Inference of Genetic Effects via Approximate Message Passing

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

† Institute of Science and Technology Austria, Klosterneuburg, Austria.
Agenda

1. What is a genome-wide association study (GWAS)?
Agenda

1. What is a genome-wide association study (GWAS)?
2. AMP overview. Making AMP approach scalable and stable for the GWAS inference task
Agenda

1. What is a genome-wide association study (GWAS)?
2. AMP overview. Making AMP approach scalable and stable for the GWAS inference task
3. Comparison to the state-of-the-art methods (regenie, GMRM)
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

- almost no limit to the amount of measured genetic variants (hundreds of millions; more genetic variants $\Rightarrow$ better generalization), but limited sample size
Modelling genetic effects on a trait

- almost no limit to the amount of measured genetic variants (hundreds of millions; more genetic variants $\implies$ better generalization), but limited sample size
- **Data format** (genotype matrices normalized column-wise): 

$$X_{ij} = \begin{cases} 
    2, & \text{aa} \\
    1, & \text{Aa} \\
    0, & \text{AA} 
\end{cases} \implies \{0, 1, 2\}^{N \times P} \ni X = \begin{bmatrix} 
    1 & 2 & \ldots & 0 \\
    0 & 0 & \ldots & 1 \\
    \vdots & \vdots & \ddots & \vdots \\
    0 & 2 & \ldots & 2 
\end{bmatrix} \sim 10^5$$
Modelling genetic effects on a trait

- almost no limit to the amount of measured genetic variants (hundreds of millions; more genetic variants $\Rightarrow$ better generalization), but limited sample size
- Data format (genotype matrices normalized column-wise):

$$X_{ij} = \begin{cases} 2, & aa \\ 1, & Aa \\ 0, & AA \end{cases} \quad \Rightarrow \quad X = \begin{bmatrix} 1.886 & 4.242 & \ldots & -0.472 \\ -0.472 & -1.414 & \ldots & 1.886 \\ \vdots & \vdots & \ddots & \vdots \\ -0.472 & 4.242 & \ldots & 4.243 \end{bmatrix} \sim 10^5$$
Modelling genetic effects on a trait

- almost no limit to the amount of measured genetic variants (hundreds of millions; more genetic variants $\implies$ better generalization), but limited sample size
- Data format (genotype matrices normalized column-wise): 

  $X_{ij} = \begin{cases} 
  2, & aa \\
  1, & Aa \\
  0, & AA 
\end{cases} \implies X = \begin{bmatrix} 
  1.886 & 4.242 & \ldots & -0.472 \\
  -0.472 & -1.414 & \ldots & 1.886 \\
  \vdots & \vdots & \ddots & \vdots \\
  -0.472 & 4.242 & \ldots & 4.243 
\end{bmatrix} \sim 10^5$

- Bayesian Linear Regression for the individual-level model:

  $y_i = \langle X(i,:), \beta \rangle + \epsilon_i$ for $i \in [N] = \{1, \ldots, N\}$
Modelling genetic effects on a trait

- almost no limit to the amount of measured genetic variants (hundreds of millions; more genetic variants $\implies$ better generalization), but limited sample size
- Data format (genotype matrices normalized column-wise):

\[
X_{ij} = \begin{cases} 
2, & \text{aa} \\
1, & \text{Aa} \\
0, & \text{AA}
\end{cases} \implies X = \begin{bmatrix}
1.886 & 4.242 & \ldots & -0.472 \\
-0.472 & -1.414 & \ldots & 1.886 \\
\vdots & \vdots & \ddots & \vdots \\
-0.472 & 4.242 & \ldots & 4.243
\end{bmatrix} \sim 10^5
\]

\[
\sim 10^6
\]

- Bayesian Linear Regression for the individual-level model:

\[
y_i = \langle X(i,:), \beta \rangle + \epsilon_i \text{ for } i \in [N] = \{1, \ldots, N\} \text{ and }
\]

\[
\beta_j \sim (1 - \lambda) \cdot \delta_0(\cdot) + \lambda \cdot \sum_{i=1}^{L} \pi_i \cdot N(\cdot, 0, \sigma_i^2), \quad \epsilon_i \sim N(0, \gamma \epsilon_{-1})
\]
Prior Work

- **P Phenotypes**
  - Common genotypes (array or sequencing)
  - Step 1 (Level 0) Ridge Regression within CV scheme, applied to block of B SNPs to reduce dimension
  - Step 1 (Level 1) Linear or Logistic Ridge Regression within CV scheme
  - P Phenotypes
    - 23 LOCO predictions for each phenotype
    - Imputed/Exome/CNV Genotypes
  - Step 2: Association testing
    - Single variant tests
    - Gene-based tests
    - Annotation files
    - GxE or GxG tests

[regenie, PLINK]

[LDpred2, SBayesR, SBayesRC, GMRM]
2. Approximate Message Passing

- family of iterative algorithms that incorporate structural information about genetic signal
2. Approximate Message Passing

- family of iterative algorithms that incorporate structural information about genetic signal
- linear models \([\text{Kab03, BM12, BM11, DMM09, KMS+12}],\) generalized linear models \([\text{BKM+19, MLKZ20, Ran11, SR14, SC19}],\) and low-rank matrix estimation
2. **Approximate Message Passing**

- family of iterative algorithms that incorporate structural information about genetic signal
- 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. (EM) Vector AMP

- Problem: correlation structure between columns of $X$?
2. (EM) Vector AMP

- Problem: correlation structure between columns of $X$?
- $X$ right-orthogonally invariant [RSF16, T17]: distributions of objects in the high-dimensional limit precisely characterized by a state evolution recursion
2. (EM) Vector AMP

- Problem: correlation structure between columns of \( X \)?
- \( X \) right-orthogonally invariant [RSF16, T17]: distributions of objects in the high-dimensional limit precisely characterized by a state evolution recursion

```
Denosing step
learns and incorporates knowledge of effects distribution \( p(\beta) \)

LMMSE step
takes into account correlation structure between genetic markers
```
genomic VAMP

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

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

2. Initialization matters (sparsity \(\sim 50k\) genetic positions, geometric sequence for prior mixture probabilities and variances)
1. **Filtering** the normalized genotype matrix for first-degree relatives to reduce the correlation between rows (≈ 400,000 out of 460,000 participants from UK Biobank study)

2. Initialization matters (sparsity ≈ 50k genetic positions, geometric sequence for prior mixture probabilities and variances)

3. **Auto-tuning** of denoising signal error precision [FSR+17] combined with EM steps [VS12, FS17] that updates estimate of $p(\beta)$

---

**genomicVAMP**
**genomic VAMP**

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

2. Initialization matters (sparsity $\sim 50k$ genetic positions, geometric sequence for prior mixture probabilities and variances)


4. Damping of denoised marker effects (momentum)
**genomic VAMP**

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

2. Initialization matters (sparsity $\sim 50k$ genetic positions, geometric sequence for prior mixture probabilities and variances)


4. Damping of denoised marker effects (momentum)

5. Warm-start of conjugate gradients for LMMSE calculation [SD20]
**genomic VAMP**

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

2. Initialization matters (sparsity ≈ 50k genetic positions, geometric sequence for prior mixture probabilities and variances)


4. Damping of denoised marker effects (momentum)

5. Warm-start of conjugate gradients for LMMSE calculation [SD20]

6. Re-using Hutchinson estimator
**Genomic VAMP**

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

2. Initialization matters (sparsity ∼ 50k genetic positions, geometric sequence for prior mixture probabilities and variances)


4. Damping of denoised marker effects (momentum)

5. Warm-start of conjugate gradients for LMMSE calculation [SD20]

6. Re-using Hutchinson estimator

7. MPI + OpenMP

8. Data processing by using a lookup table + SIMD:

   $\begin{bmatrix} 0 & 1 & 0 & 0 & 1 & 1 & 0 \end{bmatrix}$

   $\Downarrow$

   $\begin{bmatrix} \text{NaN} & 2 & 0 & 1 \end{bmatrix}$
genomicVAMP

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

2. Initialization matters (sparsity ∼ 50k genetic positions, geometric sequence for prior mixture probabilities and variances)


4. Damping of denoised marker effects (momentum)

5. Warm-start of conjugate gradients for LMMSE calculation [SD20]

6. Re-using Hutchinson estimator

7. MPI + OpenMP

8. data processing by using a lookup table + SIMD:

$$\begin{pmatrix}
0 & 1 & 0 & 0 & 1 & 1 & 1 & 1 & 0
\end{pmatrix} \downarrow
\begin{pmatrix}
\text{NaN} & 2 & 0 & 1
\end{pmatrix}$$
3. Association testing

\[ y^{(i)} := y - X_{\text{chr}(i)} \hat{\beta}_{\text{chr}(i)} \sim X(:, i) \]
Prediction accuracy for BMI (Body Mass Index)

- gVAMP 8M
- GMRM 2.17M
- SBayesR
- LDpred2

Out-of-sample $R^2$
Prediction accuracy for BMI (Body Mass Index)

\[ r := \frac{X^T y}{N} \] and approx. of correlation matrix, \( \hat{R} \)

- gVAMP 8M
- GMRM 2.17M
- SBayesR
- LDpred2

Out-of-sample \( R^2 \)
Prediction accuracy for BMI (Body Mass Index)

- gVAMP 8M
- GMRM 2.17M +12.7%
- SBayesR
- LDpred2

out-of-sample $R^2$
Prediction accuracy for BMI (Body Mass Index)

- gVAMP 8M: 0.14
- GMRM 2.17M: 0.13 + 6%
- SBayesR: 0.12
- LDpred2: 0.11
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: Eosinophil count
DBP: Diastolic blood pressure
CHOL: Cholesterol
BMI: Body mass index
BMD: Heel bone mineral density
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 cannot do this, exhibiting lower FDR, greater TPR than regenie. It is capable of analysing heterogeneous data (WES, X chromosome data).

1. Summary statistics & meta analysis
2. Time-to-event models
   \[
   \log y_i = \mu + \langle x_i, \beta \rangle + w_i + \alpha + K\alpha
   \]
3. Using GVAMP on WGS data (between 10^−12 M genetic variants)
4. Low-complexity alternatives to GVAMP?
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 FDR, greater TPR than regenie.
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 FDR, greater TPR than regenie.
- capable of analysing heterogeneous data (WES, X chromosome data).
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 FDR, greater TPR than regenie.
- capable of analysing heterogeneous data (WES, X chromosome data).

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 FDR, greater TPR than regenie
- capable of analysing heterogeneous data (WES, X chromosome data)

1. summary statistics & meta analysis models
   - access only to \( r := X^T y / N \) and an approximation of a correlation matrix, called \( \hat{R} \)
   - merging information from different databases/cohorts
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 FDR, greater TPR than regenie.
- Capable of analysing heterogeneous data (WES, X chromosome data).

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 FDR, greater TPR than regenie
- capable of analysing heterogeneous data (WES, X chromosome data)

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

\[ \log y_i = \mu + \langle x_i, \beta \rangle + \frac{w_i}{\alpha} + \frac{K}{\alpha} \]
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 FDR, greater TPR than regenie
- capable of analysing heterogeneous data (WES, X chromosome data)

1. summary statistics & meta analysis models
2. time-to-event models
   \[ \log y_i = \mu + \langle x_i, \beta \rangle + \frac{w_i}{\alpha} + \frac{K}{\alpha} \]
3. using gVAMP on WGS data (between 10 – 12M genetic variants)
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 FDR, greater TPR than regenie.
- Capable of analysing heterogeneous data (WES, X chromosome data).

1. Summary statistics & meta analysis models
2. Time-to-event models
   \[ \log y_i = \mu + \langle x_i, \beta \rangle + \frac{w_i}{\alpha} + \frac{K}{\alpha} \]
3. Using gVAMP on WGS data (between 10 - 12M genetic variants)
4. Low-complexity alternatives to VAMP?
gVAMP git repo: https://github.com/medical-genomics-group/gVAMP
gVAMP git repo: https://github.com/medical-genomics-group/gVAMP

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

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

for $\tau \in \{\tau_1, \ldots, \tau_J\}$ and block index $b$ calculate

$$\hat{\beta}_{\tau,b} = (X_b^T X_b + \tau I)^{-1} X_b^T y$$

• (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

\[ y = X\beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2_\epsilon I_N). \]

- 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

\[ y^{(i)} := y - X_{\setminus i}\hat{\beta}_{\setminus i} \quad \text{on} \quad X_i \]

($X_{\setminus i} = \text{all columns of } X \text{ except the i-th one}$)
Parallelization of the code

\[ X = \begin{pmatrix}
0 & 4.242 & \cdots & -1.414 \\
-1.414 & -1.414 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
-1.414 & 4.242 & \cdots & 1.414
\end{pmatrix} \]

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

\[ u \mapsto X u = \sum_{w=1}^{W} X_w u_w \rightarrow 2 \cdot (W - 1) \cdot N \text{ doubles sent for communication} \]

- \( X \) is being streamed-in using a lookup table (no additional memory is required, performing 4 basic operations at once):
  \[
  \begin{pmatrix}
  0 & 1 & 0 & 0 & 1 & 1 & 1 & 0
  \end{pmatrix}
  \mapsto
  \begin{pmatrix}
  \text{NaN} & 2 & 0 & 1
  \end{pmatrix}
  \]
Autosomal imputed data + X + WES analysis
3. Association testing
3. Association testing

![Graphs showing association testing results](image)

- **Graph a**: Comparison of different methods (gVAMP, REGENIE, GMRM) in terms of TPR and FDR across different genetic effect sizes.
- **Graph b**: Prediction accuracy analysis for the same methods, highlighting the performance variation.
- **Graph c**: Efficiency comparison in terms of hours required for inference across various sample sizes (887,060, 2,174,071, 8,430,446).

**Sources**
- Inference of Genetic Effects via Approximate Message Passing
Association testing: gVAMP vs GMRM

Inference of Genetic Effects via Approximate Message Passing