Efficient Bayesian mixed model analysis increases association power in large cohorts

10/20/14: ASHG 2014

Po-Ru Loh
Harvard School of Public Health

<table>
<thead>
<tr>
<th></th>
<th>Linear regression</th>
<th>Existing mixed model methods</th>
<th>New method: BOLT-LMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>$O(MN)$</td>
<td>$O(MN^2)$</td>
<td>$\approx O(MN^{1.5})$</td>
</tr>
<tr>
<td>Corrects for confounding?</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Power</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Mixed model association is hot

• **Linear mixed models (LMMs)** for GWAS have attracted intense research interest in the past few years

• High-profile computational speedups:
  – EMMAX (*Kang* et al. 2010 *Nat Gen*)
  – P3D/TASSEL (*Zhang* et al. 2010 *Nat Gen*)
  – FaST-LMM (*Lippert* et al. 2011 *Nat Meth*)
  – GEMMA (*Zhou* & *Stephens* 2012 *Nat Gen*)
  – GRAMMAR-Gamma (*Svishcheva* et al. 2012 *Nat Gen*)

• Extensions to standard LMM:
  – Multiple loci (*Segura* et al. 2012 *Nat Gen*)
  – Multiple traits (*Korte* et al. 2012 *Nat Gen*)
  – Sparse modeling (*Listgarten* et al. 2013 *Nat Gen*)
  – GCTA-LOCO (*Yang* et al. 2014 *Nat Gen*)

• See also: ASHG 2014 talk 169, Fusi & poster 1458S, Heckerman
Mixed model association corrects for confounding and increases power

<table>
<thead>
<tr>
<th></th>
<th>Linear regression</th>
<th>Existing mixed model methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>$O(MN)$</td>
<td>$O(MN^2)$</td>
</tr>
<tr>
<td>Corrects for confounding?</td>
<td>$\times$</td>
<td>$\checkmark$</td>
</tr>
<tr>
<td>Power</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

$M = \#$ SNPs
$N = \#$ samples
New mixed model method (BOLT-LMM) increases speed, further increases power

<table>
<thead>
<tr>
<th></th>
<th>Linear regression</th>
<th>Existing mixed model methods</th>
<th>New method: BOLT-LMM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time</td>
<td>$\mathcal{O}(MN)$</td>
<td>$\mathcal{O}(MN^2)$</td>
<td>$\approx \mathcal{O}(MN^{1.5})$</td>
</tr>
<tr>
<td>Corrects for confounding?</td>
<td>❌</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Power</td>
<td>⚡</td>
<td>⚡</td>
<td>⚡</td>
</tr>
</tbody>
</table>

Models non-infinitesimal genetic architectures
BOLT-LMM’s iterative algorithm increases speed; non-infinitesimal model increases power

**Speed:** New retrospective statistic allows rapid computation via iterative methods

\[
\frac{(x_m^T V^{-1} y)^2}{(V^{-1} y)^T \Theta^* (V^{-1} y)}
\]

**Power:** Flexible Bayesian prior on SNP effect sizes models genetic architectures with large-effect loci

“infinitesimal model”

“non-infinitesimal model”

Non-normal: Heavier tails
BOLT-LMM requires far less time and memory than existing methods

BOLT-LMM requires far less time and memory than existing methods.
BOLT-LMM can handle big data

<table>
<thead>
<tr>
<th>Number of samples (N)</th>
<th>BOLT-LMM: Time, memory</th>
<th>Existing methods: Time, memory</th>
</tr>
</thead>
<tbody>
<tr>
<td>60,000</td>
<td>5 hours 4.6 GB</td>
<td>~2 weeks &gt;60 GB</td>
</tr>
<tr>
<td>480,000</td>
<td>3 days 35 GB</td>
<td>~4 years &gt;3.5 TB</td>
</tr>
</tbody>
</table>

Benchmark data sets:
- Simulated genotypes generated from WTCCC2 samples
  \( M = 300K \) SNPs
- Simulated phenotypes with \( h^2_g = 0.2 \) SNP heritability
BOLT-LMM increases power in simulations (while controlling false positives)

**Increased $\chi^2$ stats at causal SNPs**

**Increased effective sample size**

Data:
- Real genotypes from N=15,633 WTCCC2 samples
- Simulated phenotypes with varying numbers of causal SNPs + environmental stratification
BOLT-LMM increases association power in WGHS phenotypes

Data:
Women’s Genome Health Study
N=23,294
European-ancestry samples
How does BOLT-LMM increase power?

Under the hood, Part 1
Mixed models increase power by conditioning on polygenic effects

• Joint modeling reveals association!

• Example:

Is phenotype $y$ associated with $x_1$?

Hard to tell...
Mixed models increase power by conditioning on polygenic effects

• **Joint modeling reveals association!**

• **Example:**

Is phenotype $y$ associated with $x_1$?

How about if we also have $x_2$?
Mixed models increase power by conditioning on polygenic effects

• Joint modeling reveals association!
• Example:

Is phenotype $y$ associated with $x_1$?

Knowledge of $x_2$ reveals assoc of $x_1$ through joint model

Note: In mixed model, $x_1$ (test SNP) is modeled as fixed effect; $x_2$ is modeled as random effect
Better joint model of phenotype $\Rightarrow$ better conditioning $\Rightarrow$ more power

**Infinitesimal model**
- Standard mixed model: all SNPs causal with normally distributed effect sizes:
  \[
  \beta \sim N\left(0, \frac{h^2}{M}\right)
  \]

**Non-infinitesimal model**
- Reality: Only a small fraction of SNPs causal with larger effects
- BOLT-LMM: SNP effect sizes modeled with mixture of two Gaussians

---

ASHG 2014 poster 1767S, Vilhjalmsson
How does BOLT-LMM increase speed?
(And what exactly does it compute?)

Under the hood, Part 2
Idea 1: New retrospective statistic + algorithm increases speed (for infinitesimal LMM)

\[
\frac{(x_m^TV^{-1}y)^2}{(V^{-1}y)^T\Theta^*(V^{-1}y)}
\]

Retrospective mixed model assoc statistic

- Compute quasi-likelihood score test similar to MASTOR:
  - Jakobsdottir & McPeek 2013 AJHG
- Calibrate using approach similar to GRAMMAR-Gamma
  - Svisheeva et al. 2012 NG

Fast iterative algorithm

- Replace expensive eigendecomposition with linear system solving
- No need to compute genetic relationship matrix (GRM) => save time, RAM
Idea 2: Bayesian extension of linear mixed model (non-inf. model) increases power

\[
\frac{(x_m^T V^{-1} y)^2}{(V^{-1} y)^T \Theta^* (V^{-1} y)} \quad \rightarrow \quad \frac{(x_m^T y_{\text{resid}})^2}{y_{\text{resid}}^T \Theta^* y_{\text{resid}}}
\]

Retrospective mixed model assoc statistic

Retrospective mixed model assoc statistic using Bayesian prior

- Compute fast variational approximation of posteriors
- Calibrate using LD score regression
  - Bulik-Sullivan et al. bioRxiv (under revision, Nat Genet)
  - ASHG 2014 talk 351, Finucane & poster 1787T, Bulik-Sullivan
New mixed model method (BOLT-LMM) increases speed, further increases power

<table>
<thead>
<tr>
<th></th>
<th>Linear regression</th>
<th>Existing mixed model methods</th>
<th>New method: BOLT-LMM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Time</strong></td>
<td>$O(MN)$</td>
<td>$O(MN^2)$</td>
<td>$\approx O(MN^{1.5})$</td>
</tr>
<tr>
<td><strong>Corrects for confounding?</strong></td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Power</strong></td>
<td>⚡</td>
<td>⚡</td>
<td>⚡</td>
</tr>
</tbody>
</table>

Models non-infinitesimal genetic architectures
Limitations and future directions

• Limitations
  – Loss of power from case-control ascertainment
    • *ASHG 2014 poster 1746S, Hayeck*
  – Potential breakdown of approximations in family studies, rare variant tests, non-human data sets

• Future directions
  – Fast heritability estimation
  – Multiple variance components
  – Multiple phenotypes
Acknowledgments

- Alkes Price
- Nick Patterson
- Bjarni Vilhjalmsson
- Hilary Finucane
- Sasha Gusev

- George Tucker
- Bonnie Berger

- Brendan Bulik-Sullivan
- Benjamin Neale
- Rany Salem

Google “BOLT-LMM”

Loh et al. bioRxiv (under revision, Nat Genet)