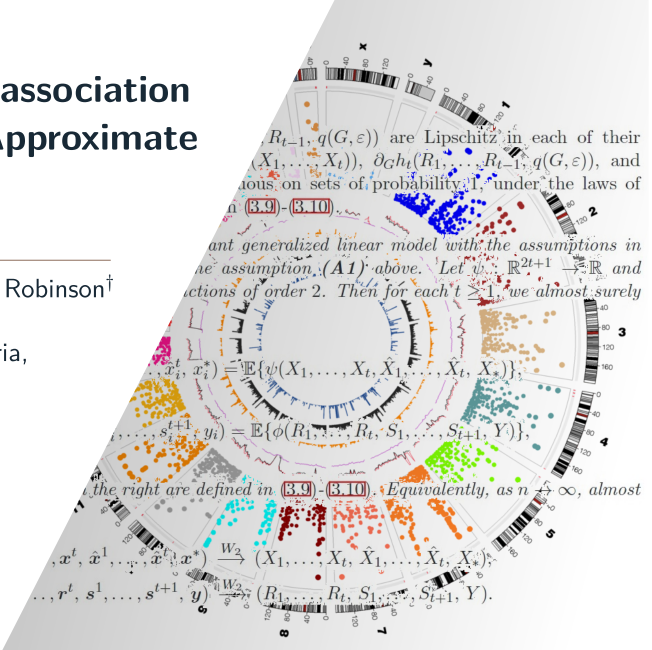
Light-speed whole genome association testing and prediction via Approximate Message Passing

Al Depope[†], Marco Mondelli[†], Matthew R. Robinson[†]

[†] Institute of Science and Technology Austria,
Klosterneuburg, Austria.



Agenda

1. Overview of GWAS

Agenda

1. Overview of GWAS
2. What AMP framework brings to the table? How can one make AMP scalable and stable for the GWAS inference task?

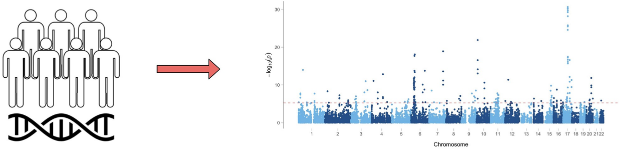
Agenda

1. Overview of GWAS
2. What AMP framework brings to the table? How can one make AMP scalable and stable for the GWAS inference task?
3. How does it compare to the existing state-of-the-art methods? What is the extent of applicability of gVAMP?

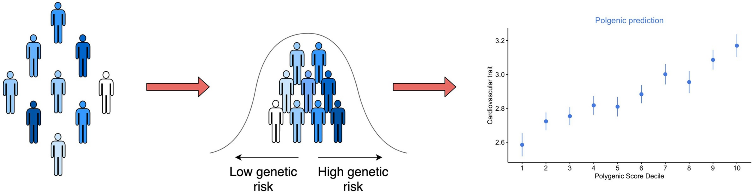
1. Genome-Wide Association Studies

1. Genome-Wide Association Studies

Step 1: Genome-wide association studies in adult populations from the UK Biobank



Step 2: Whole genome polygenic risk scores



Modelling genetic effects on a trait

Modelling genetic effects on a trait

- Bayesian Linear Regression for the **individual-level** model:

$$y_i = \langle \mathbf{x}_i, \beta \rangle + \epsilon_i \text{ for } i \in [N] = \{1, \dots, N\}$$

Modelling genetic effects on a trait

- Bayesian Linear Regression for the **individual-level** model:

$$y_i = \langle \mathbf{x}_i, \beta \rangle + \epsilon_i \text{ for } i \in [N] = \{1, \dots, N\} \quad \text{and}$$

$$\beta_j \sim (1-\lambda) \cdot \delta_0(\cdot) + \lambda \cdot \sum_{l=1}^{L} \pi_l \cdot \mathcal{N}(\cdot, 0, \sigma_l^2), \quad \epsilon_i \sim \mathcal{N}(0, {\gamma_\epsilon}^{-1})$$

Modelling genetic effects on a trait

- Bayesian Linear Regression for the **individual-level** model:

$$y_i = \langle \mathbf{x}_i, \beta \rangle + \epsilon_i \text{ for } i \in [N] = \{1, \dots, N\} \quad \text{and}$$

$$\beta_j \sim (1-\lambda)\cdot\delta_0(\cdot) + \lambda \cdot \sum_{l=1}^{L} \pi_l \cdot \mathcal{N}(\cdot, 0, \sigma_l^2), \quad \epsilon_i \sim \mathcal{N}(0, {\gamma_\epsilon}^{-1})$$

- Data format (genotype matrices normalized column-wise):

$$g_j^{(i)} = \begin{cases} 2, & aa \\ 1, & Aa \\ 0, & AA \end{cases}$$

Modelling genetic effects on a trait

- Bayesian Linear Regression for the **individual-level** model:

$$y_i = \langle \mathbf{x}_i, \beta \rangle + \epsilon_i \text{ for } i \in [N] = \{1, \dots, N\} \quad \text{and}$$

$$\beta_j \sim (1-\lambda) \cdot \delta_0(\cdot) + \lambda \cdot \sum_{l=1}^{L} \pi_l \cdot \mathcal{N}(\cdot, 0, \sigma_l^2), \quad \epsilon_i \sim \mathcal{N}(0, {\gamma_\epsilon}^{-1})$$

- Data format (genotype matrices normalized column-wise):

$$g_j^{(i)} = \begin{cases} 2, & aa \\ 1, & Aa \\ 0, & AA \end{cases} \implies \{0,1,2\}^{N \times P} \ni \mathbf{X} = \left.\underbrace{\begin{bmatrix} 1 & 2 & \dots & 0 \\ 0 & 0 & \dots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 2 & \dots & 2 \end{bmatrix}}_{\sim 10^6}\right\} \sim 10^5$$

2. (Vector) Approximate Message Passing

2. (Vector) Approximate Message Passing

- iterative algorithms that incorporate structural information about genetic effects

2. (Vector) Approximate Message Passing

- iterative algorithms that incorporate structural information about genetic effects
- linear models [Kab03, BM12, BM11, DMM09, KMS+12], generalized linear models [BKM+19, MLKZ20, Ran11, SR14, SC19] and low-rank matrix estimation

2. (Vector) Approximate Message Passing

- iterative algorithms that incorporate structural information about genetic effects
- linear models [Kab03, BM12, BM11, DMM09, KMS+12], generalized linear models [BKM+19, MLKZ20, Ran11, SR14, SC19] and low-rank matrix estimation
- achieves Bayes-optimal performance for some models [DM14, DJM13, BKM+19]

2. (Vector) Approximate Message Passing

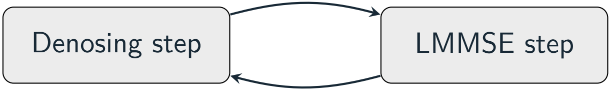
- iterative algorithms that incorporate structural information about genetic effects
- linear models [Kab03, BM12, BM11, DMM09, KMS+12], generalized linear models [BKM+19, MLKZ20, Ran11, SR14, SC19] and low-rank matrix estimation
- achieves Bayes-optimal performance for some models [DM14, DJM13, BKM+19]
- **statistical physics conjecture: AMP is optimal among polynomial-time algorithms**

2. (Vector) Approximate Message Passing

- iterative algorithms that incorporate structural information about genetic effects
- linear models [Kab03, BM12, BM11, DMM09, KMS+12], generalized linear models [BKM+19, MLKZ20, Ran11, SR14, SC19] and low-rank matrix estimation
- achieves Bayes-optimal performance for some models [DM14, DJM13, BKM+19]
- **statistical physics conjecture: AMP is optimal among polynomial-time algorithms**
- if $\mathbf{X}$ right-orthogonally invariant [RSF16, T17]: under: distributions of objects in the limit precisely characterized by a deterministic recursion called *state evolution*

2. (Vector) Approximate Message Passing

- iterative algorithms that incorporate structural information about genetic effects
- linear models [Kab03, BM12, BM11, DMM09, KMS+12], generalized linear models [BKM+19, MLKZ20, Ran11, SR14, SC19] and low-rank matrix estimation
- achieves Bayes-optimal performance for some models [DM14, DJM13, BKM+19]
- **statistical physics conjecture: AMP is optimal among polynomial-time algorithms**
- if **X** right-orthogonally invariant [RSF16, T17]: under: distributions of objects in the limit precisely characterized by a deterministic recursion called *state evolution*
-



learns and incorporates knowledge of effects distribution $p(\beta)$

takes into account correlation structure between genetic markers

genomic**VAMP**

1. *Filtering* the normalized genotype matrix for first-degree relative to reduce the correlation between rows ($\sim 400,000$ out of $460,000$ participants from the UK Biobank study)

genomic**VAMP**

1. *Filtering* the normalized genotype matrix for first-degree relative to reduce the correlation between rows ($\sim 400,000$ out of $460,000$ participants from the UK Biobank study)
2. *Auto-tuning* of $\gamma_{1,t}$ [FSR+17] combined with EM steps [VS12, FS17] that updates $p_t(\beta)$

genomic**VAMP**

1. *Filtering* the normalized genotype matrix for first-degree relative to reduce the correlation between rows ($\sim 400,000$ out of $460,000$ participants from the UK Biobank study)
2. *Auto-tuning* of $\gamma_{1,t}$ [FSR+17] combined with EM steps [VS12, FS17] that updates $p_t(\beta)$
3. Damping of denoised marker effects (momentum)

genomic**VAMP**

1. *Filtering* the normalized genotype matrix for first-degree relative to reduce the correlation between rows ($\sim 400,000$ out of $460,000$ participants from the UK Biobank study)
2. *Auto-tuning* of $\gamma_{1,t}$ [FSR+17] combined with EM steps [VS12, FS17] that updates $p_t(\beta)$
3. Damping of denoised marker effects (momentum)
4. Warm-start of conjugate gradients for LMMSE calculation [SD20]

genomic**VAMP**

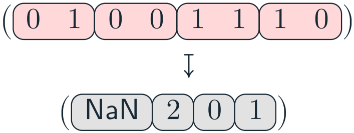
1. *Filtering* the normalized genotype matrix for first-degree relative to reduce the correlation between rows ($\sim 400,000$ out of $460,000$ participants from the UK Biobank study)
2. *Auto-tuning* of $\gamma_{1,t}$ [FSR+17] combined with EM steps [VS12, FS17] that updates $p_t(\beta)$
3. Damping of denoised marker effects (momentum)
4. Warm-start of conjugate gradients for LMMSE calculation [SD20]
5. Re-using Hutchinson estimator

genomic**VAMP**

1. *Filtering* the normalized genotype matrix for first-degree relative to reduce the correlation between rows ($\sim 400,000$ out of $460,000$ participants from the UK Biobank study)
2. *Auto-tuning* of $\gamma_{1,t}$ [FSR+17] combined with EM steps [VS12, FS17] that updates $p_t(\beta)$
3. Damping of denoised marker effects (momentum)
4. Warm-start of conjugate gradients for LMMSE calculation [SD20]
5. Re-using Hutchinson estimator
6. MPI $+$ OpenMP

genomic**VAMP**

1. *Filtering* the normalized genotype matrix for first-degree relative to reduce the correlation between rows ($\sim 400,000$ out of $460,000$ participants from the UK Biobank study)
2. *Auto-tuning* of $\gamma_{1,t}$ [FSR+17] combined with EM steps [VS12, FS17] that updates $p_t(\beta)$
3. Damping of denoised marker effects (momentum)
4. Warm-start of conjugate gradients for LMMSE calculation [SD20]
5. Re-using Hutchinson estimator
6. MPI + OpenMP
7. data streaming by using a lookup table + SIMD:



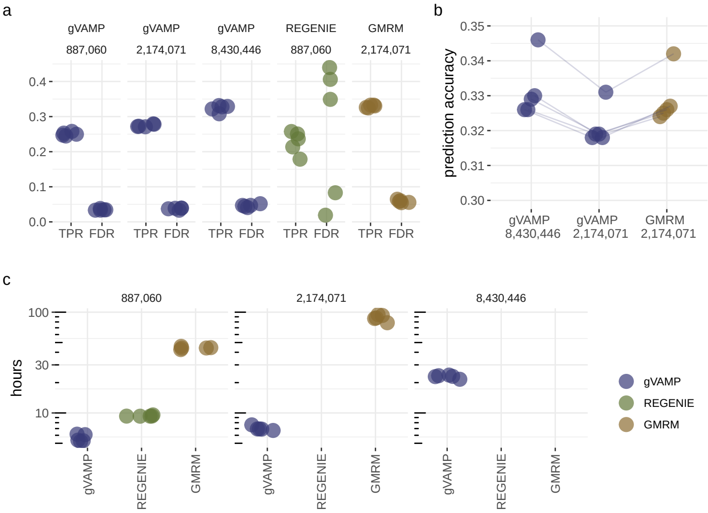
((0 1)(0 0)(1 1)(1 0))

↧

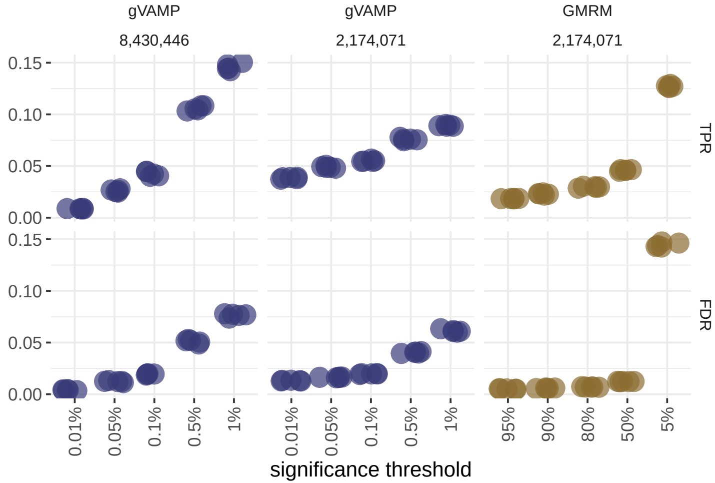
((NaN)(2)(0)(1))

3. Simulations: Association testing & prediction

3. Simulations: Association testing & prediction

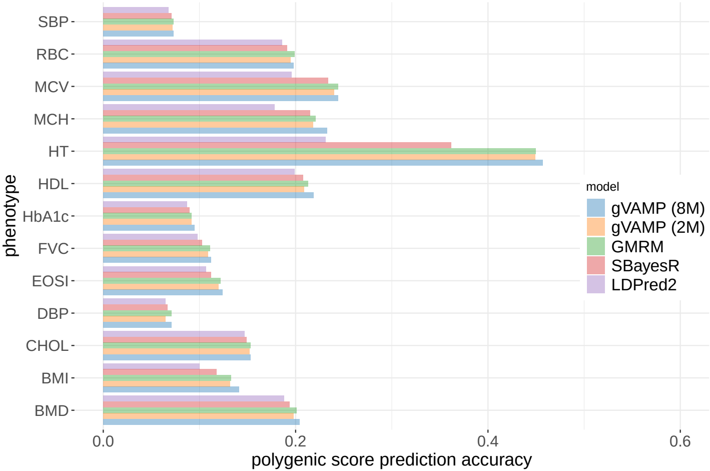


Fine mapping: gVAMP vs GMRM



Prediction accuracy

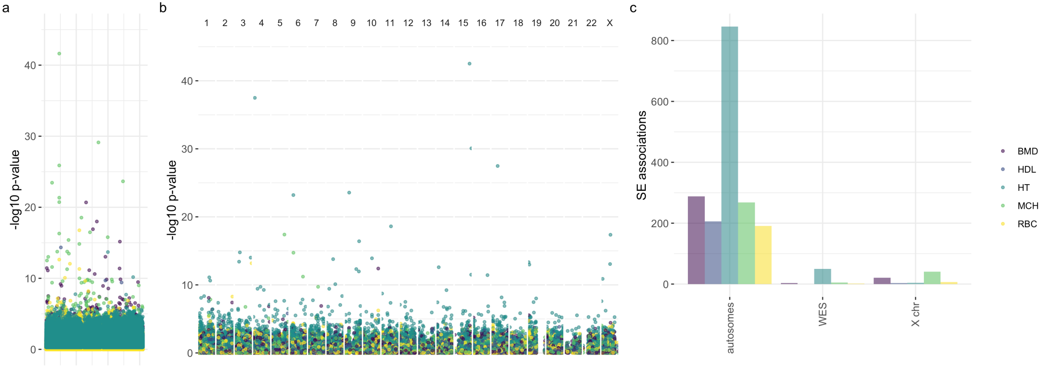
SBP: Systolic blood pressure
RBC: Red blood cell count
MCV: Mean corpuscular volume
MCH: Mean corpuscular haemoglobin
HT: Standing height
HDL: High density lipoprotein
HbA1c: Glycated haemoglobin
FVC: Forced vital capacity
EOSI: Eosinophill count
DBP: Diastolic blood pressure
CHOL: Cholesterol
BMI: Body mass index
BMD: Heel bone mineral density



Autosomal imputed data + X + WES analysis

Autosomal imputed data + X + WES analysis

- 60 genes where rare coding mutations significantly influence phenotype, and 76 associations localised to the single-locus level on chromosome X across five traits



Autosomal imputed data + X + WES analysis

Genes	Trait	Replicated
CALCR, CEP350, HSPA9, MOXD1 and SLC26A8	MCH	yes(Open Targets)
EFNA3, GRK5 and SCG2	BMD	EFNA3 - angiogenesis, GRK5 - linked to bone formation,...
COL4A4 and TFRC	RBC	recently discovered in large-scale meta-analysis
SHOX, TRIM68, TRAPPC2,...	HT	45/50 WES replicated

- $21, 3, 41, 7$, and 4 X chromosome associations that are conditional on everything else for BMD, HDL, MCH, RBC and HT
- novel associations: BMD:$20/21$, HDL:$3/3$, MCH:$40/41$, RBC:$5/7$ and HT:$0/4$

Summary & Future Directions

Summary & Future Directions

- gVAMP requires **less than a day** to model 8.4 **million** imputed genetic variants **jointly** in over $400,000$ UK Biobank participants. Other methods such as regenie, GMRM can not do this

Summary & Future Directions

- gVAMP requires **less than a day** to model 8.4 **million** imputed genetic variants **jointly** in over $400,000$ UK Biobank participants. Other methods such as regenie, GMRM can not do this
- exhibits **lower FPR**, **greater TPR** and is **more consistent** than regenie method

Summary & Future Directions

- gVAMP requires **less than a day** to model 8.4 **million** imputed genetic variants **jointly** in over $400,000$ UK Biobank participants. Other methods such as regenie, GMRM can not do this
- exhibits **lower FPR**, **greater TPR** and is **more consistent** than regenie method
- capable of analysing and doing **fine-mapping** for WES and X chromosome data jointly with imputed data (hundreds of associations)

Summary & Future Directions

- gVAMP requires **less than a day** to model 8.4 **million** imputed genetic variants **jointly** in over $400,000$ UK Biobank participants. Other methods such as regenie, GMRM can not do this
- exhibits **lower FPR**, **greater TPR** and is **more consistent** than regenie method
- capable of analysing and doing **fine-mapping** for WES and X chromosome data jointly with imputed data (hundreds of associations)

1. summary statistics & meta analysis models

Summary & Future Directions

- gVAMP requires **less than a day** to model 8.4 **million** imputed genetic variants **jointly** in over $400,000$ UK Biobank participants. Other methods such as regenie, GMRM can not do this
- exhibits **lower FPR**, **greater TPR** and is **more consistent** than regenie method
- capable of analysing and doing **fine-mapping** for WES and X chromosome data jointly with imputed data (hundreds of associations)

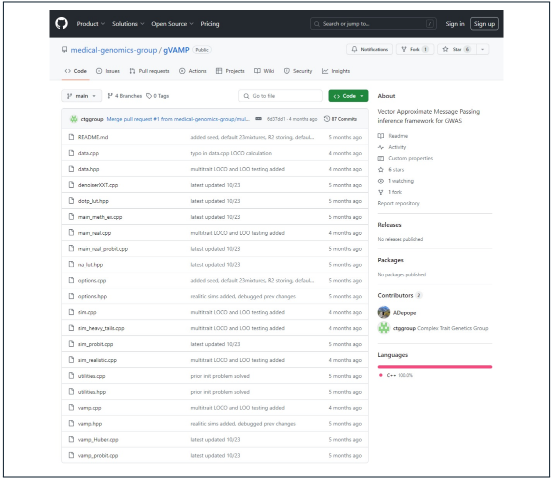
1. summary statistics & meta analysis models
2. time-to-event models

Summary & Future Directions

- gVAMP requires **less than a day** to model 8.4 **million** imputed genetic variants **jointly** in over $400,000$ UK Biobank participants. Other methods such as regenie, GMRM can not do this
- exhibits **lower FPR**, **greater TPR** and is **more consistent** than regenie method
- capable of analysing and doing **fine-mapping** for WES and X chromosome data jointly with imputed data (hundreds of associations)

1. summary statistics & meta analysis models
2. time-to-event models
3. using gVAMP on WGS data

gVAMP git repo: `https://github.com/medical-genomics-group/gVAMP`



medical-genomics-group / gVAMP Public

4 Branches 0 Tags

ctggroup Merge pull request #1 from medical-genomics-group/mul... 6d37dd1 · 4 months ago 87 Commits

README.md	added seed, default 23mixtures, R2 storing, defaul...	5 months ago
data.cpp	typo in data.cpp LOCO calculation	4 months ago
data.hpp	multitrait LOCO and LOO testing added	4 months ago
denoiserXXT.cpp	latest updated 10/23	5 months ago
dotp_lut.hpp	latest updated 10/23	5 months ago
main_meth_ex.cpp	latest updated 10/23	5 months ago
main_real.cpp	multitrait LOCO and LOO testing added	4 months ago
main_real_probit.cpp	latest updated 10/23	5 months ago
na_lut.hpp	latest updated 10/23	5 months ago
options.cpp	added seed, default 23mixtures, R2 storing, defaul...	5 months ago
options.hpp	realitic sims added, debugged prev changes	5 months ago
sim.cpp	multitrait LOCO and LOO testing added	4 months ago
sim_heavy_tails.cpp	multitrait LOCO and LOO testing added	4 months ago
sim_probit.cpp	latest updated 10/23	5 months ago
sim_realistic.cpp	multitrait LOCO and LOO testing added	4 months ago
utilities.cpp	prior init problem solved	5 months ago
utilities.hpp	prior init problem solved	5 months ago
vamp.cpp	multitrait LOCO and LOO testing added	4 months ago
vamp.hpp	realitic sims added, debugged prev changes	5 months ago
vamp_Huber.cpp	latest updated 10/23	5 months ago
vamp_probit.cpp	latest updated 10/23	5 months ago

About

Vector Approximate Message Passing inference framework for GWAS

Readme
Activity
Custom properties
6 stars
1 watching
1 fork
Report repository

Releases

No releases published

Packages

No packages published

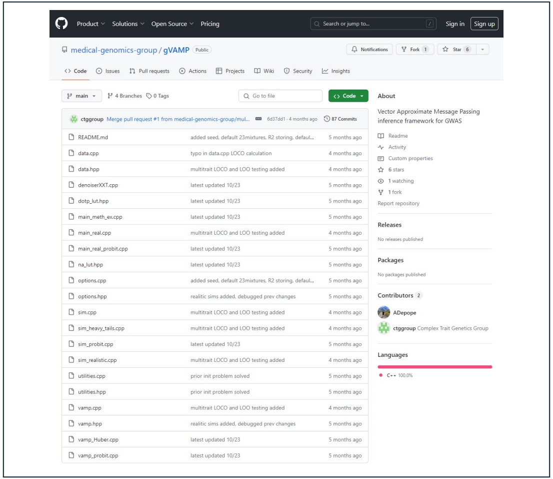
Contributors 2

ADepope
ctggroup Complex Trait Genetics Group

Languages

● C++ 100.0%

gVAMP git repo: `https://github.com/medical-genomics-group/gVAMP`



medical-genomics-group / gVAMP Public

4 Branches 0 Tags

ctggroup Merge pull request #1 from medical-genomics-group/mul... 6d37dd1 · 4 months ago 87 Commits

README.md	added seed, default 23mixtures, R2 storing, defaul...	5 months ago
data.cpp	typo in data.cpp LOCO calculation	4 months ago
data.hpp	multitrait LOCO and LOO testing added	4 months ago
denoiserXXT.cpp	latest updated 10/23	5 months ago
dotp_lut.hpp	latest updated 10/23	5 months ago
main_meth_ex.cpp	latest updated 10/23	5 months ago
main_real.cpp	multitrait LOCO and LOO testing added	4 months ago
main_real_probit.cpp	latest updated 10/23	5 months ago
na_lut.hpp	latest updated 10/23	5 months ago
options.cpp	added seed, default 23mixtures, R2 storing, defaul...	5 months ago
options.hpp	realitic sims added, debugged prev changes	5 months ago
sim.cpp	multitrait LOCO and LOO testing added	4 months ago
sim_heavy_tails.cpp	multitrait LOCO and LOO testing added	4 months ago
sim_probit.cpp	latest updated 10/23	5 months ago
sim_realistic.cpp	multitrait LOCO and LOO testing added	4 months ago
utilities.cpp	prior init problem solved	5 months ago
utilities.hpp	prior init problem solved	5 months ago
vamp.cpp	multitrait LOCO and LOO testing added	4 months ago
vamp.hpp	realitic sims added, debugged prev changes	5 months ago
vamp_Huber.cpp	latest updated 10/23	5 months ago
vamp_probit.cpp	latest updated 10/23	5 months ago

About

Vector Approximate Message Passing inference framework for GWAS

Readme
Activity
Custom properties
6 stars
1 watching
1 fork
Report repository

Releases

No releases published

Packages

No packages published

Contributors 2

ADepope
ctggroup Complex Trait Genetics Group

Languages

● C++ 100.0%

The End

Thanks for your attention!

Extra Slides

REGENIE overview

- Step 1: (Inference)
  - (Ridge regression): reads P markers in blocks of $B = 1000$ consecutive markers and

$$\mathbf{X} = \begin{array}{cccc} B & B & \dots & B \\ \left(\begin{array}{c} 0 \\ -1.414 \\ \vdots \\ -1.414 \end{array} \right. & \begin{array}{c} 4.242 \\ -1.414 \\ \vdots \\ 4.242 \end{array} & \begin{array}{c} \dots \\ \dots \\ \ddots \\ \dots \end{array} & \left. \begin{array}{c} -1.414 \\ 0 \\ \vdots \\ 1.414 \end{array} \right) \end{array}$$

    for $\tau \in \{\tau_1, \dots, \tau_J\}$ and block index b calculate $\hat{\beta}_{\tau,b} = (\mathbf{X}_b^T \mathbf{X}_b + \tau I)^{-1} \mathbf{X}_b^T y$
  - (Cross-validation): fitting model $y = W\alpha + \varepsilon$ using ridge with cross-validation, where W contains JM/B predictors stacked
- Step 2: Single-variant association testing using Leave-One-Chromosome-Out (LOCO) approach

Leave-One-Out (LOO) testing approach

- using VAMP we obtain estimators $\hat{\beta}$ for the effect sizes in a linear model

$$y = \mathbf{X}\beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 I_N).$$

- Leave-One-Out (LOO) p-values for the statistical test $H_0 : \beta_i = 0$ are calculated as a p-value from t-test for testing whether the slope of a regression line is zero when regressing

$$y^{(i)} := y - \mathbf{X}_{\setminus i}\hat{\beta}_{\setminus i} \quad \text{on} \quad \mathbf{X}_i$$

($\mathbf{X}_{\setminus i}$ = all columns of $\mathbf{X}$ except the i-th one)

Parallelization of the code

$$\mathbf{X} = \begin{pmatrix} 0 & 4.242 & \dots & -1.414 \\ -1.414 & -1.414 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ -1.414 & 4.242 & \dots & 1.414 \end{pmatrix}$$

- each MPI worker sees approximately equal number of consecutive columns ($\mathbf{X}$ is stored in a column-major format)
- $v \mapsto \mathbf{X}^T v$ operation is brought down to the level of single markers and combined with OpenMP reduction

- $u \mapsto \mathbf{X}u = \sum_{w=1}^{W} \mathbf{X}_w u_w \to$ $2 \cdot (W-1) \cdot N$ doubles sent for communication
- $\mathbf{X}$ is being streamed-in using a lookup table (no additional memory is required, performing 4 basic operations at once): $(\boxed{0 \quad 1}\boxed{0 \quad 0}\boxed{1 \quad 1}\boxed{1 \quad 0}) \mapsto$ $(\boxed{\text{NaN}}\boxed{2}\boxed{0}\boxed{1})$