IASBLUP v1.3.0

User Manual

Software for Large-Scale Genetic Evaluation

Developer: Wentao Cai

Affiliation: Chinese Academy of Agricultural Sciences (CAAS)

Homepage: iasbreeding.cn

Contact: caiwentao@caas.cn

Platforms: Linux / macOS / Windows

Table of Contents

1. Introduction

IASBLUP (Intelligent Agricultural Solutions of BLUP) is a comprehensive software package designed for large-scale genetic evaluation. It supports genomic selection (GBLUP), pedigree-based BLUP (ABLUP), and single-step GBLUP (ssGBLUP). The software integrates three core functionalities: kinship matrix construction, REML variance component estimation, and estimated breeding value (EBV) prediction—covering the full pipeline of routine genetic evaluation in animal and plant breeding programs.

Core Features

Module	Description
Kinship (Relationship Matrix)	Build G, A, A^{-1}, H, and H^{-1} matrices
REML Variance Estimation	Dense & sparse, single/multi-component, univariate & bivariate REML
Breeding Value Prediction	GBLUP, multi-component prediction, and bivariate joint prediction

What's New in v1.3.0

- Sparse A^{-1} / H^{-1} MME-based REML
- PCG + Stochastic Lanczos Quadrature (SLQ) for large-scale variance estimation
- Low-rank approximation for accelerated computation
- Bivariate REML with genetic correlation analysis
- Multi-component variance estimation and HE regression
- Trait-specific covariate specification
- Automatic kinship file format detection (binary / text)

2. Installation and System Requirements

System Requirements

- Operating System: Linux (recommended), macOS, Windows
- Memory: ≥ 8 GB recommended (more for large datasets)
- Dependencies: Eigen, OpenMP (for multi-threading), BLAS/LAPACK

Basic Usage

```
IASBLUP <mode> [options]
```

Help

```
IASBLUP --help
```

3. Quick Start

Example 1: Build Genomic Relationship Matrix (G)

```
IASBLUP --kinship --bedfile data_prefix --out G
```

Example 2: Build Pedigree Relationship Matrix (A)

```
IASBLUP --kinship --pedigree ped.txt --out A
```

Example 3: Build Inverse Pedigree Matrix (A^{-1})

```
IASBLUP --kinship --pedigree ped.txt --Ainv --out Ainv_full
```

Example 4: Build Inverse H Matrix (H^{-1})

```
IASBLUP --kinship --bedfile data_prefix --pedigree ped.txt --Hinv --out Hinv_full
```

Example 5: Univariate REML with Dense Kinship

```
IASBLUP --reml --kin-file G.bin --phefile pheno.txt --phe-pos 2 --out uni
```

Example 6: Univariate REML with Sparse A^{-1}/H^{-1} (MME Mode)

```
IASBLUP --reml --MME --kin-file Ainv_full --phefile pheno.txt --phe-pos 2 --out
uni_mme
```

Example 7: Predict Breeding Values (EBV)

```
IASBLUP --predict --kin-file G.bin --phefile pheno.txt --phe-pos 2 --out ebv
```

Example 8: Bivariate REML

```
IASBLUP --two-traits --kin-file G.bin --phefile pheno.txt --phe-pos1 2 --phe-pos2
3 --out bi
```

4. Input File Formats

4.1 Genotype Files (PLINK Binary Format)

Specify the file prefix with --bedfile. The program automatically reads:

File	Description
prefix.bed	Binary genotype data
prefix.bim	SNP marker information
prefix.fam	Individual information

4.2 Pedigree File

Specify with --pedigree. Three whitespace-delimited columns, no header:

```
IndividualID    SireID    DamID
```

Unknown parents are coded as 0.

Example:

```
A001    S001    D001
A002    S001    D002
A003    S002    D001
A004    0       D003
```

4.3 Phenotype File

Specify with --phefile. Whitespace-delimited with the first column as individual ID:

```
ID      Fixed1    Trait1    Trait2    Covar1    Covar2
A001    1         12.5      8.3       2         1.5
A002    2         15.2      NA        1         2.3
A003    1         NA        9.1       2         1.8
A004    2         13.8      7.6       1         2.1
```

- Missing values (NA) are automatically filtered.
- Use --phe-pos to specify the phenotype column (1-based indexing).
- Covariates are specified with --covar-pos (factor) and --qcovar-pos (quantitative).

4.4 Kinship Matrix Files

IASBLUP auto-detects the following formats:

Binary Format (recommended)

File	Description
prefix.bin	Matrix data (binary float)
prefix.id	Individual ID list

Text Format

Full matrix in whitespace-delimited text, with or without ID row/column.

5. Module Reference

5.1 Kinship Matrix Construction (--kinship)

5.1.1 G Matrix (Genomic Relationship Matrix)

Build a genomic relationship matrix from SNP marker data.

```
IASBLUP --kinship --bedfile data_prefix --out G
```

Optional parameters:

Parameter	Description	Default
--kin-method <int>	GRM construction method	1
--kin-lambda <number>	Lambda weighting parameter	1.0
--block-size <int>	SNP block size for streaming (0=auto)	0
--threads <int>	Number of CPU threads	1
--kin-bin	Write output in GCTA binary format	No
--also-text	Also export text format	No

5.1.2 A Matrix (Pedigree Relationship Matrix)

Build the additive genetic relationship matrix from pedigree data.

```
IASBLUP --kinship --pedigree ped.txt --out A
```

5.1.3 A^{-1} Matrix (Inverse Pedigree Relationship Matrix)

Build a sparse A^{-1} directly from pedigree, for use in MME-based REML.

```
IASBLUP --kinship --pedigree ped.txt --Ainv --out Ainv_full
```

Output format: Sparse binary (prefix.bin + prefix.id).

5.1.4 H Matrix (Combined Relationship Matrix)

Build the H matrix combining pedigree and genomic information (single-step method).

```
IASBLUP --kinship --bedfile data_prefix --pedigree ped.txt --out H
```

Tuning parameters:

Parameter	Description	Default
--kin-tau <number>	Scaling factor τ for G_{22}	1.0
--kin-omega <number>	Scaling factor ω for A_{22}	1.0

5.1.5 H^{-1} Matrix (Inverse Combined Relationship Matrix)

Build a sparse H^{-1} for large-scale single-step BLUP evaluation.

```
IASBLUP --kinship --bedfile data_prefix --pedigree ped.txt --Hinv --out Hinv_full
```

Note: Both genotype and pedigree files are required for this mode.

5.2 Variance Component Estimation (--reml)

IASBLUP provides multiple REML methods to accommodate datasets of different scales.

5.2.1 Dense Single-Component REML

```
IASBLUP --reml --kin-file G.bin --phefile pheno.txt --phe-pos 2 --out uni
```

Variance Estimation Method (--var-method)

Method	Code	Suitable For	Description
Direct	1	n < 2,000	Direct matrix inversion; exact but slow
Eigen	2	n < 10,000	Eigen-decomposition; moderate scale
Cholesky	3	n < 10,000	Cholesky decomposition; stable and efficient
Low-rank	4	n < 50,000	Low-rank approximation; large scale
PCG+SLQ (default)	5	n > 10,000	PCG iteration + stochastic trace; best for large scale

Examples:

```
# Eigen-decomposition method
IASBLUP --reml --kin-file G.bin --phefile pheno.txt --phe-pos 2 \
        --var-method eigen --out uni

# Low-rank approximation with custom parameters
IASBLUP --reml --kin-file G.bin --phefile pheno.txt --phe-pos 2 \
        --var-method lowrank --lowrank-ratio 0.95 --lowrank-maxrank 800 --out uni
```

Low-Rank Method Parameters

Parameter	Description	Default	Recommended Range
--lowrank-ratio <num>	Target variance explained	0.95	0.90 – 0.99
--lowrank-maxrank <int>	Maximum retained rank	800	50 – 2000

5.2.2 Sparse MME-Based REML

When using the pedigree inverse A^{-1} or combined inverse H^{-1}, employ MME-based AI-REML.

```
IASBLUP --reml --MME --kin-file Ainv_full --phefile pheno.txt --phe-pos 2 --out
mme_uni
```

MME-REML parameters:

Parameter	Description	Default

--MME	Enable MME mode	No
--normalize	Hutchinson normalization of A^{-1}/H^{-1}	No
--no-hutch	Skip Hutchinson normalization even if --normalize is set	No
--hutch-score <int>	Number of Hutchinson probes for AI/score estimation	48
--step0 <num>	Initial line-search step size	1.0

Full example with covariates:

```
IASBLUP --reml --MME --kin-file Ainv_full \
        --phefile pheno.txt --phe-pos 2 \
        --covar-pos 3,4 --qcovar-pos 5 \
        --normalize --hutch-score 80 \
        --max-iter 50 --out mme_uni
```

5.2.3 Multi-Component REML

Jointly estimate variance components from multiple kinship matrices (e.g., additive + dominance).

```
IASBLUP --reml --multi-kins G_add.bin,G_dom.bin \
        --component-names Additive,Dominance \
        --phefile pheno.txt --phe-pos 2 --out multi
```

Parameter	Description
--multi-kins <files>	Multiple kinship files, comma-separated (≥ 2)
--component-names <names>	Component names, comma-separated (optional)

5.2.4 Haseman–Elston (HE) Regression

A fast and robust alternative to REML for variance component estimation.

```
IASBLUP --reml --he --multi-kins G1.bin,G2.bin \
        --phefile pheno.txt --phe-pos 2 --out he_result
```

Parameter	Description	Default
--he or --he-regression	Use HE regression instead of REML	No
--he-max-pairs <int>	Maximum number of HE pairs	100,000

5.2.5 Covariates

IASBLUP supports two types of covariates:

Parameter	Type	Description
--covar-pos <list>	Factor	Categorical variables (e.g., sex, herd); automatically dummy-

		coded
--qcovar-pos <list>	Quantitative	Continuous variables (e.g., weight, age); entered directly

Multiple columns are comma-separated:

```
--covar-pos 3,4 --qcovar-pos 5,6
```

5.3 Breeding Value Prediction (--predict)

5.3.1 Standard Prediction Workflow

Estimate variance components first, then predict breeding values for all individuals:

```
IASBLUP --predict --kin-file G.bin --phefile pheno.txt --phe-pos 2 --out ebv
```

5.3.2 Prediction with Known Heritability

Skip REML and use a user-supplied heritability value:

```
IASBLUP --predict --kin-file G.bin --phefile pheno.txt --phe-pos 2 --h2 0.35 --
out ebv
```

Parameter	Description	Range
--h2 <number>	Provide heritability directly; skip REML	0.0 – 1.0

Note: --h2 is only available in --predict mode for single-component analysis.

5.3.3 Multi-Component Prediction

```
IASBLUP --predict --multi-kins G_add.bin,G_dom.bin \
        --component-names Additive,Dominance \
        --phefile pheno.txt --phe-pos 2 --out multi_ebv
```

Multi-component prediction estimates the variance contribution of each component and outputs the corresponding random effects.

5.3.4 Prediction Output

After prediction, the program outputs:

- Fixed effect estimates with standard errors and ANOVA table
- Breeding values (EBV) for both phenotyped and non-phenotyped individuals
- Output file: prefix.random.txt

5.4 Bivariate Analysis (--two-traits)

5.4.1 Dense Bivariate REML

```
IASBLUP --two-traits --kin-file G.bin --phefile pheno.txt \
        --phe-pos1 2 --phe-pos2 3 --out bi
```

5.4.2 Sparse MME Bivariate REML

```
IASBLUP --two-traits --MME --kin-file Ainv_full \
        --phefile pheno.txt --phe-pos1 2 --phe-pos2 3 \
        --normalize --out bi_mme
```

5.4.3 Trait-Specific Covariates

Bivariate analysis supports separate covariate specifications for each trait:

Parameter	Description
--covar-pos1 <list>	Factor covariates for trait 1
--covar-pos2 <list>	Factor covariates for trait 2
--qcovar-pos1 <list>	Quantitative covariates for trait 1
--qcovar-pos2 <list>	Quantitative covariates for trait 2

Full example:

```
IASBLUP --two-traits --MME --kin-file Ainv_full \
        --phefile pheno.txt --phe-pos1 2 --phe-pos2 3 \
        --covar-pos1 4,5 --covar-pos2 4,6 \
        --qcovar-pos1 7 --qcovar-pos2 7,8 \
        --normalize --hutch-score 80 \
        --out bi_mme --predict
```

Backward compatibility: If only --covar-pos / --qcovar-pos (without the 1/2 suffix) are used, they are applied to both traits.

5.4.4 Bivariate Joint Prediction

Append --predict to also output EBVs for both traits:

```
IASBLUP --two-traits --kin-file G.bin --phefile pheno.txt \
        --phe-pos1 2 --phe-pos2 3 --out bi --predict
```

5.4.5 Bivariate Output

Estimate	Description
Var(G_trait1)	Additive genetic variance for trait 1
Var(G_trait2)	Additive genetic variance for trait 2
Cov(G_12)	Additive genetic covariance
Var(E_trait1)	Residual variance for trait 1
Var(E_trait2)	Residual variance for trait 2
Cov(E_12)	Residual covariance

h^2(Trait 1)	Heritability of trait 1
h^2(Trait 2)	Heritability of trait 2
rG	Genetic correlation
rE	Residual correlation
rP	Phenotypic correlation

All estimates are reported with standard errors (SE).

6. Complete Parameter Reference

Mode Selection

Parameter	Description
--kinship	Build kinship / relationship matrix
--reml	REML variance component estimation
--predict	Predict breeding values
--two-traits	Bivariate analysis
--help	Display help message

Input Files

Parameter	Description
--bedfile <prefix>	PLINK binary genotype file prefix
--pedigree <file>	Pedigree file (3 columns: ID Sire Dam)
--phefile <file>	Phenotype file
--kin-file <file>	Kinship matrix file (prefix or path)
--multi-kins <files>	Multiple kinship files, comma-separated

Phenotype and Covariates

Parameter	Description
--phe-pos <int>	Phenotype column position (1-based)
--phe-pos1 <int>	Trait 1 column (bivariate mode)
--phe-pos2 <int>	Trait 2 column (bivariate mode)
--covar-pos <list>	Factor covariate columns, comma-separated
--qcovar-pos <list>	Quantitative covariate columns, comma-separated
--covar-pos1 <list>	Trait 1 factor covariates (bivariate)
--covar-pos2 <list>	Trait 2 factor covariates (bivariate)
--qcovar-pos1 <list>	Trait 1 quantitative covariates (bivariate)
--qcovar-pos2 <list>	Trait 2 quantitative covariates (bivariate)

Kinship Matrix Parameters

Parameter	Description	Default
--Ainv	Build inverse pedigree matrix (sparse)	–
--Hinv	Build inverse H matrix (sparse)	–
--kin-method <int>	GRM construction method	1
--kin-lambda <number>	Lambda parameter for G matrix	1.0
--kin-tau <number>	Scaling factor τ for G_{22} in H/H^{-1}	1.0

--kin-omega <number>	Scaling factor ω for A_{22} in H/H^{-1}	1.0
--block-size <int>	SNP block size (0=auto)	0
--kin-bin	Write in GCTA binary format	–
--also-text	Also export text format	–

REML Parameters

Parameter	Description	Default
--var-method <method>	Variance estimation method (1–5)	5 (PCG+SLQ)
--max-iter <int>	Maximum REML iterations	30
--h2 <number>	User-supplied heritability (--predict only)	–
--lowrank-ratio <num>	Low-rank target variance ratio	0.95
--lowrank-maxrank <int>	Low-rank maximum rank	800
--he / --he-regression	Use HE regression	–
--he-max-pairs <int>	Maximum HE pairs	100,000
--component-names <str>	Multi-component names, comma-separated	G1,G2,…

MME / Sparse Matrix Parameters

Parameter	Description	Default
--MME	Enable MME mode (sparse inverse matrix)	–
--normalize	Hutchinson normalization of A^{-1}/H^{-1}	–
--no-hutch	Skip Hutchinson normalization	–
--hutch-score <int>	Number of Hutchinson probes	48
--step0 <num>	Initial line-search step size	1.0

General Parameters

Parameter	Description	Default
--threads <int>	Number of CPU threads	1
--out <prefix>	Output file prefix	Results

7. Output Files

7.1 Kinship Matrix Output

Dense Matrix

File	Description
prefix.bin	Matrix data (binary float)
prefix.id	Individual ID list
prefix.txt (optional)	Text format matrix

Sparse Inverse Matrix (A^{-1} / H^{-1})

File	Description
prefix.bin	Sparse matrix data (binary)
prefix.id	Individual ID list

7.2 REML Output

File	Description
prefix.log	Full analysis log (iterations, variance components, heritability, etc.)

The log file contains:

- Iteration-by-iteration parameter updates
- Final variance component estimates with standard errors
- Heritability estimates with standard errors
- (Bivariate mode) genetic, residual, and phenotypic correlations

7.3 Prediction Output

File	Description
prefix.random.txt	Breeding values for all individuals (ID + EBV)
prefix.ebv	EBV file (MME mode)
prefix.sol	Fixed effect solutions and ANOVA table
prefix.log	Analysis log

EBV File Format (Univariate)

```
ID      EBV
A001    0.5234
A002    -0.3127
A003    0.1893
```

EBV File Format (Bivariate)

```
ID      EBV_T1    EBV_T2
A001    0.5234    0.3891
A002    -0.3127   0.1456
A003    0.1893    -0.2345
```

8. Typical Workflows

Workflow 1: Genomic Selection (GBLUP)

```
# Step 1: Build G matrix
IASBLUP --kinship --bedfile geno --threads 4 --out G

# Step 2: REML variance component estimation
IASBLUP --reml --kin-file G.bin --phefile pheno.txt --phe-pos 2 \
        --covar-pos 3,4 --qcovar-pos 5 --var-method 5 --out reml_result

# Step 3: Predict breeding values
IASBLUP --predict --kin-file G.bin --phefile pheno.txt --phe-pos 2 \
        --covar-pos 3,4 --qcovar-pos 5 --out ebv_result
```

Workflow 2: Single-Step GBLUP (ssGBLUP)

```
# Step 1: Build H⁻¹ matrix
IASBLUP --kinship --bedfile geno --pedigree ped.txt --Hinv \
        --kin-tau 1.0 --kin-omega 1.0 --threads 4 --out Hinv

# Step 2: MME-based REML
IASBLUP --reml --MME --kin-file Hinv --phefile pheno.txt --phe-pos 2 \
        --covar-pos 3,4 --normalize --hutch-score 80 --out ssGBLUP

# Step 3: BLUP prediction (add --predict flag)
IASBLUP --reml --MME --kin-file Hinv --phefile pheno.txt --phe-pos 2 \
        --covar-pos 3,4 --normalize --predict --out ssGBLUP
```

Workflow 3: Bivariate Genetic Correlation Analysis

```
# Build G matrix
IASBLUP --kinship --bedfile geno --threads 4 --out G

# Bivariate REML + prediction
IASBLUP --two-traits --kin-file G.bin --phefile pheno.txt \
        --phe-pos1 2 --phe-pos2 3 \
        --covar-pos1 4,5 --covar-pos2 4,6 \
        --qcovar-pos1 7 --qcovar-pos2 7,8 \
        --max-iter 200 --out bivar --predict
```

Workflow 4: Multi-Component Variance Estimation

```
# Multi-component REML
IASBLUP --reml --multi-kins G_add.bin,G_dom.bin \
        --component-names Additive,Dominance \
        --phefile pheno.txt --phe-pos 2 --out multi_var

# Multi-component prediction
IASBLUP --predict --multi-kins G_add.bin,G_dom.bin \
        --component-names Additive,Dominance \
        --phefile pheno.txt --phe-pos 2 --out multi_ebv
```

9. Statistical Methods

9.1 Genomic Relationship Matrix (G)

Based on VanRaden (2008) Method 1:

$$G = ZZ' / 2\Sigma p_i(1-p_i)$$

where Z is the centered and standardized genotype matrix and p_i is the allele frequency of the i-th SNP.

9.2 REML Variance Component Estimation

Based on the mixed linear model:

$$y = X\beta + Zu + e$$

where:

- y: phenotype vector
- X: fixed-effect design matrix
- β: fixed effects
- Z: random-effect incidence matrix
- u ~ N(0, Gσ²_a): additive genetic effects
- e ~ N(0, Iσ²_e): residual effects

REML iteratively estimates σ²_a and σ²_e using the Average Information (AI) algorithm.

9.3 Comparison of Variance Estimation Methods

Method	Time Complexity	Space Complexity	Accuracy	Suitable Scale
Direct	$O(n^3)$	$O(n^2)$	Exact	n < 2K
Eigen	$O(n^3)$	$O(n^2)$	Exact	n < 10K
Cholesky	$O(n^3)$	$O(n^2)$	Exact	n < 10K
Low-rank	$O(nk^2)$	O(nk)	Approximate	n < 50K
PCG+SLQ	$O(n^2)$ per iter	$O(n^2)$	Approximate	n > 10K

Recommendation: For n > 10,000, the default PCG+SLQ method is recommended.

9.4 MME-Based REML

For sparse inverse matrices (A^{-1} or H^{-1}), the Mixed Model Equations (MME) framework is used. Within the AI-REML iteration, the Hutchinson stochastic trace estimator is used to approximate the score vector and AI matrix.

9.5 Bivariate Model

The bivariate model simultaneously analyzes two traits, estimating the genetic covariance matrix G_o and residual covariance matrix R_o.

Genetic correlation:

$$r_G = \sigma_{g12} / \sqrt{(\sigma^2_{g1} \times \sigma^2_{g2})}$$

10. Frequently Asked Questions (FAQ)

Q1: How do I choose the right --var-method?

Data Size (n)	Recommended Method
< 2,000	Direct (1) or Eigen (2)
2,000 – 10,000	Cholesky (3) or Eigen (2)
10,000 – 50,000	Low-rank (4) or PCG+SLQ (5)
> 50,000	PCG+SLQ (5) (default)

Q2: When should I use --MME mode?

--MME mode is required when using sparse inverse matrices (A^{-1} or H^{-1}). It is particularly useful for: (1) large-scale pedigree data (tens to hundreds of thousands of individuals); (2) single-step GBLUP evaluation (ssGBLUP); (3) scenarios where non-genotyped individuals also need breeding value estimates.

Q3: What does --normalize do?

It normalizes the A^{-1} or H^{-1} matrix by the estimated mean of diag(A), computed via the Hutchinson stochastic trace estimator. This improves REML convergence and numerical stability.

Q4: How are missing phenotype values handled?

Records with missing values (NA) are automatically excluded. In bivariate analysis, missing values are handled independently for each trait, allowing different effective sample sizes.

Q5: How can I speed up large-scale analysis?

- Multi-threading: Increase threads with --threads
- Block computing: Use --block-size during G matrix construction to manage memory
- Method selection: Use PCG+SLQ (--var-method 5) for large data
- Sparse matrices: Use --MME mode with A^{-1}/H^{-1} for very large populations

Q6: What if REML does not converge?

- Increase maximum iterations: --max-iter 100
- Adjust initial step size: --step0 0.5
- Increase Hutchinson probes: --hutch-score 100 (MME mode)
- Check data quality: ensure no extreme phenotype outliers and that the kinship matrix is positive definite

Q7: How do I tune Low-rank parameters?

- --lowrank-ratio 0.95: Retains 95% of variance; suitable for most cases
- --lowrank-ratio 0.99: Higher accuracy at increased computational cost
- --lowrank-maxrank: Typically need not exceed 1000; depends on sample size and population structure

11. Citation

If you use IASBLUP in your research, please cite:

Cai, W.T. IASBLUP: An integrated software for large-scale genetic evaluation. iasbreeding.cn

Related References

1. VanRaden, P.M. (2008). Efficient methods to compute genomic predictions. Journal of Dairy Science, 91(11), 4414–4423.
2. Henderson, C.R. (1984). Applications of Linear Models in Animal Breeding.
3. Aguilar, I. et al. (2010). Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal of Dairy Science, 93(2), 743–752.
4. Misztal, I. et al. (2009). Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. Journal of Dairy Science, 92(9), 4648–4655.