons are aided by drawing pairs of varieties (often a test and control variety) on the same graph. Figure 4 contains the latent regression plots for the ten varieties labelled on Figure 3, with variety V1 appearing on every panel for comparison. Varieties V2 and V3 were chosen to be graphed as they are potential new parents. The remaining varieties are existing parents and V5, V6, V8 and V10 were chosen as they are the four most widely deployed varieties; the other varieties were chosen because they show extremes in terms of stability and responsiveness. Note that the panels are ordered on OP, with the last panel corresponding to V10 which has the lowest OP of the ten selected varieties. It is clear from Figure 4 that the OP of varieties V2 and V3 is very similar to that of V1 since the regression lines are almost co-incident (and thence the fitted values at the mean of the loadings are very similar). The regression lines for varieties V7 - V10 reflect fitted values that are consistently lower than for V1, with substantial differences at the mean of the loadings and hence much lower OPs compared with V1. In terms of stability, varieties V9 and V7 have a large scatter of points about their regression lines, and hence have large RMSD, whereas the points for V4 and V3 lie very close to the lines, so they have small RMSD (also see Figure 3). As with Figure 3, interpretation should take into account major variations in accuracy. Hence individual EBVs on Figure 4 have been shaded lighter if their accuracy is less than 0.8. We note then, for example, the accuracy of many of the EBVs for V2 and V3 are much lower than those for the existing parents due to the relatively limited testing of V2 and V3.

In addition to examining overall variety stability using RMSD, it may be of interest to consider the individual responsiveness measures. Figures 5 and 6 plot OP against responsiveness for the second and third factors. These measures may be particularly useful if the associated factors represent meaningful environmental characteristics that can be exploited for specific adaptation. A method that is often proposed to assess this is to compute correlations between individual rotated vectors of estimated loadings and measured environmental covariates. In our experience this is rarely successful since the factors typically reflect complex combinations of environmental stresses. A potentially more fruitful approach involves a reversal of the focus in the sense of using variety rather than environment information. In this strategy, reference or probe varieties which are known to have differential performance in the presence of certain environmental conditions, are used to characterise the environments in the MET (Mathews et al, 2011). This has been successful for domesticated crops such as wheat, but is currently of limited use for *Pinus radiata* since it has a more complex genome and has only undergone one or two selection cycles. However, for pedagogical reasons, in the following we discuss the method in the context of the motivating example.

Figure 5 shows that V9 has a large positive response to the second factor. Thus V9 had relatively better EBVs for environments with large positive loadings for this factor (eg. E83, E5, E61 as shown in the Appendix) compared with environments with large negative loadings (eg. E65, E64, E17). In contrast, V7 exhibits

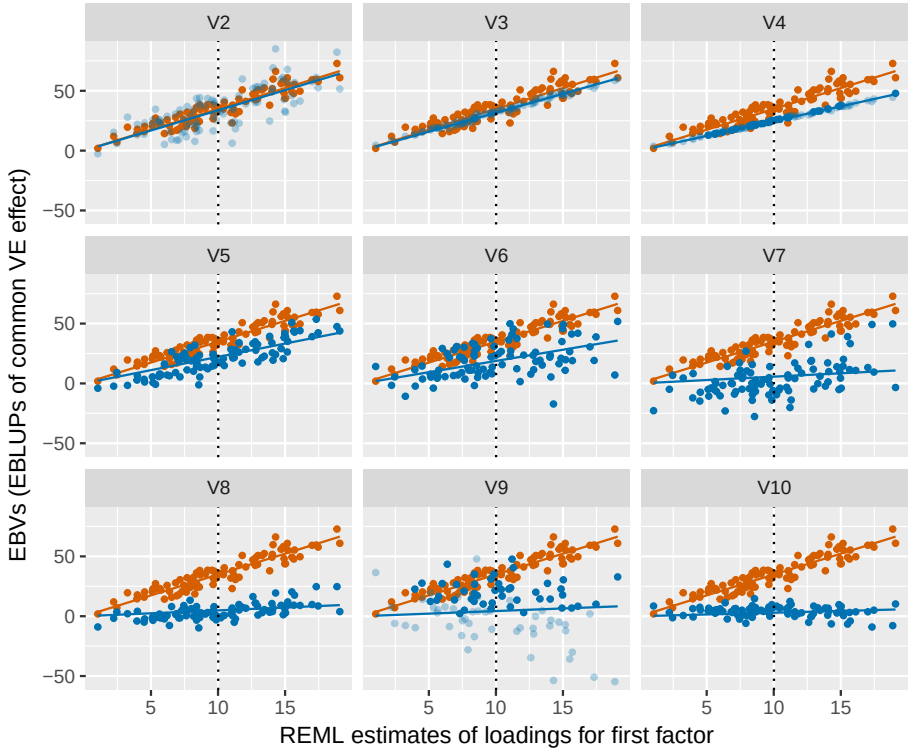

Fig. 4 Motivating example: first latent regression plots for DBH for 10 varieties. Panels correspond to different varieties (V2-V10, coloured blue) and variety V1 is shown on every panel (coloured orange). Lighter coloured points have an accuracy of less than 0.8. The vertical dotted line corresponds to the mean value of the estimated loadings for the first factor.

the opposite behaviour. Additionally, neither V7 nor V9 show much response to the third factor (see Figure 6). Thus the genetic architecture of V7 compared with V9 could be used to characterise the environments with extreme positive and negative loadings for the second factor. We note that these are very old varieties and have only average OP for DBH so their responsiveness is of interest mainly for characterising environments and possibly for the purpose of maintaining genetic diversity. In contrast we consider variety V2 which has a high OP and also a large negative response to the third factor (see Figure 6) which means it had relatively better EBVs for environments with large negative loadings for this factor (eg. E17, E83, E79 as shown in the Appendix) compared with environments with large positive loadings (eg. E21, E76, E58). Thus the previously mentioned high RMSD for this variety may represent exploitable variation rather than undesirable instability. It will be particularly interesting to track this variety as more data are collected since it has the potential for both high OP and boosted performance under specific environmental conditions.

As previously mentioned, the interpretation of individual responsiveness measures in this example is problematic so that the focus in terms of stability is RMSD. In the RPBC breeding program it has been suggested that both OP and RMSD

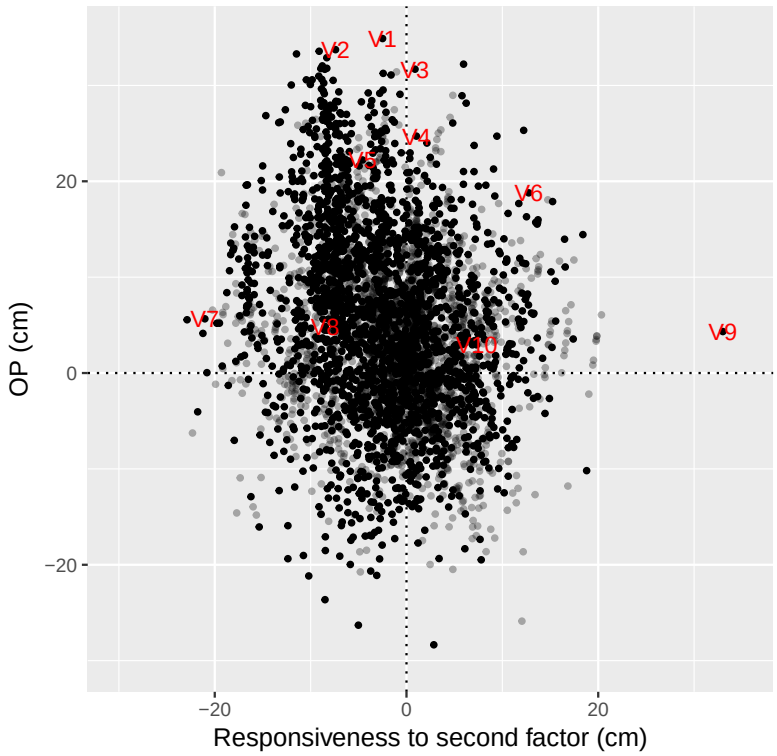

Fig. 5 Motivating example: overall performance (OP) vs responsiveness to second factor for DBH for all 4608 varieties under consideration for selection as parents. Varieties V1-V10 labelled. Lighter coloured points correspond to varieties with an accuracy for OP of less than 0.8.

represent characteristics with economic importance. The fact that they are on the same scale, namely the scale of the trait being analysed, should aid in assigning economic weights. Given the weights, it will be straight-forward to combine OP and RMSD measures across key traits (including DBH) to form a simple, concise selection index.

In terms of information for industry, a certificate is issued for each seedlot purchased by forest owners and provides a rating for individual traits, including DBH. The rating is based on the genetic quality of the parents and their proportion in the seedlot. Currently the genetic quality of a parent is obtained in a piecemeal manner using the results of the MET analysis. In the first step, the estimated additive genetic correlations between trials are clustered in order to identify trials that lack correlation with others. EBVs for an individual parent are then averaged across all remaining trials to provide a single value which is then converted to a rating. With the advent of FAST as described in this paper, there is potential for a more coherent, objective and informative scheme for industry. The natural choice for a single measure of the genetic quality of a parent is OP. This avoids the need to ignore trials, which is a some-what subjective procedure, and also avoids the loss of information incurred by converting to a rating. In addition to OP, it

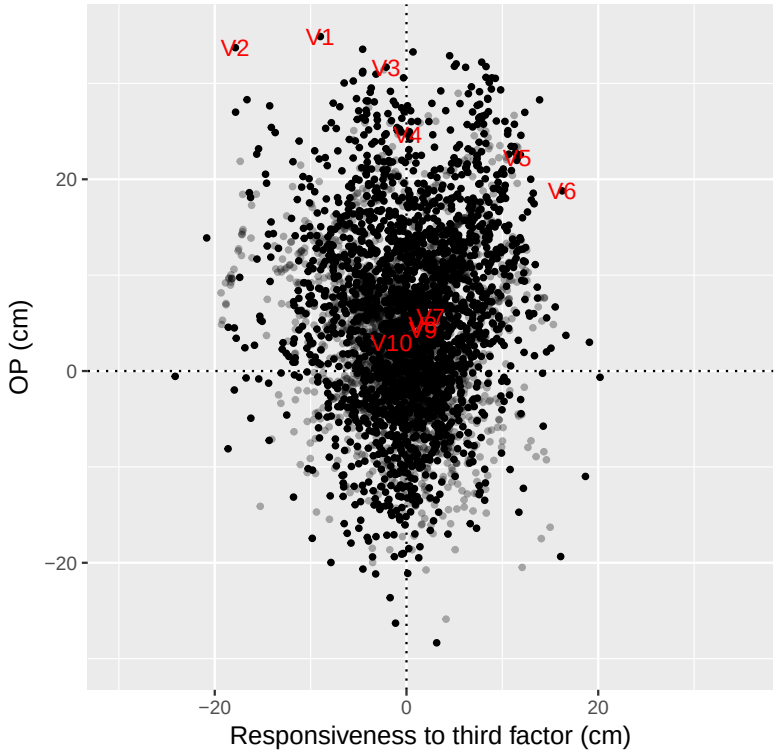

Fig. 6 Motivating example: overall performance (OP) vs responsiveness to third factor for DBH for all 4608 varieties under consideration for selection as parents. Varieties V1-V10 labelled. Lighter coloured points correspond to varieties with an accuracy for OP of less than 0.8.

may be important to consider stability, so that RMSD may also be reported to industry. We note that it is straight-forward to compute OP, RMSD and associated measures of accuracy on a seedlot basis given the specific mixture of parents. The challenge remains as to how this information may be disemminated and thence adopted by industry.

4.2 General applicability of FAST

The development and application of FAST thus far has assumed that all of the rotated estimated loadings for the first factor are positive. Here we discuss implications and departures from this assumption. First, it is instructive to make the comparison between OP and the more traditional concept of a variety main effect. The latter is typically obtained by fitting a linear mixed model that partitions VE effects into variety main effects and VEI. Thus the additive VE effects are given by

$$\mathbf{u_a} = (\mathbf{1}_p \otimes \mathbf{I}_m)\,\mathbf{u_v} + \mathbf{u_{ve}} \tag{14}$$

where $\mathbf{1}_p$ is the $p-$vector with all values equal to unity, $\mathbf{u_v}$ is the $m-$vector of additive variety main effects (which has associated variance $\sigma_v^2\mathbf{A}$) and $\mathbf{u_{ve}}$ is the $mp-$vector of additive VEI effects (which has associated variance $\sigma_{ve}^2\mathbf{I}_p \otimes \mathbf{A}$). We note that the model in equation (14) can be re-written in the form of a factor analytic model, namely

$$\mathbf{u_a} = (\sigma_v \mathbf{1}_p \otimes \mathbf{I}_m)\, \mathbf{f}_{\mathbf{a}_1} + \boldsymbol{\delta}_\mathbf{a}$$

where $\mathbf{f}_{\mathbf{a}_1} = \mathbf{u_v}/\sigma_v$ and $\boldsymbol{\delta}_\mathbf{a} = \mathbf{u_{ve}}$. This is a special case of the FA1 model in which all the loadings are equal, with $\lambda_{a_{1j}} = \sigma_v$, and all the specific variances are equal, with $\psi_{a_j} = \sigma_{ve}^2$. Fitting this model would therefore result in estimated loadings for the first (and only) factor that are equal and non-negative (given by the square root of the REML estimate of the variety main effect variance). Thus if FAST was applied to this model the OP for a variety would be identical to the EBLUP of the variety main effect.

Experience has shown that the general FAk_a model provides a far superior fit to most MET data-sets and hence our development of FAST in this paper. If all rotated estimated loadings for the first factor are positive, then the first factor represents a generalised version of the main effect part of the model in equation (14) in which the loadings are positive but no longer equal. OP for a variety may then be thought of as a "generalised main effect" which allows for heterogeneity of scale between environments.

Another key link with the model in equation (14) is in terms of the between environment additive genetic covariance structure. The model in equation (14) leads to a genetic variance matrix, $\mathbf{G_a}$, in which all diagonal elements are given by $\sigma_v^2 + \sigma_{ve}^2$ and all off-diagonal elements by σ_v^2. Thus the variety main effect variance can also be interpreted as the (common) genetic covariance between every pair of environments. In terms of the FAk_a model, the use of the principal component rotation (see section 3.1) means that the first rotated factor accounts for the maximum amount of genetic covariance in the data. The estimate of this source of covariance for two environments j and h is given by the product of the corresponding two estimated loadings for the first factor, namely $\hat{\lambda}^*_{a_{1j}}\hat{\lambda}^*_{a_{1h}}$. Thus if all the rotated estimated loadings for the first factor are positive, there is a dominant component of positive genetic covariance between all pairs of environments. Once again this is a generalisation of the main effect scenario in the sense that heterogeneity is accommodated.

In our experience in analysing MET data, the vast majority of results are characterised by rotated estimated loadings for the first factor that are either all positive or include a few small negative values. In terms of the analyses of the five key traits for RPBC, four traits, including DBH as presented in this paper, had all positive loadings in the first factor and the remaining trait had a single negative value. In the most recent annual analyses of grain yield data from four Australian pulse breeding programs, three out of eight analyses had all positive loadings in the first factor and the remainder had a few small negative values.

The presence of small negative estimated loadings in the first factor indicates that the first (and therefore dominant) factor contains cross-over VEI, but of such a small magnitude to be of no practical importance. We illustrate this graphically using the results of the eight pulse breeding analyses. Figure 7 contains schematic representations of first latent regression plots for these analyses. In each panel,

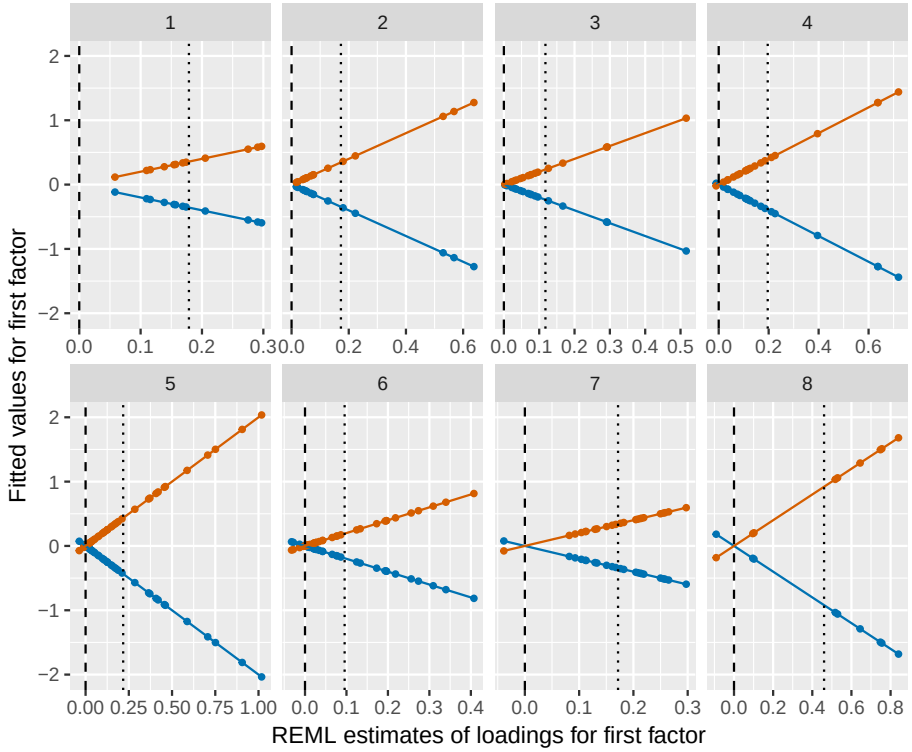

Fig. 7 Schematic first latent regression plots from the analyses of eight pulse breeding datasets (panels labelled 1-8). Two extreme varieties chosen for each and labelled A (coloured orange) and B (coloured blue). Slopes of the solid lines are given by the EBLUPs of the (rotated) variety scores for the first factor. Points along solid lines are fitted values for first factor at individual environment loadings. The dashed vertical line is positioned at zero so environments with negative loadings are to the left of this line. The overall performance measure for each variety is the fitted value at the mean value of the estimated loadings (vertical dotted line).

two extreme varieties (A and B) are graphed to show the maximum cross-over VEI represented in the first factor. Only the fitted values (and not the EBLUPs of the common VE effects) are shown on the panels and they are plotted both as the regression lines and also the underlying points (to show the position of the environment loadings). Due to the presence of negative loadings in analyses 4-8, Figure 7 shows that this causes the lines to intersect, which translates to cross-over VEI. However, the magnitude of the cross-over is very small, so that the superiority of B over A is probably not biologically or economically important and the dominant feature is the superiority of A over B, both in terms of frequency of environments and magnitude of differences. This is captured in OP which is higher for A than B (see fitted values at mean loadings in Figure 7). Thus in cases where there are a few small negative loadings in the first rotated factor, OP and RMSD are still meaningful and we would proceed with FAST as described in this paper.

As the proportion and magnitude of the negative loadings in the first factor increases, there reaches a point where OP and RMSD are no longer meaningful.

For example, if roughly half of the loadings in the first factor are negative, then the dominant feature in the VE effects is cross-over VEI so that an overall performance measure is inappropriate. The analogy with the main effect model in equation (14) is a reduction in the main effect variance estimate to a point where it is zero and thence the variety main effects effectively "do not exist", that is, the associated EBLUPs are all zero. This is a very rare scenario and requires a different approach for selection. Within the framework of FAST, one possibility is to compute a responsiveness to the first factor in addition to the remaining factors. Then graphs of all pairs of responsiveness measures, that is, of the form given in Figures 5 and 6, could be used to aid with selection decisions. The magnitude of VEI dictates that selection must be for specific rather than broad adaptation. It is key to note that FAST not only identifies that this is the case, but also offers a way forward for informed selection decisions, particularly if an interpretation can be ascribed to the loadings.

Conflict of interest The authors declare that they have no conflict of interest.

Acknowledgements The authors thank RPBC for use of their data. We thank the referees whose comments have led to an improved manuscript.

Appendix

Rotated REML estimates of loadings for individual environments (E1-E92) from FA3 model fitted to additive VE effects in motivating example.

Env	Load 1	Load 2	Load 3	Env	Load 1	Load 2	Load 3
E1	13.0	-1.8	2.2	E47	12.8	-4.8	3.2
E2	10.5	3.9	-4.9	E48	14.0	-4.2	0.6
E3	11.8	-3.9	-8.4	E49	14.1	2.6	-3.3
E4	14.7	-3.1	-3.8	E50	6.2	1.8	-4.5
E5	11.2	12.5	2.0	E51	17.0	-1.6	0.4
E6	2.2	2.2	-4.7	E52	13.3	2.9	-5.1
E7	1.0	10.7	-1.0	E53	9.2	0.2	-6.2
E8	11.1	5.6	-0.6	E54	8.6	4.2	-9.0
E9	7.6	4.9	5.1	E55	10.4	5.3	0.1
E10	4.5	7.7	-2.2	E56	16.1	0.0	6.3
E11	5.6	3.4	-3.8	E57	14.9	-0.1	3.9
E12	8.2	3.5	-4.8	E58	13.9	5.8	8.6
E13	7.1	6.6	4.8	E59	7.4	-5.8	1.5
E14	5.1	3.2	-3.3	E60	5.7	4.3	-2.9
E15	9.1	2.1	-4.2	E61	6.4	12.1	-1.6
E16	6.9	2.8	6.2	E62	7.2	4.2	-6.5
E17	14.3	-16.9	-11.8	E63	6.5	-3.9	7.8
E18	4.0	6.3	-2.3	E64	17.3	-17.6	5.2
E19	5.6	0.3	-3.6	E65	18.9	-18.4	-2.3
E20	7.7	-1.3	-4.9	E66	12.1	-0.7	1.0
E21	11.1	1.6	14.7	E67	10.2	6.8	-1.7
E22	13.5	0.0	-5.4	E68	8.4	9.1	2.7
E23	11.6	2.1	7.1	E69	4.5	3.9	-5.3
E24	11.3	10.2	6.0	E70	12.8	-5.9	-2.7
E25	9.6	0.5	1.7	E71	15.2	-5.5	-0.9
E26	15.6	-0.4	6.7	E72	13.0	4.7	-6.6
E27	15.0	6.7	6.7	E73	7.9	-9.6	5.9
E28	19.1	7.0	3.7	E74	2.5	-2.2	2.0
E29	9.9	9.1	0.3	E75	7.3	-0.6	5.0
E30	8.6	3.7	4.6	E76	8.3	-0.6	10.1
E31	6.8	1.2	-4.2	E77	7.6	0.4	2.3
E32	8.4	6.7	1.5	E78	6.0	7.0	4.8
E33	12.6	-11.8	-0.2	E79	14.1	-2.2	-9.8
E34	15.5	-13.0	7.9	E80	15.2	-3.6	-6.9
E35	9.7	2.3	-4.1	E81	17.5	0.6	2.9
E36	5.3	1.9	-8.9	E82	9.9	5.2	-1.3
E37	4.0	-3.2	-3.0	E83	8.6	13.6	-10.5
E38	15.0	2.9	7.5	E84	8.4	4.5	0.6
E39	11.1	0.9	0.5	E85	14.9	5.3	-7.5
E40	8.0	-2.8	-0.2	E86	15.7	-11.0	2.1
E41	8.4	2.4	2.7	E87	11.6	-5.1	-1.5
E42	8.7	0.4	-4.3	E88	9.6	0.4	-5.1
E43	5.7	1.0	-3.2	E89	5.2	5.4	-6.3
E44	9.6	7.6	5.0	E90	3.3	-2.3	-7.5
E45	8.4	-5.8	-5.4	E91	11.0	5.0	4.9
E46	7.5	-2.4	5.6	E92	9.7	-4.3	-2.5

References

Beeck C, 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

Butler DG, Cullis BR, Gilmour AR, Gogel BJ (2009) Mixed models for S language environments, ASReml-R reference manual. Training and development series, No QE02001, QLD Department of Primary Industries and Fisheries, Brisbane, QLD.

Cullis BR, Smith AB, Beeck C, Cowling WA (2010) Analysis of yield and oil from a series of canola breeding trials. Part II: Exploring VxE using factor analysis. Genome 53:1002–1016

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

Finlay KW, Wilkinson GN (1963) The analysis of adaptation in a plant breeding programme. Australian Journal of Agricultural Research 14:742–754

Gauch J HG (1992) Statistical analysis of regional yield trials: AMMI analysis of factorial designs. Elsevier, Amsterdam

GF Plus (2006) Radiata Pine Breeding Company Ltd., New Zealand, URL www.rpbc.co.nz/gfscheme.htm

Gogel BJ, Smith AB, Cullis BR (2018) Comparison of a one- and two-stage mixed model analysis of Australia's National Variety Trial Southern Region wheat data. Euphytica URL https://doi.org/10.1007/s10681-018-2116-4

Henderson C (1950) Estimation of genetic parameters (abstract). Annals of Mathematical Statistics 21:309–310

Kelly A, Smith A, Eccleston J, Cullis B (2007) The accuracy of varietal selection using factor analytic models for multi-environment plant breeding trials. Crop Science 47:1063–1070

Kempton RA (1984) The use of biplots in interpreting variety by environment interactions. Journal of Agricultural Science, Cambridge 103:123–135

Lin C, Binns M, Leftkovitch L (1986) Stability analysis: where do we stand? Crop Science 26:894–900

Mathews KL, Trethowan R, Milgate AW, Payne T, van Ginkel M, Crossa J, DeLacy I, Cooper M, Chapman S (2011) Indirect selection using reference and probe genotype performance in multi-environment trials. Crop and Pasture Science 62:313–327

Nelder JA (1994) The statistics of linear models: back to basics. Statistics and Computing 4:221–234

Oakey H, Verbyla A, Cullis B, Wei X, Pitchford W (2007) Joint modelling of additive and non-additive (genetic line) effects in multi-environment trials. Theoretical and Applied Genetics 114:1319–1332

Shukla GK (1972) Some statistical aspects of partitioning genotype-environmental components of variability. Heredity 29:237–245

Smith A, Ganesalingam A, Kuchel H, Cullis B (2015) Factor analytic mixed models for the provision of grower information from national crop variety testing programmes. Theoretical and Applied Genetics 128:55–72

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

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, Cambridge 143:449–462

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 45:445–460

Welham S, Gogel B, Smith A, Thompson R, Cullis B (2010) A comparison of analysis methods for late-stage variety evaluation trials. Australian and New Zealand Journal of Statistics 52:125–149

Yates F, Cochran WG (1938) The analysis of groups of experiments. Journal of Experimental Science, Cambridge 28:556–580