Efficient Bayesian mixed-model analysis increases association power in large cohorts

Loh et al. 2015. Nature Genetics

April 21, 2015

Mixed-models in GWAS

- Linear mixed models (LMM) have become increasingly popular for genome-wide association studies (GWAS)

Mixed-models in GWAS

- Linear mixed models (LMM) have become increasingly popular for genome-wide association studies (GWAS)
- Variance components can account for inherently hierarchical data

Mixed-models in GWAS

- Linear mixed models (LMM) have become increasingly popular for genome-wide association studies (GWAS)
- Variance components can account for inherently hierarchical data
  - Observations are, by definition, not independent
  - Dependence structure is not always known *a priori*
    - (sometimes cousins unwittingly wed)
    - Dependence structure can be estimated from the data

Mixed-models in GWAS

- Linear mixed models (LMM) have become increasingly popular for genome-wide association studies (GWAS)
- Variance components can account for inherently hierarchical data
  - Observations are, by definition, not independent
  - Dependence structure is not always known *a priori*
    - (sometimes cousins unwittingly wed)
    - Dependence structure can be estimated from the data
- LMMs can reduce the false discovery rate by accounting for pseudo-replication

Mixed-models in GWAS

- Linear mixed models (LMM) have become increasingly popular for genome-wide association studies (GWAS)
- Variance components can account for inherently hierarchical data
  - Observations are, by definition, not independent
  - Dependence structure is not always known *a priori*
    - (sometimes cousins unwittingly wed)
    - Dependence structure can be estimated from the data
- LMMs can reduce the false discovery rate by accounting for pseudo-replication
- LMMs can be computationally expensive, $O(MN^2)$ or $O(M^2N)$

Standard LMM

The standard LMM takes the form:

$$y_i = x_i\beta + \sigma_G^2 K + \sigma_E^2 I$$

Where y_i is the phenotype of individual i, xi are the covariates of interest (SNPs, age, etc.), σ_G^2 is the variance due to genetic similarity, K is the *genetic similarity matrix (GSM)*, E is the variance due to environment (chance), and I is the NxN identity matrix.

- The GSM (K) accounts for the hierarchical structure of the data and can be estimated in various ways
- The $\sigma_E^2 I$ term is similar to a typical regression error term

LMM Model Assumptions

- Assumes the ”infinitesimal model” of genetic effects

LMM Model Assumptions

- Assumes the "infinitesimal model" of genetic effects
  - Proposed by Fisher in 1930
  - Many loci contribute small additive effects for a phenotype
  - Appears consistent with some traits (size) but not others (Huntington's disease)

LMM Model Assumptions

- Assumes the "infinitesimal model" of genetic effects
  - Proposed by Fisher in 1930
  - Many loci contribute small additive effects for a phenotype
  - Appears consistent with some traits (size) but not others (Huntington's disease)
- Assumes correctly specified variance components

$$\sigma_G^2 \text{ and } \sigma_E^2$$

LMM Model Assumptions

- Assumes the "infinitesimal model" of genetic effects
  - Proposed by Fisher in 1930
  - Many loci contribute small additive effects for a phenotype
  - Appears consistent with some traits (size) but not others (Huntington's disease)
- Assumes correctly specified variance components

$$\sigma^2_G \text{ and } \sigma^2_E$$

- Assumes linearity in the parameters (β)

LMM Test Statistic

For efficiency the current methodology incorporates a 2 step process:

LMM Test Statistic

For efficiency the current methodology incorporates a 2 step process:

1. Fit the variance components via REML ($\sigma_G^2 K$ and $\sigma_E^2 I$).

LMM Test Statistic

For efficiency the current methodology incorporates a 2 step process:

1. Fit the variance components via REML ($\sigma_G^2 K$ and $\sigma_E^2 I$).
2. Use score test to compute test statistics for each SNP

$$\chi_1^2 = \frac{\left(x'_{SNP} V^{-1} y\right)^2}{x'_{SNP} V^{-1} x_{SNP}}$$

where V is the $cov(y) = \sigma_G^2 K + \sigma_E^2 I$

(Some models use simultaneous likelihood ratio tests to achieve exact statistics)

Proposed Model

Proposed model has the same basic components:

$$y_i = x_i\beta + \sigma_G^2 K + \sigma_E^2 I$$

Proposed Model

Proposed model has the same basic components:

$$y_i = x_i\beta + \sigma_G^2 K + \sigma_E^2 I$$

- Several "improvements":
  1. Adjust current score statistic with calibration factor, c.
  2. Test against "denoised" residuals for increased power
  3. Gaussian mixture extension to accommodate heterogeneous effect sizes

Proposed Model

Proposed model has the same basic components:

$$y_i = x_i\beta + \sigma_G^2 K + \sigma_E^2 I$$

- Several "improvements":
  1. Adjust current score statistic with calibration factor, c.
  2. Test against "denoised" residuals for increased power
  3. Gaussian mixture extension to accommodate heterogeneous effect sizes
- Not purely Bayesian – actual test statistic is frequentist

Calibration Factor: c_{inf}

The BOLT_LMM statistic (for infinitesimal model)

$$\chi_1^2 = \frac{\left(x'_{SNP} V_{LOCO}^{-1} y\right)^2}{c_{inf}}$$

where,

$$c_{inf} = \frac{\text{mean}\left(x'_{SNP} V_{LOCO}^{-1} y\right)^2}{\text{mean}\chi_1^2}$$

This statistic is similar to others (GRAMMAR-gamma and MASTOR) but avoids proximal contamination via the LOCO approach.

Key Insight

The key insight in this paper is this approach:

- The numerator from the χ^2 statistic is a scalar multiple of $\sigma_E^2 V_{LOCO}^{-1} y$

Key Insight

The key insight in this paper is this approach:

- The numerator from the χ^2 statistic is a scalar multiple of $\sigma_E^2 V_{LOCO}^{-1} y$
  - The test statistic uses residuals where other effects have been "conditioned out"

Key Insight

The key insight in this paper is this approach:

- The numerator from the χ^2 statistic is a scalar multiple of $\sigma_E^2 V_{LOCO}^{-1} y$
  - The test statistic uses residuals where other effects have been "conditioned out"
  - Less noise $=$ more power to detect a signal

Key Insight

The key insight in this paper is this approach:

- The numerator from the χ^2 statistic is a scalar multiple of $\sigma_E^2 V_{LOCO}^{-1} y$
    - The test statistic uses residuals where other effects have been “conditioned out”
    - Less noise $=$ more power to detect a signal
    - Can manipulate the model producing the residual without changing the test statistic

Key Insight

The key insight in this paper is this approach:

- The numerator from the χ^2 statistic is a scalar multiple of $\sigma_E^2 V_{LOCO}^{-1} y$
  - The test statistic uses residuals where other effects have been "conditioned out"
  - Less noise = more power to detect a signal
  - Can manipulate the model producing the residual without changing the test statistic
- The test statistic is still a quasi-likelihood score so is asymptotically χ^2

The Bayesian Part

Assume the effects, β have a distribution:

- Infinitesimal model:

$$\beta \sim N(0, \sigma_G^2)/M$$

where M is the proximally-controlled SNPs under evaluation

The Bayesian Part

Assume the effects, β have a distribution:

- Infinitesimal model:

$$\beta \sim N(0, \sigma_G^2)/M$$

  where M is the proximally-controlled SNPs under evaluation

- The general (heterogeneous effects) model utilizes a Gaussian mixture:

$$\beta \sim N(0, \sigma_{\beta,1}^2) \text{ with probability } p$$

$$\beta \sim N(0, \sigma_{\beta,2}^2) \text{ with probability } 1 - p$$

  If p is small and $\sigma_{\beta,2}^2$ is much smaller than $\sigma_{\beta,1}^2$ you will get a many βs of small (near 0) effect and a few of large effect.

From Bayesian to Frequentist

The method uses Bayesian model to obtain a frequentist statistic:

- First fit the Bayesian mixture model
- Calculate the posterior mean to obtain the residual
- Use the residual to calculate the χ^2 statistic

Computational Efficiency

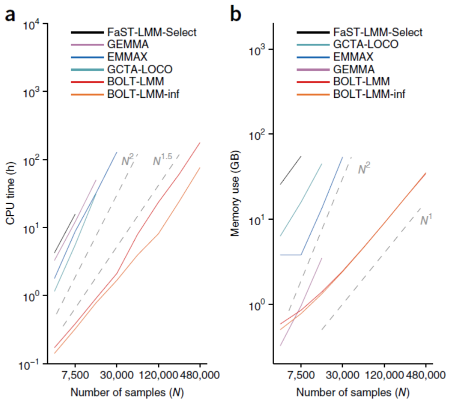
Table 1 Comparison of fast mixed-model association methods that model all SNPs

Method[a]	Requires $O(MN^2)$ time	Avoids proximal contamination	Models non-infinitesimal genetic architecture
EMMAX (ref. 3)	X		
FaST-LMM (ref. 5)	X[b]	X	
FaST-LMM-Select (refs. 9,11,15)	X[b]	X	X[c]
GEMMA (ref. 6)	X		
GRAMMAR-Gamma (ref. 10)	X[d]		
GCTA-LOCO (ref. 12)	X	X	
BOLT-LMM		X	X

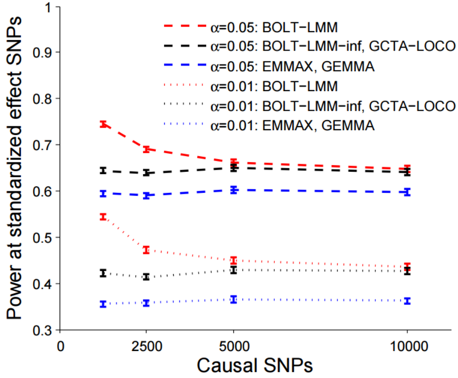
- Computes on $\sim O(MN)$ time
- Avoids proximal contamination by borrowing methods from GCTA-LOCO
- Accommodates non-infinitesimal genetic architecture

Computational Efficiency

- Fastest of the group*
- CPU time scales to $\sim O(MN^{1.5})$
- Memory use scales to $\sim O(MN)$



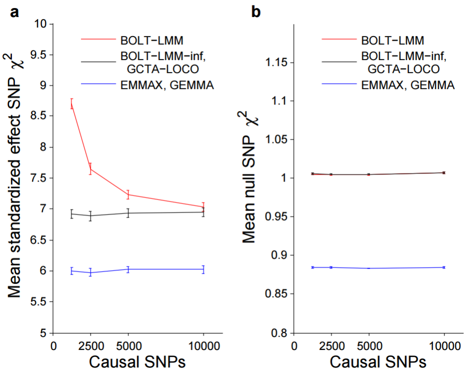
Simulations



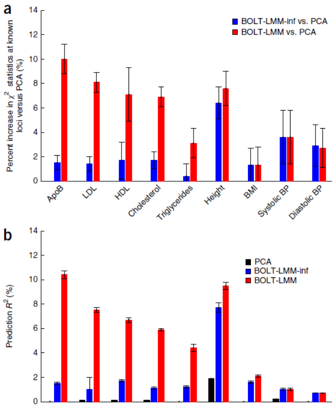
- Increased power especially when the number of causal SNPs is small
- Increased power is stable at various α levels (consistent with theory)
- BOLT-LMM is virtually identical to GCTA-LOCO under infinitesimal architecture assumption

Simulations

- Effective false positive control
- Large gains in power with small number of causal SNPs
- No associated increase in type I error with power increase



Women's Genome Health Study



- Similar increases in power in non-simulated data
- Power increases are not homogeneous (as predicted by theory)
- Substantial increases in predictive ability

Comments from the Peanut Gallery

- Not sure about mixing Bayesian and Frequentist paradigms
- Room to improve the modelling of genetic architecture
- Carry the Bayesian theme through to the test statistic to get a credible interval

Questions?