Recent updates in the BLUPF90 software suite

Daniela Lourenco

S. Tsuruta, I. Aguilar, Y. Masuda, M. Bermann,

A. Legarra, and I. Misztal

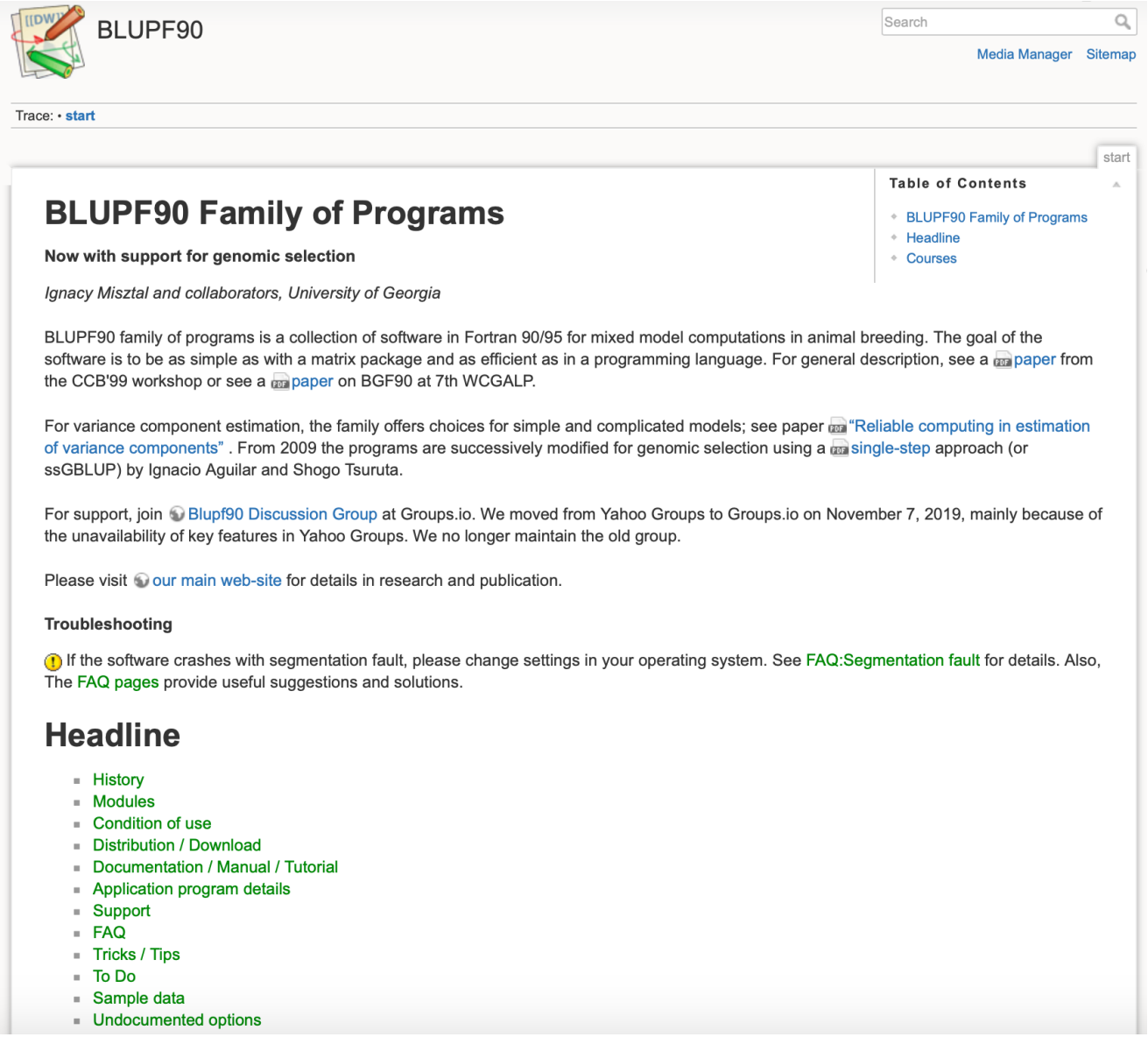
July 6, 2022

UNIVERSITY OF GEORGIA

College of Agricultural & Environmental Sciences

Animal Breeding and Genetics Group

BLUPF90 software suite


BLUPF90

Search

Media Manager Sitemap

Trace: • start

start

Table of Contents

- BLUPF90 Family of Programs
- Headline
- Courses

BLUPF90 Family of Programs

Now with support for genomic selection

Ignacy Misztal and collaborators, University of Georgia

BLUPF90 family of programs is a collection of software in Fortran 90/95 for mixed model computations in animal breeding. The goal of the software is to be as simple as with a matrix package and as efficient as in a programming language. For general description, see a paper from the CCB'99 workshop or see a paper on BGF90 at 7th WCGALP.

For variance component estimation, the family offers choices for simple and complicated models; see paper "Reliable computing in estimation of variance components" . From 2009 the programs are successively modified for genomic selection using a single-step approach (or ssGBLUP) by Ignacio Aguilar and Shogo Tsuruta.

For support, join Blupf90 Discussion Group at Groups.io. We moved from Yahoo Groups to Groups.io on November 7, 2019, mainly because of the unavailability of key features in Yahoo Groups. We no longer maintain the old group.

Please visit our main web-site for details in research and publication.

Troubleshooting

If the software crashes with segmentation fault, please change settings in your operating system. See FAQ:Segmentation fault for details. Also, The FAQ pages provide useful suggestions and solutions.

Headline

- History
- Modules
- Condition of use
- Distribution / Download
- Documentation / Manual / Tutorial
- Application program details
- Support
- FAQ
- Tricks / Tips
- To Do
- Sample data
- Undocumented options

- Collection of software
  - Fortran ≥ 90
  - Computations in AB & G
- Since 1997/1998 by Ignacy Misztal
- Several developers + collaborators
- Simple, efficient, and comprehensive
  - Very general models

BLUPF90 software suite

blupf90+

- blupf90 — BLUP with explicit equations
- remlf90 — Expectation Maximization REML
- airemlf90 — Average Information REML

gibbsf90+

- gibbsXf90 — Bayesian Analyses – linear traits
- thrgibbsXf90 — Bayesian Analyses – categorical traits

Genomics

- preGSf90 — Processing of SNP data (QC + matrices)
- qcf90 — QC of large SNP data
- postGSf90 — Estimation of SNP effects and GWAS
- predf90 — Prediction of GEBV based on SNP effects
- seekparentf90 — Parentage verification (SNP and pedigree)
- predictf90 — Adjusted and predicted phenotypes + residuals

Large-scale

- blup90iod2
- blup90iod2OMP1
- blup90iod3MPI1
- cblup90iod2
- cblup90iod2OMP1
- accf90
- accf90GS

blupf90+

- `blupf90`: MME solver
- `airemlf90`: variance components using Average Information REML
- `remlf90`: variance components using Expectation Maximization REML

Mixed Model Equations Solver

Variance Components Estimation

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{W}'\mathbf{R}^{-1}\mathbf{W}+\mathbf{A}^{-1}\otimes\mathbf{G}_0^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

blupf90+

MME Solver

Default

VC Estimation

- AI-REML:

```
OPTION method VCE
```

- EM-REML:

```
OPTION method VCE
OPTION EM-REML xx
```

xx → # of EM rounds

blupf90+

MME Solver

- Storing reliabilities based on PEV

  `OPTION store_accuracy X`

  Number of animal effect

$$Rel = 1 - \frac{PEV}{\sigma_u^2(1+f)}$$

  - Adjusts for f (inbreeding) from **A**, **G**, or **H**

- Storing solutions with original ID if renumf90 was used to renumber the data

  `OPTION origID`

  - Only *solutions.original* is created

gibbsf90+

- `gibbs1f90`: stores single trait matrices once – fast for multi-trait models
- `gibbs2f90`: gibbs1f90 with joint sampling of correlated effects – Maternal effects and RRM
- `gibbs3f90`: gibbs2f90 with heterogeneous residual variance
- `thrgibbs1f90`: for linear-threshold models
- `thrgibbs3f90`: thrgibbs1f90 with heterogeneous residual variance

Variance Components Estimation

Mixed Model Equations Solver

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{W}'\mathbf{R}^{-1}\mathbf{W}+\mathbf{A}^{-1}\otimes\mathbf{G}_0^{-1} \end{bmatrix} \begin{bmatrix} \hat{\boldsymbol{\beta}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}'\mathbf{R}^{-1}\mathbf{y} \end{bmatrix}$$

gibbsf90+

Linear

Default

Threshold (-Linear)

```
OPTION cat 0 2 5
```

Programs

Available for research (free)

- BLUPF90+ - a combined program of blupf90, remlf90, and airemlf90
- GIBBSF90+ - a combined program of gibbs1f90, gibbs2f90, gibbs3f90, thrgibbs1f90, and thrgibbs3f90
- POSTGIBBSF90 - statistics and graphics for post-Gibbs analysis (S. Tsuruta)
- RENUMF90 - a renumbering program that also can check pedigrees and assign unknown parent groups; supports large data sets
- PREGSF90 – genomic preprocessor that combines genomic and pedigree relationships (I. Aguilar)
- POSTGSF90 – genomic postprocessor that extracts SNP solutions after genomic evaluations (single step, GBLUP) (I. Aguilar)
- PREDICTF90 - a program to calculate adjusted y, y_hat, and residuals (I. Aguilar)
- PREDF90 - a program to predict direct genomic value (DGV) for animals based on genotypes and SNP solution
- QCF90 - a quality-control tool on genotypes and pedigree information (Y. Masuda)
- INBUPGF90 - a program to calculate inbreeding coefficients with incomplete pedigree (I. Aguilar)
- SEEKPARENTF90 - a program to verify paternity and parent discovery using SNP markers (I. Aguilar)

No longer updated (as of May 2022)

- BLUPF90 - BLUP in memory
- REMLF90 - accelerated EM REML
- AIREMLF90 - Average Information REML with several options including EM-REML and heterogeneous residual variances (S. Tsuruta)
- GIBBSF90 - simple block implementation of Gibbs sampling - no genomic
- GIBBS1F90 - as above but faster for creating mixed model equations only once
- GIBBS2F90 - as above but with joint sampling of correlated effects
- GIBBS3F90 - as above with support for heterogeneous residual variances
- THRGIBBSF90 - Gibbs sampling for any combination of categorical and linear traits (D. Lee) - no genomic
- THRGIBBS1F90 - as above but simplified with several options (S. Tsuruta)
- THRGIBBS3F90 - as above with heterogeneous residual variances for linear traits

renumf90

Renumbering software for the BLUPF90 suite

- **Renumbers data and pedigree**
- **Creates a parameter file for BLUPF90 family**
  - **Parameter file can be modified by the users for new models**
- **Traces back pedigree for individuals in the data**
- **Performs comprehensive pedigree checks**
- **Provides data statistics**
- **Creates an Xref file for genotyped individuals**
- **Computes inbreeding by default** 

Inbreeding in ssGBLUP

Computed using Henderson-Quaas' algorithm **with inbreeding**

Computed using VanRaden's formula, which considers inbreeding

Computed using Colleau's algorithm, which considers inbreeding

before now

$$\mathbf{H}^{-1}=\mathbf{A}^{-1}+\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}^{-1}-\mathbf{A}_{22}^{-1}\end{bmatrix}$$

$$\mathbf{H}^{-1}=\mathbf{A}^{-1}+\begin{bmatrix}\mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{G}^{-1}-\mathbf{A}_{22}^{-1}\end{bmatrix}$$

```
DATAFILE
phenotypes.txt
TRAITS
 3
FIELDS_PASSED TO OUTPUT

WEIGHT(S)

RESIDUAL_VARIANCE
0.60
EFFECT
2 cross alpha
EFFECT
1  cross alpha
RANDOM
animal
FILE
pedigree.txt
FILE_POS
1 2 3 0 0
SNP_FILE
genotypes.txt
PED_DEPTH
0
INBREEDING
pedigree
(CO)VARIANCES
0.40
```

RENUMF90 ≥ v1.157

```
DATAFILE
phenotypes.txt
TRAITS
 3
FIELDS_PASSED TO OUTPUT

WEIGHT(S)

RESIDUAL_VARIANCE
0.60
EFFECT
2 cross alpha
EFFECT
1  cross alpha
RANDOM
animal
FILE
pedigree.txt
FILE_POS
1 2 3 0 0
SNP_FILE
genotypes.txt
PED_DEPTH
0
(CO)VARIANCES
0.40
```

Accounting for missing parents / base

- QP-transformation for $\mathbf{A}^{-1}$ (Quaas & Pollack, 1981;Westell et al., 1988)

$$\mathbf{A}^{*} = \begin{bmatrix} \mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{Q} \\ -\mathbf{Q}'\mathbf{A}^{-1} & \mathbf{Q}'\mathbf{A}^{-1}\mathbf{Q} \end{bmatrix}$$

- QP-transformation for $\mathbf{H}^{-1}$ (Misztal et al., 2013)

$$\mathbf{H}^{*} = \mathbf{A}^{*} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} & -(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1})\mathbf{Q}_2 \\ 0 & -\mathbf{Q}_2'(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}) & \mathbf{Q}_2'(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1})\mathbf{Q}_2 \end{bmatrix}$$

```
OPTION exact_upg
```

- Altered QP-transformation for $\mathbf{H}^{-1}$ (Tsuruta et al., 2019)

$$\mathbf{H}^{*} = \mathbf{A}^{*} + \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1} & -(-\mathbf{A}_{22}^{-1})\mathbf{Q}_2 \\ 0 & -\mathbf{Q}_2'(-\mathbf{A}_{22}^{-1}) & \mathbf{Q}_2'(-\mathbf{A}_{22}^{-1})\mathbf{Q}_2 \end{bmatrix}$$

```
OPTION exact_upg
OPTION TauOmegaQ2 0 1
```

Accounting for missing parents / base

- Metafounders (Legarra et al., 2015)
  - **A** is modified to be compatible with **G** with 0.5 AF

$$\mathbf{H}_{\Gamma}^{-1} = \mathbf{A}_{\Gamma}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_{05}^{-1} - \mathbf{A}_{\Gamma 22}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}$$

MF implemented in `blupf90+`

- $\boldsymbol{\Gamma}$ = matrix of relationships within and across MF
- `gammaf90` under development & test

Improving efficiency for large data

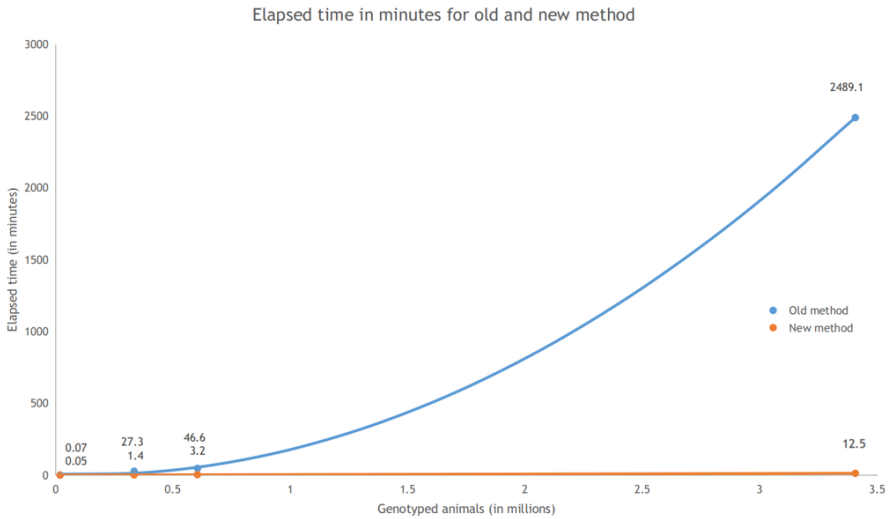
APY $\mathbf{G}^{-1}$: $$\mathbf{G}_{APY}^{-1} = \begin{bmatrix} \mathbf{G}_{cc}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} + \begin{bmatrix} -\mathbf{G}_{cc}^{-1}\mathbf{G}_{cn} \\ \mathbf{I} \end{bmatrix} \mathbf{M}_{nn}^{-1} \begin{bmatrix} -\mathbf{G}_{nc}\mathbf{G}_{cc}^{-1} & \mathbf{I} \end{bmatrix}$$

(Misztal et al., 2014)

$\mathbf{A}_{22}^{-1}$ components: $$\mathbf{A}_{22}^{-1} = \mathbf{A}^{22} - \mathbf{A}^{21}(\mathbf{A}^{11})^{-1}\mathbf{A}^{12}$$

(Standen & Mantysaari, 2014)
(Masuda et al., 2017)

- Blending: $\mathbf{G} = \alpha\, \mathbf{G}^{*} + (1- \alpha)\, \mathbf{A}_{22}$
  - Colleau (2002)
  - 3.4M genotyped Holsteins 15k core: 41 hours
  - Rearranging Colleau for core and noncore: 12 minutes
  - Improvements for > 1M genotyped animals



Improving efficiency for large data

- Bit-wise storage and operations
- Transposed SNP storage

Genotype	Character	ASCII (8bits)	Re-coded (2bits)
Homozygote (AA)	“0”	00110010	10
Heterozygote (Aa)	“1”	00110001	11
Homozygote (aa)	“2”	00110000	01
Missing	“5”	00110101	00

- zlib – compression library – reads zipped genotypes
- Can read PLINK format
- Multibreed evaluations with 30M animals and 3.9M genotyped
  - ~ 1 day
  - 1 TB

P-values for SNP effects - ssGWAS

1) Factorize and Invert LHS of ssGBLUP with YAMS (Masuda et al., 2014)

blupf90+

2) Solve the MME for $\begin{bmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{u}}\end{bmatrix}$ using the sparse Cholesky factor

```
OPTION snp_p_value
```

3) Extract coefficients for genotyped animals ($\mathbf{C}^{u_2u_2}$) from LHS^{-1}

4) Obtain individual prediction error variance of SNP effects:

$$Var(\hat{a}_i) = \alpha b \frac{1}{2\sum p_i(1-p_i)} \mathbf{z}_i' \mathbf{G}^{-1} (\mathbf{G}\sigma_u^2 - \mathbf{C}^{u_2u_2}) \mathbf{G}^{-1} \mathbf{z}_i \frac{1}{2\sum p_i(1-p_i)} \alpha b$$

(Gualdron-Duarte et al., 2014)

postGSf90

5) Backsolve GEBV to SNP effects ($\hat{a}$): $\hat{a} = \alpha b \frac{1}{2\sum p_i q_i} \mathbf{Z}'\mathbf{G}^{-1}\hat{u}$

```
OPTION snp_p_value
```

6) p-value$_i$ $= 2\left(1 - \Phi\left(\left|\frac{\hat{a}_i}{sd(\hat{a}_i)}\right|\right)\right)$ [same p-values as e.g. EMMAX]

Φ is the cumulative standard normal function

Command line arguments

- Help
  - `renumf90 --help`
  - `renumf90 --show-template`
  - `blupf90+ --help`
  - `blupf90+ --help-genomic`

- Running the programs
  - `blupf90+ parfile.par`
  - `gibbsf90+ parfile.par --samples i --burnin j --interval k`

Final remarks

- BLUPF90 software suite under constant development for 25 years
  - Efficiency
  - Flexibility
  - Simplicity
  - Fits any models for AB&G
- Five active programmers
  - I. Aguilar, A. Legarra, M. Bermann, S. Tsuruta, Y. Masuda
- Active discussion group
  - https://groups.io/g/blupf90
- Comprehensive website and documentation
  - http://nce.ads.uga.edu/wiki

Acknowledgments