DWReml: An R package for fitting the linear mixed model

David Butler
Brian Cullis
Alison Smith

28 November 2023

Mixed Models and Experimental Design Lab (MMaDEL)
National Institute for Applied Statistics Research Australia
University of Wollongong
bcullis@uow.edu.au

Background

Motivation

- Over 25 years since the core of ASReml-R was devised.
- It has stood the test of time with the implementation of the average information algorithm and allowing users to fit a wide range of variance models, most importantly in recent times including those with genetic relatedness
- The R functional user interface has proved to be useful in specifying the terms in the linear mixed model and as well providing access to the R computing environment.
- Increasing demand for fitting models with genetic relatedness based on marker information has seriously impacted on the capability of ASReml-R to fit even simpler variance models to moderately sized data-sets in real time

Background

An open-source R *package which fits the linear mixed model and estimates variance components by residual maximum likelihood*

Specifications/Requirements

- An ASReml-R-like R functional user interface.
- Efficient ordering, analysis, factorization and solution of the mixed model equations - the current stumbling block with ASReml-R
- A modern computing paradigm written in C++ - move away from Fortran 90 (coding language used in ASReml-R)
- Match or exceed the number and types of variance models which are available in ASReml-R
- Address problems with convergence associated with fitting more complex variance models often experienced in ASReml-R.
- To become available in the public domain.

Introduction

The DWReml project: Dependent Wollongong Residual Maximum Likelihood

- Commenced about 18 months ago
- UOW team: David Butler and Sue Welham lead statistical computer scientists; Brian Cullis and Alison Smith lead statistical researchers; Robin Thompson maintains strong contact; Luke Mazur (Post Doc) statistical computing scientist
- In this talk we will semi-officially announce the first release of DWReml, viz V0.0.92 and
  - Provide brief overview of the main components and architecture
  - Describe the anatomy of a DWReml call
  - Summarize available variance models
  - Illustrate the calls for several tree breeding examples
  - Demonstrate performance against ASReml-R for MET examples in tree and chickpea breeding.

Introduction

DWReml fits the linear mixed model and estimates variance components by residual maximum likelihood using the Average Information algorithm and a supernodal sparse linear solver.

Features

- An ASReml-like R functional user interface.
- Efficient ordering, analysis, factorization and solution of the mixed model equations using the MUMPS (MUltifrontal Massively Parallel) Sparse direct solver.
- A modern computing paradigm written in C++.

Introduction

DWReml fits the linear mixed model and estimates variance components by residual maximum likelihood using the Average Information algorithm and a supernodal sparse linear solver.

Companion packages

- Numerous companion packages and methods to aid examination of model fit, tools to summarise results from the final model and form special design matrices.
- These include
    - `TPSbits`: Author Sue Welham - Creates structures to enable fit of 1D spline and 2D tensor-product splines of Rodriguez-Alvarez et al (2018)
    - `asv`, `variogram`, `dwrPlus`, `fasum.dwreml`, `iplot.dwreml`, `startmeup.dwreml`, `aom.dwreml`, `ss2met.dwreml`
- To become available in the public domain.

Solving the mixed model equations (MME)

- Direct methods for solving the MME use some form of Gaussian elimination.
- Matrix factorization and inversion are typically the most computationally demanding tasks when fitting the LMM.
- Cholesky factorization of the coefficient matrix $\boldsymbol{C} = \boldsymbol{LDL}^{\top}$ can yield computational efficiencies.
- A *supernode* is defined as a block of contiguous columns of $\boldsymbol{L}$ with the same sparsity pattern.
- Supernodal methods exploit this construct and are faster than traditional methods for solving the MME (see for example Mazur PhD 2022).
- Modern solvers also exploit parallel computing architectures.
- DWReml embeds the MUMPS equation solver https://mumps-solver.org/index.php in the AI algorithm for REML estimation.

Anatomy of a DWReml call

```
dwr <- dwreml(
 fixed = y ~Site,
   random = ~diag(Site):Genotype,
     residual=~dsum(~units|Site))
```

- The returned object
- Fixed effects terms
- Random effects terms

  Site:Genotype effects $\sim N\left(\mathbf{0}, \begin{bmatrix} \sigma^2_{s_1} & & & \\ & \sigma^2_{s_2} & & \\ & & \ddots & \\ & & & \sigma^2_{s_S} \end{bmatrix} \otimes \boldsymbol{I}_g\right)$

- Residual structure

  error effects $\sim N\left(\mathbf{0}, \begin{bmatrix} \sigma^2_1 \boldsymbol{I}_{n_1} & & & \\ & \sigma^2_2 \boldsymbol{I}_{n_2} & & \\ & & \ddots & \\ & & & \sigma^2_s \boldsymbol{I}_{n_s} \end{bmatrix}\right)$

Available variance models in common with ASReml-R

- Identity
  id(), *idv()*, *idh()*
- Correlation
  cor(), *corv()*, *corh()*
- General correlation
  corg(), *corgv()*, *corgh()*
- Time series
  ar1(), *ar1v()*, *ar1h()*
- General (co)variance
  diag(), *us()*, *rr()*
- General structure
  str()
- Known (genetic) structures
  vm(), *ide()*, *ric()*

Model constructor functions in common with ASReml-R

lin()	Form a variate from a factor.
pol()	Orthogonal polynomials.
at()	Forms binary factors.
grp()	A factor from contiguous data columns.
mbf()	A factor from columns in a linked file much easier syntax to use.

Performance/convenience extensions: cut

cut() Similar in operation to *at()* except that unused levels of the resulting factor(s) are dropped at each level of a *conditioning* factor. For example, if `Site:Variety` is sparse

```
cut(∼Variety | Site)
```

generates a `Variety` model term with a subset of levels for each level of `Site`. This reduces the column dimension of the design matrix and has performance implications when retrieving the elements of the inverse coefficient matrix.

ref Mazur, Cullis and Thompson (in prep)

use See chickpea example later in this talk

Performance/convenience extensions: (nested) dsum

dsum() Allows an additional — operator specifying a nested residual structure. For example, `dsum( ∼ar1(Column):ar1(Row)|Trial|Environment)` specifies levels of `Trial` are nested within `Environment`, and fits common correlation and section variance parameters to the *Trials* in an *Environment* grouping.

ref Jordan, Smith and Cullis (in prep)

use See chickpea example later here & Lu's talk

Performance/convenience extensions: xpr

xpr() Creates columns in the design matrix for a factor (or variate) that is the result of an algebraic expression with existing model terms as the operands.
The expression is given in an R formula object and the levels of all participating terms must conform (in size).
Allowed operators are ’+’, ’-’, ’*’ and ’/’ with any constants or coefficients given explicitly; all other symbols are expected to resolve to model terms.

use Fitting a reduced animal model for the tree example in this talk

Example model call

- Douglas Fir example (thanks to Trevor Doerksen)
- Multi-environment trial (MET) data-set with
  - 47 progeny trials conducted in 39 environments (locations)
  - Height data for 247628 trees
  - Trees derived from 1876 parents
  - 13.0% parental fill-in (% possible parent × environment combinations present in data)

Analysis of single environment: “fish”

```
fish.dat <- droplevels(tmp.dat[tmp.dat$site.grp=='fish',])
nrow(fish.dat) # 3994
length(levels(fish.dat$fem.gg)) # 150
length(levels(fish.dat$mal.gg)) # 19
length(levels(fish.dat$rep)) # 8
length(levels(fish.dat$rep_set)) # 4
```

DWReml call for Approximate Reduced Animal Model

```
library(dwreml)
fish.dwr <- dwreml(ht ~1,
random = ~rep + rep:rep_set +
           vm(xpr(~0.5*fem.gg + 0.5*mal.gg, Astar.sparse), Astar.sparse),
residual = ~units,
data = fish.dat)
```

Timing comparisons with ASReml-R

- Chickpea example (thanks to Kristy Hobson from Chickpea Breeding Australia)
- Multi-environment trial (MET) data-set with
  - 46 variety trials conducted in 2019 and 2020
  - 29 environments (trial location × year of conduct)
  - 4256 varieties with data
  - 4919 varieties with pedigree records (used to create numerator relationship matrix: NRM)
  - 10691 markers available for 4256 varieties (used to create genomic relationship matrix: GRM)
  - 26426 records (plots measured for grain yield)
  - 13.7% variety fill-in
  - 37.1% parental fill-in

Timing comparisons with ASReml-R

- MET analysis involved sequential fitting of 4 LMMs to individual plot yield data
    - factor analytic model (FA) for additive variety by environment effects (FA orders 1-4)
    - FA model of order 1 for non-additive variety by environment effects
    - Autoregressive spatial correlation models for each trial
    - Random effects for design factors for each trial

DWReml call for rr4rr1 model: not using cut()

```
rr4rr1 .dwr <- dwreml(yield~Environment + Environment:GDrop +
          at(Environment, cov.lst [[1]]): PlotLength ,
          random = ~rr(Environment,4):vm(GKeep, GSmet.ainv) +
          diag(Environment):vm(GKeep , GSmet.ainv) +
           rr(Environment):ide(GKeep) +
          diag(Environment):ide(GKeep) +
          at(Environment, expt.fit ): Trial  +
          at(Environment, col.fit ): Trial :Column +
          at(Environment, row.fit): Trial :Row +
          residual = ~dsum(~ar1(Column):ar1(Row) +
                    id(Column):ar1(Row) | Trial  | Environment,
                    levels  =  list (  ar1ar1  ,  idar1 )),
          data = nest.df  ,  na.action = na.method(x='include'),
          R.param=gam$R.sv, G.param=gam$G.sv)
```

Timing comparisons with ASReml-R: NRM

- At convergence, DWReml and ASReml-R reached same residual log-likelihood and same parameter estimates for each model
- But timings very different. Seconds per iteration:

Additive	Iteration		ASReml-R	
Model	Number	DWReml	4.1	4.2
FA1	1	34	3356	106
FA1	2+	31	666	52
FA2	1	40	4734	235
FA2	2+	38	1589	120
FA3	1	55	7449	513
FA3	2+	51	3410	284
FA4	1	73	?	775
FA4	2+	69	?	438

Timing comparisons with ASReml-R: GRM

- Models not run to convergence for this talk
- But timings very, very different. Seconds per iteration, where BB stands for Blue Bomb and would run out of memory / or computer crash:

Additive	Iteration	DWReml		ASReml-R	
Model	Number	Nocut	Cut	4.1	4.2
FA0	1	116	13	3232	312
FA0	2+	103	10	3083	258
FA1	1	314	37	BB	672
FA1	2+	262	26	?	449

Convergence comparisons with ASReml-R

- MET analysis of Douglas Fir data: Approximate Reduced Animal Models with 6 forms for additive parent by environment effects

				Variance
	Residual log-likelihood			Parameter
Model	DWReml		ASReml-R 4.1	Estimates
diag	-39931.56	$=$	-39931.56	equal
corh	-38423.86	$=$	-38423.86	equal
FA1 (def start)	-38219.20	$>>$	-38539.78	large diffs
FA1 (diag start)	-38219.20	$=$	-38219.19	equal
FA2 (FA1 start)	-38052.18	$<$	-38051.89	small diffs
FA3 (FA2 start)	-37999.66	$>$	-38003.94	large diffs
FA4 (FA3 start)	-37951.83	$>>$	-37960.34	large diffs