Variety selection using interaction classes derived from factor analytic linear mixed models in a single step multi-environment trial analysis with information on genetic relatedness

Lu Wang[1]

Alison Smith[1] Brian Cullis[1] & Adam Norman[2]

Biometrics 2023

[1]Mixed Models and Experimental Design Lab (MMaED)
National Institute for Applied Statistics Research Australia (NIASRA)
University of Wollongong
luw@uow.edu.au

[2]Australian Grain Technologies
adam.norman@agtbreeding.com.au

Outline

I. Motivations

II. What do we do and why

III. How do we do it

IV. What we have achieved

Motivations

Motivations

- Australian Grain Technologies (AGT) is Australia's largest plant breeding company, and the market leader in wheat genetics.

- Their work is driven by developing new field crop varieties that are more productive, better quality and cost less to grow.
- Millions of potential new crop varieties tested each year; but only a very special few make it to release.
- As a plant breeding research company, AGT focuses on innovation, and is a fast adopter of new technologies and statistical methods.

Motivations

Plant breeding objectives

- **Yield** is a key trait that is routinely examined in plant breeding programs via field testing.
  - possesses a complex genetic architecture and often exhibits low heritability.
  - being influenced by many sources of non-genetic variation.
  - possesses large variety by environment interaction (VEI), representing the differential performance of varieties in response to a change in environment[†].

[†] an "environment" is defined to be the combination of a geographic location and year of the trial(s) present in the data-set.

Motivations

Plant breeding objectives

- **Yield** is a key trait that is routinely examined in plant breeding programs via field testing.
  - possesses a complex genetic architecture and often exhibits low heritability.
  - being influenced by many sources of non-genetic variation.
  - possesses large variety by environment interaction (*VEI*), representing the differential performance of varieties in response to a change in environment[†].
- Lines enter the yield testing phase at stage 1 (S1) and progress through to later stages towards eventual release as a commercial cultivar after approximately 8 years.

[†] an "environment" is defined to be the combination of a geographic location and year of the trial(s) present in the data-set.

Motivations

Plant breeding objectives

- **Yield** is a key trait that is routinely examined in plant breeding programs via field testing.
  - possesses a complex genetic architecture and often exhibits low heritability.
  - being influenced by many sources of non-genetic variation.
  - possesses large variety by environment interaction (VEI), representing the differential performance of varieties in response to a change in environment[†].
- Lines enter the yield testing phase at stage 1 (S1) and progress through to later stages towards eventual release as a commercial cultivar after approximately 8 years.
- Aim: at each stage of testing, accurately select the best lines to progress to the next stage.

[†] an "environment" is defined to be the combination of a geographic location and year of the trial(s) present in the data-set.

What do we do and why

What do we do and why

Aim: making selections with the presence of *VEI*

To address all these challenges and improve selection accuracy (Smith et al., 2021), we:

- generate multi-environment trial (MET) data.
  - a series of designed selection experiments conducted across a range of targeted geographic locations and typically over several years.

What do we do and why

Aim: making selections with the presence of *VEI*

To address all these challenges and improve selection accuracy (Smith et al., 2021), we:

- generate multi-environment trial (MET) data.
  - a series of designed selection experiments conducted across a range of targeted geographic locations and typically over several years.
- adopt a fully efficient one stage factor analytic linear mixed model (FALMM) analysis approach:
  - allows a separate FA variance structure for the variety effects in individual environments (VE effects) (see Smith et al. (2005), Smith et al. (2021) and Gogel et al. (2018), for example).
  - has the ability to process imbalanced data and
  - to incorporate genetic relatedness through ancestral (pedigree) information or genomic (marker) data.
  - appropriate modelling of all sources of variation, including spatial correlations.
  - provides more accurate predictions of variety effects across environments.

How do we do it

Motivating example

Task: Selecting the best subset of 20-30 lines from 144 lines tested in stage 3 (S3), 2022.

- In 2022, these 144 lines were only tested in 12 trials.

Motivating example

Task: Selecting the best subset of 20-30 lines from 144 lines tested in stage 3 (S3), 2022.

- In 2022, these 144 lines were only tested in 12 trials.
- For the purpose of this selection, the MET data-set comprised:
  - a series of 103 trials over the period 2019-2022, which covered the full selection history of these 144 S3 lines.
  - 38 environments across 14 locations in South Australia, Western Australia and Victoria.
  - a total of 9399 varieties and 40,514 plot yields.
  - Pedigree information was available for all varieties, with 11,786 records including 2,387 for parents.
  - The numerator relationship matrix (NRM) was formed using the *pedicure* package of Butler (2019) in *R* (R Core Team, 2022).
  - The variety by environment "fill-in"[†] is 8.5%.

[†] the number of V×E combinations present in the data expressed as a percentage of the number of possible V×E combinations.

Motivating example

Trial layouts

- In our context, a field trial is a physical block of plots onto which a valid experimental design (with replication and randomization) is imposed.
  - 82 trials were partially replicated (Cullis et al., 2006), in which a number of varieties were tested on a single plot each without replication; 21 trials were fully replicated with 2 replicates for each variety.

Motivating example

Trial layouts

- In our context, a field trial is a physical block of plots onto which a valid experimental design (with replication and randomization) is imposed.
  - 82 trials were partially replicated (Cullis et al., 2006), in which a number of varieties were tested on a single plot each without replication; 21 trials were fully replicated with 2 replicates for each variety.
- Each trial comprised a two-dimensional arrangement of plots indexed by rows and columns. Blocking was employed across all trials in the column direction.
  - The smallest trial comprised 192 plots arranged as 8 rows by 24 columns; the largest comprised 768 plots arranged as 32 rows by 24 columns. The number of varieties per trial ranged from 144 to 519.

Motivating example

Trial layouts

- In our context, a field trial is a physical block of plots onto which a valid experimental design (with replication and randomization) is imposed.
  - 82 trials were partially replicated (Cullis et al., 2006), in which a number of varieties were tested on a single plot each without replication; 21 trials were fully replicated with 2 replicates for each variety.
- Each trial comprised a two-dimensional arrangement of plots indexed by rows and columns. Blocking was employed across all trials in the column direction.
  - The smallest trial comprised 192 plots arranged as 8 rows by 24 columns; the largest comprised 768 plots arranged as 32 rows by 24 columns. The number of varieties per trial ranged from 144 to 519.
- Each environment involved either a single field trial or multiple trials, called co-located trials (Smith et al., 2021). These arose due to management challenges and the conduct of trials of different stages.
  - 17 environments had co-located trials and 21 had a single trial in each environment.

Single step MET analysis using FALMM in *DWReml*

Fator analytic models for VE effects

- The use of pedigree information allows for the partition of VE effects into additive and non-additive effects. Therefore,

$$\boldsymbol{u_g} = \boldsymbol{u_a} + \boldsymbol{u_e}.$$

- Following Smith et al. (2023 - *In-prep*), separate FA models were fitted for each set of these effects.
- Each set of these effects can be partitioned into *common VE* (CVE) effects and the lack of fit effects, also known as *specific VE* (SVE) effects.

$$\begin{aligned} \boldsymbol{u_a} &= (\boldsymbol{\Lambda_a} \otimes \boldsymbol{I}_m)\boldsymbol{f_a} + \boldsymbol{\delta_a} = \boldsymbol{\beta_a} + \boldsymbol{\delta_a} \\ \boldsymbol{u_e} &= (\boldsymbol{\Lambda_e} \otimes \boldsymbol{I}_m)\boldsymbol{f_e} + \boldsymbol{\delta_e} = \boldsymbol{\beta_e} + \boldsymbol{\delta_e}. \end{aligned}$$

- In FA models, it is assumed that

$$\text{var}\begin{pmatrix} \boldsymbol{f_a} \\ \boldsymbol{f_e} \end{pmatrix} = \begin{bmatrix} \boldsymbol{D_a} \otimes \boldsymbol{A} & 0 \\ 0 & \boldsymbol{D_e} \otimes \boldsymbol{I}_m \end{bmatrix} \qquad \text{var}\begin{pmatrix} \boldsymbol{\delta_a} \\ \boldsymbol{\delta_e} \end{pmatrix} = \begin{bmatrix} \boldsymbol{\Psi_a} \otimes \boldsymbol{A} & 0 \\ 0 & \boldsymbol{\Psi_e} \otimes \boldsymbol{I}_m \end{bmatrix}$$

Single step MET analysis using FALMM in *DWReml*

FA model syntax in *DWReml*

Therefore, the variance matrices for the CVE effects are given by

$$\text{var}\begin{pmatrix}\boldsymbol{\beta_a}\\ \boldsymbol{\beta_e}\end{pmatrix} = \begin{bmatrix}\boldsymbol{\Lambda_a D_a \Lambda_a^\intercal} \otimes \boldsymbol{A} & 0\\ 0 & \boldsymbol{\Lambda_e D_e \Lambda_e^\intercal} \otimes \boldsymbol{I}_m\end{bmatrix}$$

```
fa1fa1.dwr <- dwreml(yield ~ Env,
    random = ~ rr(Env,1):vm(Variety, A.inv) + cut(~vm(Variety, A.inv)|Env, rds="Acut") +
               rr(Env,1):ide(Variety, A.inv) + cut(~ide(Variety, A.inv)|Env, rds="Acut") +
               at(Env, env.coloc):Covblk + at(Env, env.coloc):Covblk:ColRep +
               at(Env, env.coloc):Covblk:Column + at(Env, env.coloc):Covblk:Row +
               at(Env, env.single):ColRep +
               at(Env, env.single):Column + at(Env, env.single):Row,
    data = df, na.action = na.method(x='include'),
    residual = ~ dsum(~ar1(Column):ar1(Row) | Covblk | Env))
```

Single step MET analysis using FALMM in *DWReml*

FA model syntax in *DWReml*

		Variance Models	Terms fitted in *DWReml*
Genetic effects	$\boldsymbol{\beta_a}$	$\boldsymbol{\Lambda_a D_a \Lambda_a^\intercal} \otimes \boldsymbol{A}$	`rr(Env,1):vm(Variety, A.inv)`
	$\boldsymbol{\delta_a}$	$\oplus_{j=1}^{38} \psi_{\boldsymbol{a}_j} \boldsymbol{A}_j$	`cut(~vm(Variety, A.inv)|Env, rds="Acut")`
	$\boldsymbol{\beta_e}$	$\boldsymbol{\Lambda_e D_e \Lambda_e^\intercal} \otimes \boldsymbol{I}_m$	`rr(Env,1):ide(Variety, A.inv)`
	$\boldsymbol{\delta_e}$	$\oplus_{j=1}^{38} \psi_{\boldsymbol{e}_j} \boldsymbol{I}_{m_j}$	`cut(~ide(Variety, A.inv)|Env, rds="Acut")`

Peripheral effects		$\oplus_{j=1}^{38} \boldsymbol{G}_{\boldsymbol{p}_j}$	`at(Env, env.single):ColRep`
			`at(Env, env.single):Column`
			`at(Env, env.single):Row`

			`at(Env, env.coloc):Covblk`
			`at(Env, env.coloc):Covblk:ColRep`
			`at(Env, env.coloc):Covblk:Column`
			`at(Env, env.coloc):Covblk:Row`

Residuals		$\oplus_{j=1}^{38} \boldsymbol{R}_{\boldsymbol{e}_j}$	`dsum(~ar1(Column):ar1(Row) | Covblk | Env)`

What we have achieved

Results and interaction classes

Table 1: Summary of model fits. Likelihood ratio chi-square; Akaike information criterion (AIC); percentage of genetic variance accounted for by k_a additive factors; by $k_e = 1$ non-additive factor and by all $k = k_a + k_e$ factors.

Models	LR chi-square	df	AIC	Genetic variance accounted for (%)		
				Additive	Non-additive	Total
DIAG (Ind.)	0[†]	-	14958	-	-	-
DIAG-DIAG	11326	38	3708	80.6	19.4	-
FA1FA1	3860	76	0[‡]	60.4	59.8	60.3
FA2FA1	246	37	-172	78.8	41.3	74.5
FA3FA1	142	36	-243	82.8	43.1	78.3
FA4FA1	115	35	-287	87.0	43.7	82.2

[†] log-likelihood for DIAG (Ind.) is 23352.

[‡] AIC for FA1FA1 is -61152.94, used as the subtrahend for differences.

Results and interaction classes

Task: Selecting the best subset of 20-30 lines from 144 stage 3 (S3) test lines in 2022.

- Smith et al. (2021) addressed the issue of summarising variety performance in the presence of interaction by using "interaction classes" (iClasses).
  - The factor loadings represent the latent environmental covariates that are driving the *VEI*.
  - The estimated loadings for factor r of environment j can only be positive ("p") or negative ("n").
  - Groups of environments formed on the basis of the signs of their estimated loadings in individual factors discriminate varieties with differential patterns of *VEI*.
  - The maximum number of possible iClasses is 2^k.

Results and interaction classes

Task: Selecting the best subset of 20-30 lines from 144 stage 3 (S3) test lines in 2022.

- Smith et al. (2021) addressed the issue of summarising variety performance in the presence of interaction by using "interaction classes" (iClasses).
  - The factor loadings represent the latent environmental covariates that are driving the *VEI*.
  - The estimated loadings for factor r of environment j can only be positive ("p") or negative ("n").
  - Groups of environments formed on the basis of the signs of their estimated loadings in individual factors discriminate varieties with differential patterns of *VEI*.
  - The maximum number of possible iClasses is 2^k.

- Following Smith et al. (2021), iClasses were formed for each FA model by
  - firstly, ordering the factors by the percentage variance accounted for;
  - then pasting the values ("p" or "n") of the rotated REML estimates of the loadings of each factor.

Results and interaction classes

Table 2: Rotated REML estimates of loadings ($\times 1000$) for each factor in FA4FA1 model; iClasses based on all five factors (iClass5) and only the first four factors (iClass4). Numbers within brackets show the percentage of genetic variance accounted for by the individual factors.

	load:add1 (51%)	load:add2 (18.8%)	load:ide1 (4.9%)	load:add3 (4.1%)	load:add4 (3.4%)	iClass5	iClass4
Env26	82	-93	-4	-235	-45	pnnnn	pnnn
Env1	85	-18	-162	146	-183	pnnpn	pnnp
Env9	151	-65	279	-73	156	pnpnp	pnpn
Env5	53	-69	67	97	-236	pnppn	pnpp
Env28	101	32	-32	-205	79	ppnnp	ppnn
Env6	236	202	-135	27	53	ppnpp	ppnp
Env14	23	168	64	-98	-181	pppnn	pppn
Env38	264	406	45	97	153	ppppp	pppp
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Results and interaction classes

Task: Selecting the best subset of 20-30 lines from 144 stage 3 (S3) test lines in 2022.

- In each iClass,
  - check the unique numbers of S3 lines present.
  - obtain the rankings of lines using their mean predicted total *CVE* effects .
- Select the best 40 (total) by proportional sampling based on iClass sizes.

Table 3: Within each iClass, number of S3 lines present; number of environments; number of S3 lines selected.

	pnnn	pnpp	pnpn	pnpp	ppnn	ppnp	pppn	pppp
number of S3 lines present	6	7	144	135	144	144	144	144
number of environments	1	3	12	10	2	4	2	4
number of S3 lines selected	0	0	14	12	2	5	2	5

Results and interaction classes

Task: Selecting the best subset of 20-30 lines from 144 stage 3 (S3) test lines in 2022.

Table 4: Numbers of unique lines selected from each model by proportional sampling across iClasses; numbers of unique lines selected in common between models.

	FA1FA1	FA2FA1	FA3FA1	FA4FA1	usable iClasses
FA1FA1	24				2
FA2FA1	19	24			4
FA3FA1	19	21	24		6
FA4FA1	19	22	22	24	6

Results and interaction classes

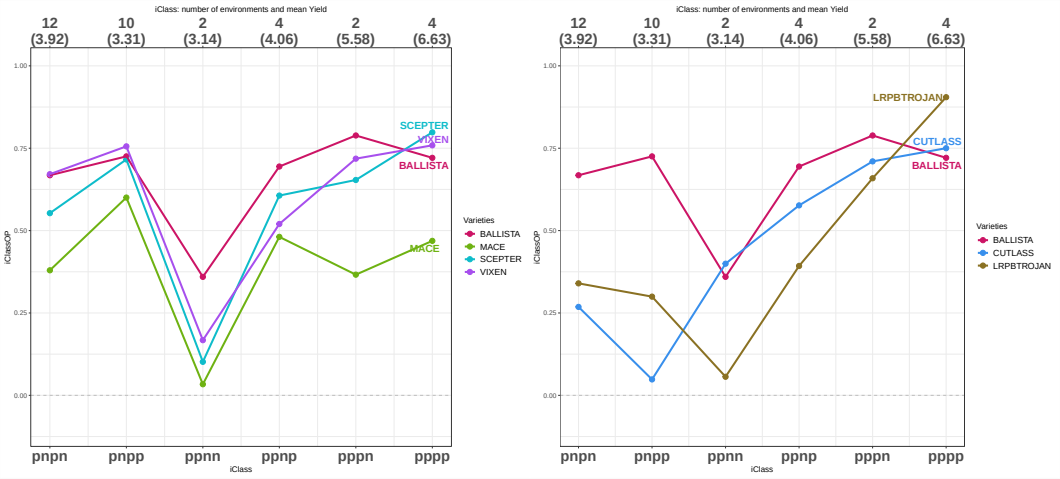

Figure 1: Interaction plot of iClassOP (t/ha) for six check varieties. The number of environments in each iClass and their associated mean yield (t/ha) is given along the top axis.

Take home messages

- Demonstrated how we implemented a fully efficient single step factor analytic linear mixed model approach in the MET analysis in *DWReml*.

Take home messages

- Demonstrated how we implemented a fully efficient single step factor analytic linear mixed model approach in the MET analysis in *DWReml*.
- iClass approach provides meaningful summaries for VE effects with the presence of *VEI*. It could be used not only to select the best varieties within each iClass but also match varieties in terms of their patterns of *VEI* across iClasses.

Take home messages

- Demonstrated how we implemented a fully efficient single step factor analytic linear mixed model approach in the MET analysis in *DWReml*.
- iClass approach provides meaningful summaries for VE effects with the presence of *VEI*. It could be used not only to select the best varieties within each iClass but also match varieties in terms of their patterns of *VEI* across iClasses.
- Innovation is driven by the purposes.

A. Smith, A. Norman, D. Butler and B. Cullis. ***Plant variety selection using interaction classes derived from Factor Analytic Linear Mixed Models: models with information on genetic relatedness****. In-prep, 2023.*

References i

David Butler. *Package 'pedicure': pedigree tools*, 2019.

B. Cullis, A. Smith, and N. Coombes. On the design of early generation variety trials with correlated data. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(4):381–393, 2006. ISSN 1085-7117. doi: 10.1198/108571106X154443.

B. Gogel, A. Smith, and B. Cullis. Comparison of a one- and two-stage mixed model analysis of australia's national variety trial southern region wheat data. *Euphytica*, 214(44), 2018.

R Core Team. R: A Language and Environment for Statistical Computing, 2022.

A. Smith, B. Cullis, and R. Thompson. The analysis of crop cultivar breeding and evaluation trials: an overview of current mixed model approaches. *Journal of Agricultural Science*, 143:449–462, 2005.

A. Smith, A. Norman, H. Kuchel, and B. Cullis. Plant variety selection using interaction classes derived from factor analytic linear mixed models: models with independent variety effects. *Frontiers in Plant Science*, 12, 2021.

A. Smith, A. Norman, H. Kuchel, and B. Cullis. Plant variety selection using interaction classes derived from Factor Analytic Linear Mixed Models: models with information on genetic relatedness. 2023 - *In-prep*.