BOLT-LMM v1.0 User Manual

Po-Ru Loh

August 8, 2014

1 Overview

The BOLT-LMM software package computes statistics for association between phenotype and genotypes using a linear mixed model (LMM) [1]. By default, BOLT-LMM assumes a Bayesian mixture-of-normals prior for the random effect attributed to SNPs other than the one being tested. This model generalizes the standard "infinitesimal" mixed model used by existing mixed model association methods (e.g., EMMAX [2], FaST-LMM [3–6], GEMMA [7], GRAMMAR-Gamma [8], GCTA-LOCO [9]), providing an opportunity for increased power to detect associations while controlling false positives. Additionally, BOLT-LMM applies algorithmic advances to compute association statistics much faster than existing methods, both when using the Bayesian mixture model and when specialized to standard mixed model association.

1.1 Citing BOLT-LMM

BOLT-LMM is described in ref. [1]:

> Loh P-R, Tucker G, Bulik-Sullivan BK, Vilhjálmsson BJ, Finucane HK, Chasman DI, Ridker PM, Neale BM, Berger B, Patterson N, and Price AL. Efficient Bayesian mixed model analysis increases association power in large cohorts. Submitted and available on bioRxiv.

2 Installation

We provide a standalone (i.e., statically linked) 64-bit Linux executable, `bolt`, which we have tested on several Linux systems. We strongly recommend using this static executable because it is well-optimized and no further installation is required.

If you wish to compile your own version of the BOLT-LMM software from the source code (in the `src/` subdirectory), you will need to ensure that library dependencies are fulfilled and make appropriate modifications to the `Makefile`:

- Library dependencies:

  - BLAS/LAPACK numerical libraries. The speed of the BOLT-LMM software depends critically on the efficiency of the BLAS/LAPACK implementation it is linked against. We recommend the Intel Math Kernel Library (MKL) if available; otherwise, ATLAS may be a good alternative.
  - Boost C++ libraries. BOLT-LMM links against the Boost `program_options` and `iostreams` libraries, which need to be installed after downloading and unzipping Boost.

- `Makefile`: Paths to libraries need to be modified appropriately. Note that the released version of the `Makefile` does not set the flag `-DUSE_MKL_MALLOC`. This flag turns on the Intel MKL's fast memory manager (replacing calls to `_mm_malloc` with `mkl_malloc`), which may improve memory performance, but we have observed crashes on some systems when using `mkl_malloc`.

For reference, the provided `bolt` executable was created on the Harvard Medical School "Orchestra" research computing cluster using Intel Composer XE 12.0.5 (with Intel MKL 10.3.5) and the Boost C++ libraries 1.42 by invoking `make cluster=orchestra linking=static`.

2.1 Running BOLT-LMM

To run the `bolt` executable, simply invoke `./bolt` on the Linux command line (within the BOLT-LMM install directory) with parameters in the format `--optionName=optionValue`.

2.2 Example

The `example/` subdirectory contains a bash script `run_example.sh` that demonstrates basic use of BOLT-LMM on a small example data set.

2.3 Help

To get a list of basic options, run:

```
./bolt -h
```

To get a complete list of basic and advanced options, run:

```
./bolt --helpFull
```

3 Computing requirements

3.1 Operating system

At the current time we have only compiled and tested BOLT-LMM on Linux computing environments; however, the source code is available if you wish to try compiling BOLT-LMM for a different operating system.

3.2 Memory

For typical data sets (M, N $>$ 10,000), BOLT-LMM uses approximately MN/4 bytes of memory, where M is the number of SNPs and N is the number of individuals. More precisely:

- M = # of SNPs in `bim` file(s) that satisfy all of the conditions:
  - not listed in any `--exclude` file
  - passed QC filter for missingness
  - listed in `--modelSnps` file(s), if specified
- N = # of individuals in `fam` file and not listed in any `--remove` file (but pre-QC; i.e., N includes individuals filtered due to missing genotypes or covariates)

3.3 Running time

In practice, BOLT-LMM has a running time that scales roughly with $MN^{1.5}$. Our largest analyses of real data (M = 600K SNPs, N = 60K individuals) took $\approx$1 day using a single computational core. We have also tested BOLT-LMM on simulated data sets containing up to N = 480K individuals; for more details, please see the BOLT-LMM manuscript [1].

3.3.1 Multi-threading

On multi-core machines, running time can be reduced by invoking multi-threading using the `--numThreads` option.

4 Input/output file naming conventions

4.1 Automatic gzip [de]compression

BOLT-LMM assumes that input files ending in `.gz` are gzip-compressed and automatically decompresses them on-the-fly (i.e., without creating a temporary file). Similarly, BOLT-LMM writes gzip-compressed output to any output file ending in `.gz`.

4.2 Arrays of input files and covariates

Arrays of sequentially-numbered input files and covariates can be specified by the shorthand `{i:j}`. For example,

```
data.chr{1:22}.bim
```

is interpreted as the list of files

```
data.chr1.bim, data.chr2.bim, ..., data.chr22.bim
```

5 Input

5.1 Genotypes

BOLT-LMM takes genotype input is in PLINK [10] binary format (`bed/bim/fam`). For file conversion and data manipulation in general, we highly recommend the PLINK 2 project, which is providing a comprehensive, much more efficient update to PLINK [11].

If all genotypes are contained in a single `bed/bim/fam` file triple with the same file prefix, you may simply use the command line option `--bfile=prefix`. Genotypes may also be split into multiple `bed` and `bim` files containing consecutive sets of SNPs (e.g., one `bed/bim` file pair per chromosome) either by using multiple `--bed` and `--bim` invocations or by using the file array shorthand described above (e.g., `--bim=data.chr{1:22}.bim`).

5.1.1 Reference genetic maps

The BOLT-LMM package includes reference maps that you can use to interpolate genetic map coordinates from SNP physical (base pair) positions in the event that your PLINK `bim` file does not contain genetic coordinates (in units of morgans). To use a reference map, use the option

```
--geneticMapFile=tables/genetic_map_hg##.txt.gz
```

selecting the build (hg17, hg18, or hg19) corresponding to the physical coordinates of your `bim` file. You may use the `--geneticMapFile` option even if your PLINK `bim` file does contain genetic coordinates; in this case, the genetic coordinates in the `bim` file will be ignored, and interpolated coordinates will be used instead.

5.1.2 Imputed dosages

The current BOLT-LMM release does not support imputed dosage data, but we plan to include support for imputed dosages in a future release. Please contact us if you are interested in using this functionality; we will try to provide support for the most popular dosage data format.

5.2 Phenotypes

Phenotypes may be specified in either of two ways:

- `--phenoUseFam`: This option tells BOLT-LMM to use the last (6th) column of the `fam` file as the phenotypes. This column must be numeric, so case-control phenotypes should be `0,1` coded and missing values should be indicated with `-9`.

- `--phenoFile` and `--phenoCol`: Alternatively, phenotypes may be provided in a separate whitespace-delimited file (specified with `--phenoFile`) with the first line containing column headers and subsequent lines containing records, one per individual. The first two columns must be `FID` and `IID` (the PLINK identifiers of an individual). Any number of columns may follow; the column containing the phenotype to analyze is specified with

`--phenoCol`. Values of `-9` and `-NA` are interpreted as missing data. All other values in the column should be numeric. The records in lines following the header line need not be in sorted order and need not match the individuals in the genotype data (i.e., `fam` file); BOLT-LMM will analyze only the individuals in the intersection of the genotype and phenotype files and will output a warning if these sets do not match.

5.3 Covariates

Covariate data may be specified in a file (`--covarFile`) with the same format as the alternate phenotype file described above. (The same file may be used for both phenotypes and covariates.) Each covariate to be used must be specified using either a `--covarCol` (for categorical covariates) or a `--qCovarCol` (for quantitative covariates) option. Categorical covariate values are allowed to be any text strings not containing whitespace; each unique text string in a column corresponds to a category. Quantitative covariate values must be numeric (with the exception of `NA`). In either case, values of `-9` and `-NA` are interpreted as missing data. If groups of covariates of the same type are numbered sequentially, they may be specified using array shorthand (e.g., `--qCovarCol=PC{1:10}` for columns `PC1, PC2, ..., PC10`).

5.4 Missing data treatment

Individuals with missing phenotypes are ignored. By default, individuals with any missing covariates are also ignored; this approach is commonly used and referred to as "complete case analysis." As an alternative, BOLT-LMM also offers the "missing indicator method" (via the `--covarUseMissingIndic` option), which adds indicator variables demarcating missing status as additional covariates. Missing genotypes are replaced with per-SNP averages.

5.5 QC

BOLT-LMM automatically filters out SNPs and individuals with missing rates exceeding thresholds of 10%. These thresholds may be modified using the options `--maxMissingPerSnp` and `--maxMissingPerIndiv`. Note that BOLT-LMM does **not** perform automatic filtering based on minor allele frequency or deviation from Hardy-Weinberg equilibrium. Allele frequency and missingness of each SNP are included in the BOLT-LMM output, however, and we recommend checking these values and Hardy-Weinberg p-values (which are easily computed using PLINK 2 `--hardy`) when following up on significant associations.

5.6 User-specified filters

Individual to remove from the analysis may be specified in one or more `--remove` files listing `FID` and `IID`s (one individual per line). Similarly, SNPs to exclude from the analysis may be specified in one ore more `--exclude` files listing SNP IDs (typically rs numbers).

6 Association analysis

6.1 Mixed model association tests

BOLT-LMM computes two association statistics, $\chi^2_{\text{BOLT-LMM}}$ and $\chi^2_{\text{BOLT-LMM-inf}}$, described in detail in our manuscript [1].

- **BOLT-LMM: Association test on residuals from Bayesian modeling using a mixture-of-normals prior on SNP effect sizes.** This approach can fit "non-infinitesimal" traits with loci having moderate to large effects, allowing increased association power.
- **BOLT-LMM-inf: Standard (infinitesimal) mixed model association.** This statistic approximates the standard approach used by existing software.

6.2 BOLT-LMM mixed model association options

The BOLT-LMM software offers the following options for mixed model analysis:

- `--lmm`: Performs default BOLT-LMM analysis, which consists of (1a) estimating heritability parameters, (1b) computing the BOLT-LMM-inf statistic, (2a) estimating Gaussian mixture parameters, and (2b) computing the BOLT-LMM statistic *only if an increase in power is expected*. If BOLT-LMM determines based on cross-validation that the non-infinitesimal model is likely to yield no increase in power, the BOLT-LMM (Bayesian) mixed model statistic is not computed.
- `--lmmInfOnly`: Computes only infinitesimal mixed model association statistics (i.e., steps 1a and 1b).
- `--lmmForceNonInf`: Computes both the BOLT-LMM-inf and BOLT-LMM statistics *regardless of whether or not an increase in power is expected* from the latter.

6.2.1 Reference LD Score tables

A table of reference LD Scores is needed to calibrate the BOLT-LMM statistic; this table is provided with the BOLT-LMM package and can be specified using the option

```
--LDscoresFile=tables/LDSCORE.1000G_EUR.tab.gz
```

6.2.2 Restricting SNPs used in the mixed model

If millions of SNPs are available from imputation, we suggest including $\leq$1 million SNPs at a time in the mixed model (using the `--modelSnps` option) when performing association analysis. Using an LD pruned set of $\leq$1 million SNPs should achieve near-optimal power and correction for confounding while reducing computational cost and improving convergence. Note that even when a file of `--modelSnps` is specified, all SNPs in the genotype data are still tested for association; only the random effects in the mixed model are restricted to the `--modelSnps`. Also

note that BOLT-LMM automatically performs leave-one-chromosome-out (LOCO) analysis, leaving out SNPs from the chromosome containing the SNP being tested in order to avoid proximal contamination [4, 9].

6.3 Standard linear regression

Setting the `--verboseStats` flag will output standard linear regression chi-square statistics and p-values in additional output columns `CHISQ_LINREG` and `P_LINREG`. Note that unlike mixed model association, linear regression is susceptible to population stratification, so you may wish to include principal components (computed using other software) as covariates when performing linear regression.

7 Output

Association statistics are output in a tab-delimited file with the following fields, one line per SNP:

- `SNP`: rs number or ID string
- `CHR`: chromosome
- `BP`: physical (base pair) position
- `GENPOS`: genetic position either from `bim` file or interpolated from genetic map
- `ALLELE1`: first allele in `bim` file (usually the minor allele), used as the effect allele
- `ALLELE0`: second allele in `bim` file, used as the reference allele
- `A1FREQ`: frequency of first allele
- `F_MISS`: fraction of individuals with missing genotype at this SNP
- `BETA`: effect size from BOLT-LMM approximation to infinitesimal mixed model
- `SE`: standard error of effect size
- `P_BOLT_LMM_INF`: infinitesimal mixed model association test p-value
- `P_BOLT_LMM`: non-infinitesimal mixed model association test p-value

7.1 Optional additional output

To output chi-square statistics for all association tests, set the `--verboseStats` flag.

7.2 Diagnostics and logging

As BOLT-LMM proceeds, output is written to the terminal (`stdout` and `stderr`). If you wish to save this output while simultaneously viewing it on the command line, you may do so using

```
./bolt [... list of options ...] 2>&1 | tee output.log
```

8 Change log

Version 1.0: Initial release.

9 Website and contact info

Software updates will be posted at the following website:

```
http://www.hsph.harvard.edu/alkes-price/software/
```

If you have comments or questions about the BOLT-LMM software, please contact Po-Ru Loh, `loh@hsph.harvard.edu`.

10 Software copyright notice agreement

References

1. Loh, P.-R. *et al.* Efficient bayesian mixed model analysis increases association power in large cohorts. *bioRxiv* (2014).

2. Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics* **42**, 348–354 (2010).

3. Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nature Methods* **8**, 833–835 (2011).

4. Listgarten, J. *et al.* Improved linear mixed models for genome-wide association studies. *Nature Methods* **9**, 525–526 (2012).

5. Listgarten, J., Lippert, C. & Heckerman, D. FaST-LMM-Select for addressing confounding from spatial structure and rare variants. *Nature Genetics* **45**, 470–471 (2013).

6. Lippert, C. *et al.* The benefits of selecting phenotype-specific variants for applications of mixed models in genomics. *Scientific Reports* **3** (2013).

7. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nature Genetics* **44**, 821–824 (2012).

8. Svishcheva, G. R., Axenovich, T. I., Belonogova, N. M., van Duijn, C. M. & Aulchenko, Y. S. Rapid variance components-based method for whole-genome association analysis. *Nature Genetics* (2012).

9. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nature Genetics* **46**, 100–106 (2014).

10. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81**, 559–575 (2007).

11. Purcell, S. & Chang, C. PLINK 1.9. URL `https://www.cog-genomics.org/plink2`.