BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 14 2012, pages 1818–1822
doi:10.1093/bioinformatics/bts291

Genome analysis Advance Access publication May 15, 2012

# Fast computation for genome-wide association studies using boosted one-step statistics

Arend Voorman[1,*], Ken Rice[1] and Thomas Lumley[2]

[1]Department of Biostatistics, University of Washington, Seattle, WA 98195, USA and [2]Department of Statistics, University of Auckland, Auckland 1142, New Zealand

Associate Editor: Alex Bateman

**ABSTRACT**

**Motivation:** Statistical analyses of genome-wide association studies (GWAS) require fitting large numbers of very similar regression models, each with low statistical power. Taking advantage of repeated observations or correlated phenotypes can increase this statistical power, but fitting the more complicated models required can make computation impractical.

**Results:** In this article, we present simple methods that capitalize on the structure inherent in GWAS studies to dramatically speed up computation for a wide variety of problems, with a special focus on methods for correlated phenotypes.

**Availability:** The R package 'boss' is available on the Comprehensive R Archive Network (CRAN) at http://cran.r-project.org/web/packages/boss/

**Contact:** voorma@u.washington.edu

**Supplementary information:** Supplementary data are available at *Bioinformatics* online.

Received on February 3, 2012; revised on April 17, 2012; accepted on May 10, 2012

## 1 INTRODUCTION

In analysis of genome wide association studies (GWAS), it is standard to fit very large numbers of very similar regression models. A general form of commonly used models is

$$y_i \sim \beta_X \mathbf{X}^T + \beta_g^T g + \epsilon_i,$$

where $1 \leq i \leq n$ indexes subjects in the study, $y_i$ is the subject's outcome of interest, $\mathbf{X}_i$ is a vector of subject-specific adjustment variables, $g_i$ is the set of genotype-dependent variables and $\epsilon_i$ is an error term. The genotype-dependent variables $g_i$ vary from model to model, within one analysis, but the adjustment variables $X_i$ remain identical within each analysis. If the errors are uncorrelated, standard ordinary least squares (OLS) or generalized linear models (GLM) efficiently estimate genetic effects $\beta_g$. However, when the errors are correlated notably, greater efficiency can be obtained, by incorporating this structure into the model. For instance, if repeated measures on an individual are available, either generalized estimating equations (GEE) or mixed models can increase power relative to an analysis that uses a single phenotype measure for each individual (Liang and Zeger, 1993). Similarly, when subjects are related, kinship can be incorporated into a mixed model to increase power (Kang *et al.*, 2008). A single GEE or mixed model analysis involves negligible computing effort, but the number of analyses carried out often extends into several millions, meaning more complicated GEE and mixed models computation with standard methods can be infeasible.

Some computational speedups are available, by capitalizing on the similarity of each regression, meaning one can obtain genome-wide estimates at a substantially reduced computational cost. For example, Auchenko *et al.* (2007) introduced an approximation to the Maximum Likelihood Estimate for mixed models which is currently implemented in ProbABEL. In their approach, both phenotypic correlation and non-genetic effects are estimated once and assumed constant across all models in an analysis. When the inclusion of genotype does not change these parameters, the inference is close to that obtained from the full fit, but can be misleadingly inaccurate when genotype is correlated with non-genetic covariates.

In another example, rather than fixing both non-genetic effects and variance structure, Kang *et al*. (2008, 2001) proposed mixed model methods where the variance structure of phenotype is fixed over all markers in which case simple modifications to Generalized Least Squares procedures can give fast computation of the results Meyer and Tier, 2011. This is more complicated to implement and slower than the ProbABEL approximation, but makes less assumptions about the nature of the genetic association. However, one drawback of a mixed model approach is that results may be invalid if either the correlation structure or the covariate–phenotype relationship is not well specified.

If the phenotype correlation is due to repeated observations on an individual, then GEE models provide powerful testing, which does not require us to correctly specify either the correlation structure or the phenotype–covariate relationships, that may be complicated (Diggle, 2002). However, there are no software packages that can currently implement this on a genome-wide scale in reasonable amounts of time.

In this article, we explore two approaches that substantially reduce the computation required for a broad class of statistical models. Each exploits two important features of GWAS analyses: (i) few covariates are different in each regression and (ii) little variation in phenotype is explained by genotype at a particular locus. Our first approach exploits symmetry in the estimation of single marker models. For ordinary least squares, it is ~20 times faster than current implementations, and for mixed models it can reduce computation by a factor of a few hundred.

Our second approach is based on matrix decompositions and applies to a broader class of models, including those with multiple markers, and allows model-robust variance estimation. In the case of linear mixed models, it aligns with a proposal by Meyer and

*To whom correspondence should be addressed.

Tier (2011). The speedup for this approach is not large for GEE and GLM models with few covariates or observations, but in our tests estimation in GEE models was 60–80 times faster than currently implemented GEE packages in R. Although these computational approaches are approximations, estimation in practice is almost identical to traditional methods, as we demonstrate. The two approaches have the unifying feature that they are based on classical one-step approximations, suggesting the name boosted one-step statistics (BOSS).

## 2 APPROACH

First, consider estimation of the effect of a single genotype variable $g=[g_1 \ldots g_n]^T$ on a phenotype $Y=[y_1 \ldots y_n]^T$ adjusted for $p$ subject-specific covariates $x_j=[x_{j1} \ldots x_{jn}]^T, j=1 \ldots p$. Denote the model matrix $\mathbf{X}_g=[1_n, x_1, \ldots x_p, g]=[\mathbf{X}, g]$. If the outcomes are independent and continuous, we get coefficients by computing the OLS estimate

$$\beta=(\mathbf{X}_g^T\mathbf{X}_g)^{-1}\mathbf{X}_g^T Y.$$

This is sufficiently simple to perform on a genome-wide scale, and there is software available to do so (Aulchenko *et al.*, 2007; Purcell *et al.*, 2007). For more complicated models, such as mixed models, GLM or GEE, we can estimate the coefficients through iteratively re-weighted least squares (IWLS) that repeatedly calculates

$$\beta^{k+1}=(\mathbf{X}_g^T W^k \mathbf{X}_g)^{-1}\mathbf{X}_g^T W^k \tilde{Y}^k,$$

where $W^k$ is a matrix of weights and $\tilde{Y}^k$ is a modified phenotype at the $k$th iteration of this procedure. The details of how this arises and the exact scheme by which these are updated are not important for this discussion, and the interested reader can refer e.g. Diggle (2002) and McCullagh and Nelder (1989). It suffices to note that, at each iteration, $W^k$ and $Y^k$ differ from $W^{k-1}$ and $Y^{k-1}$ in ways that only depend on the fitted values at the current iteration. Since it can be assumed that genotype has little effect on these fitted values, we can estimate $W^k$ and $\tilde{Y}^k$ in the absence of genotype and substantially lower the number of iterations required to converge. In fact, for the models we consider one iteration of the IWLS algorithm is typically sufficient. This can be considered an instance of a so-called 'one-step approximation', which are well known to be asymptotically equivalent to fully converged solutions (Lehmann and Casella, 1998; Lipsitz *et al.*, 1994). Here, since the one step is based on a model with nearly identical fitted values to the model of interest, the practical performance is superior to classically implemented one-step estimators, as we demonstrate in Section 4. The accuracy decreases with the strength of the genetic effect, but our simulations suggest that this loss in accuracy occurs far beyond the threshold of genome-wide significance. That is when there are genome-wide significant results, both BOSS and traditional tests will identify them, but the exact size of the effect may differ slightly when it is very large.

This procedure is already used in mixed models, although is not identified as such. There $W$ is proportional to the inverse variance–covariance matrix of phenotype induced by random effects. The details of this model are explored by (Kang *et al.*, 2010). BOSS applies the same idea to GLM and GEE.

The use of one-step estimators takes advantage of the fact that genotype at a particular loci explains little of the variation in complicated phenotypes, but does not use the fact that only one variable changes in each regression. Sections 3 and 4 outline two methods used by BOSS which perform one-step estimation quickly by avoiding calculations that do not involve $g$. The first applies to single marker models and allows only model-based standard errors. The second is slower, but applies more generally and allows computation of robust test statistics.

To make the remaining discussion simpler, we first express this weighted least-squares problem as an OLS problem. If we decompose the matrix $W$ as a square root, i.e. $W=D^T D$, we see that the above procedure is equivalent to regressing $D\mathbf{X}_g=[D\mathbf{X}, Dg]$ on $D\tilde{Y}$. In general, $W$ is an $n \times n$ matrix, making this factorization difficult, slow and memory-intensive. However, $W$ will often have structure. For GLM it is a diagonal matrix, so we only need to store a vector of length $n$. For GEE and mixed models, $W$ is typically composed of small blocks corresponding to independent clusters of observations, so we need only to store and factorize these blocks. For some mixed model problems, $W$ may be dense, and difficult to factorize in this way, but as shown in the Supplementary Appendix we can reduce the effective dimension to the number of random effects terms using the spectral decomposition. The factorization will need to be performed only once, but the multiplication $Dg$ will need to be performed for each marker. So, where the weight matrix is sufficiently simple to factorize and handle, speed in one-step estimation will come from fast implementations of OLS.

## 3 METHODS

In this section, we describe fast implementations for one-step estimation. In the previous section, we demonstrated how this can be phrased in terms of fast OLS, so we will restrict our discussion to that case.

### 3.1 A fast algorithm for GWAS with single markers

Typically, we model a quantitative trait $Y$ as an outcome variable with genotype $g$ and covariates $\mathbf{X}$ as predictors. That is we form the model

$$Y=\beta_x\mathbf{X}+\beta_g g+\epsilon,$$

where $\epsilon \sim (0, \mathrm{var}[Y|\mathbf{X}+g])$. With model-based approaches, we assume that $\mathrm{var}[Y|\mathbf{X}+g]$ does not change with $\mathbf{X}$ and $g$. After fitting the model via OLS, the output is used to test the hypothesis $H_0^a: \beta_g=0$.

However, if we swap $Y$ and $g$ to form the model

$$g=\gamma_x\mathbf{X}+\gamma_y Y+\xi.$$

The hypothesis test $H_0^b: \gamma_y=0$ is equivalent to the previous test that $\beta_g=0$ and yields identical $P$-values when one uses model-based standard errors. (A straightforward proof of this is given in the Supplementary Appendix.) While this formulation may seem unnatural, it leads to simplified calculations. Almost all of the computational effort involved in regression is spent on the matrix of covariates, and using $g$ as an outcome allows us to do this work only once, before we examine genotype.

To use this result in a GWAS setting, we first denote the new matrix of predictors $\mathbf{X}_y=[\mathbf{X}, y]$ and $\tilde{G}$ to be an $n \times m$ matrix of genotypes for the entire GWAS (in practice we may break this up into more manageable chunks). The test statistics can then be computed by:

$$\hat{\gamma}=(\mathbf{X}_y^T\mathbf{X}_y)^{-1}\mathbf{X}_y^T\tilde{G}$$

$$R=\tilde{G}-\mathbf{X}_y\hat{\gamma}$$

$$\hat{\sigma}^2=\mathbf{1}_n^T R\cdot R/(n-p)$$

$$T=\hat{\gamma}_y^2/(\mathbf{X}_y^T\mathbf{X}_y)_{y,y}^{-1},$$

where $R \cdot R$ indicates element-wise multiplication. Here, $T$ is a vector of $\chi^2$ test statistics, identical to those computed with traditional methods.

It is notable that this technique yields test statistics, but not regression coefficients. In typical GWAS, only a tiny fraction of these coefficients might be of interest, and they could be computed without special techniques, at minor computational cost. However, regression coefficients could be approximated, genome wide, by using the identity (Peng *et al.*, 2009)

$$\hat{\beta}_g = \hat{\gamma}_y \frac{\widehat{\text{var}}[G|\mathbf{X}+Y]}{\widehat{\text{var}}[Y|\mathbf{X}+G]} \approx \hat{\gamma}_y \frac{\widehat{\text{var}}[G|\mathbf{X}+Y]}{\widehat{\text{var}}[Y|\mathbf{X}]} \equiv \tilde{\beta}_g.$$

By setting $\tilde{\sigma}^2 = \tilde{\beta}_g^2/T$, we can get an estimate of the variance and report a coefficient-standard error pair that are close approximations to the truth (because the genetic effect can be assumed to be small) and whose ratio gives the same test statistic as the full regression. The approximation is worse when $\widehat{\text{var}}[Y|\mathbf{X}+G]$ is much different than $\widehat{\text{var}}[Y|\mathbf{X}]$, but this difference is typically negligible, as we demonstrate in Section 4. Furthermore, as noted by Zhong and Prentice (2010), interpretation of coefficients near the significance threshold is difficult due to the 'winner's curse' bias. Thus, when differences between this method and OLS exist, they are likely small compared to bias.

### 3.2 Efficient GWAS using Cholesky updates

The method described in the previous section, while very fast for the models described, has some limitations. It only allows us to use model-based standard errors and thus relies on parametric assumptions; in GWAS settings, these are not likely to be checked, nor is there good power to detect important model misspecification (Janssen, 2000). When our model is not correct, either in how, we specify the variance structure or the mean structure, results can be misleading. In main-effects analysis with independent outcomes, the model-based analysis may be appropriate, but there is little work exploring the impacts of misspecification of variance structure (Voorman *et al.*, 2011). Furthermore, it does not give us fitted values that we can use in subsequent IWLS iterations, if desired. In this section, we explore a method based on well-known matrix decompositions which, while slower, overcomes these limitations.

We start with the case of a single marker, which is later generalized to multiple marker models. In order to distinguish between those variables that are the same in each regression and those that differ, denote the model matrix as $\mathbf{X}_g = [1_n, x_1, \ldots x_{p-1}, g] = [\mathbf{X}, g]$. The bottleneck calculation of each fit is a matrix inversion of the form

$$(\mathbf{X_g}^T \mathbf{X_g})^{-1}.$$

This matrix inversion is typically done by first factorizing $\mathbf{X_g}^T \mathbf{X_g}$ to a form which makes inversion very simple. This factorization comprises the bulk of the computation involved in a regression. In this case, all but one column of $\mathbf{X}_g$ remains unchanged between markers. Rather than computing the inverse for each marker, we can simply update the factorization calculated without genotype; this can be done easily using the Cholesky decomposition of $\mathbf{X}^T\mathbf{X}$.

Formally, we denote $\mathbf{L}$ as the lower-triangular Cholesky factorization of $\mathbf{X}^T\mathbf{X}$. We want to augment $\mathbf{L}$ with rows corresponding composed of a $p$-vector $l_g^T$ and a scalar $c$ to form the updated Cholesky decomposition $\mathbf{L}_g$

$$\mathbf{L}_g\mathbf{L}_g^T = \begin{matrix} p \\ 1 \end{matrix} \begin{matrix} \overset{p}{} & \overset{1}{} \end{matrix} \begin{pmatrix} \mathbf{L}\mathbf{L}^T & \mathbf{L}l_g \\ l_g^T\mathbf{L}^T & l_g^T l_g + c^2 \end{pmatrix} = \begin{matrix} p \\ 1 \end{matrix} \begin{matrix} \overset{p}{} & \overset{1}{} \end{matrix} \begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T g \\ g^T\mathbf{X} & g^T g \end{pmatrix}.$$

Since $\mathbf{L}$ is lower triangular, solving the system of equations $\mathbf{L}l_g = \mathbf{X}^T g$ is straightforward. We then set $c^T c = g^T g - l_g^T l_g$. For $p$ variables and $n$ observations, this updating can be done with $O(p^2+np)$ calculations, rather than $O(p^3+np^2)$ calculations, the cost of a typical OLS fit. [The same approach is used by Efron *et al.* (2004) in the LARS algorithm to successively add covariates to a model]. The benefit of this approach is small when there are few covariates or observations, but becomes substantial in larger models and studies, which is precisely when computational speed becomes an issue in GWAS. The application of this procedure to mixed model equations was recently noticed by Meyer and Tier (2011). Our discussion extends the results to a still broader class of models.

The approach for main-effects GWAS outlined above easily extends to the situation where multiple markers or interactions are included. These terms can be subsequently added using the same method, which gives the matrix inverse using the same order of computation as standard matrix inversion (Efron *et al.*, 2004; Golub and Van Loan, 1996).

### 3.3 Missing data

One drawback of pre-computing before genotype is observed is that it is more difficult to adapt the analysis to peculiarities for any single regression. For example, if a subject is missing genotype data, the quantities computed without genotype will still contain this subject's information. If this constitutes only one observation, rank-one downdating algorithms can compute the appropriate factorization efficiently, to which the genotype of the remaining observations can be added as described above (Dongarra, 1979). If there are many missing observations, as would be the case with repeated measures, this downdating may be applied repeatedly; however, it may be more straightforward to conduct a naive analysis. In practice, linkage disequilibrium allows one to accurately impute missing genotype data, so this is not a large concern (Browning and Browning, 2009; Howie *et al.*, 2009; Scheet and Stephens, 2006). The methods used by BOSS can accommodate uncertain genotype, given by a continuous variable, without modification. Currently, BOSS ignores missing genotype, which is equivalent to imputing a value of 0. Observations with missing non-genetic covariates are dropped from the analysis by BOSS.

### 3.4 Meta-analyses

It is common in GWAS analyses to pool effect estimates and associated standard errors from multiple study centers. Since meta-analyses typically do not require that all study centers use the same estimator of an effect, the one-step estimators produced by BOSS can be considered compatible with those produced by other software. Furthermore, in the range of scenarios where it would be beneficial to meta-analyse results, that is, when the effect is modest, one should expect BOSS estimates to nearly identical to traditional estimators.

## 4 RESULTS

### 4.1 Timing comparisons

We implemented the above methods in R and compared the timings using synthetic data (R Development Core Team, 2009). We simulated data for $n = 2000$ subjects and averaged time over $m = 1000$ simulated genetic markers, which is sufficient to accurately estimate the relative speeds. For GEE and mixed models, we generated three observations per patient. For GEE, we modeled the variance structure with an exchangeable working correlation matrix, and for mixed models we used a random intercept and one random slope. To make calculations as fair as possible, we implemented OLS directly with matrix inversion, which is twice as fast as the specialized code in the 'lm()' function; we used 'glm.fit()' for logistic regression, 'geese()' from the 'geepack' package for GEE calculations and 'lmer()' from the 'lme4' package for mixed models (Bates and Maechler, 2009; R Development Core Team, 2009; Yan, 2002; Yan and Fine, 2004). All comparisons were carried out on a Macbook pro with a 2.8 GHz Intel Dual Core i7. The timing comparisons are given in Table 1.

For OLS, the Cholesky updating method is twice as fast as traditional methods, while the swap method is 20 times as fast. The advantage of the Cholesky updating is less pronounced for GxE analyses, but more so for logistic regression. For logistic regression, the swap method is about 125 times faster than the naive method.

**Table 1.** Timing comparisons, given in seconds per 1000 SNPs

| | Naive method | BOSS-chol | BOSS-swap |
|---|---|---|---|
| OLS | 1.29 | 0.74 | 0.059 |
| OLS - GxE | 1.39 | 1.07 | – |
| Logistic reg. | 7.68 | 0.82 | 0.062 |
| GEE - linear, exch. | 325 | 5.38 | 0.31 |
| GEE - logistic, exch. | 383 | 5.27 | 0.31 |
| Linear mixed models | 360 | 2.62 | 0.48 |

For repeated measures, the Cholesky method is 60 times faster than traditional methods for GEE with continuous outcomes 70 times faster for GEE with binary outcomes, and about 137 times faster for linear mixed models. If we do not want to use robust variance estimates, the 'swap' method gives GEE and mixed model estimates 1000 and 750 times faster, respectively, than standard code.

### 4.2 Accuracy of the one-step estimator

In general, we found that the *P*-values from BOSS approximations are extremely accurate, which agrees with what (Kang *et al.*, 2010) demonstrated for mixed models. Figure 1 displays results for logistic regression and GEE with exchangeable working correlation. Genotype was simulated away from the null to demonstrate that there is little difference between the two methods, even for small *P*-values. The simulations are described in more detail in the Supplementary Appendix. For logistic regression, the traditional one-step approximation is not sufficiently accurate for genome-wide significant associations. For GEE, traditional one-step approximation is the same as estimation with an independence working correlation. Although both the independence and exchangeable working correlations are consistent for arbitrary correlation structures, we see that correctly modelling correlation structure increases power.

There is no clear choice for a threshold to identify when further iterations may be beneficial. Since the accuracy of the approach depends on the change in the fitted values, iterating further either when within-cluster or overall fitted values change relative to the model fitted without genotype. However, since the approximations are so close, the simplest and most practical approach is performing subsequent iterations when the approximate *P*-values fall below a certain threshold, say $p^{one-step} < 1/m$ where $m$ is the number of tests. In this way, we would not expect to perform subsequent iterations in a study where genotype has no effect, and when an association is present in the data the *P*-value reported will be as accurate as possible.

### 4.3 Accuracy of approximations in the single marker method

First, as discussed in the previous section, the single marker method calculates a coefficient-standard error pair, which is different from the OLS estimate, but gives identical Wald statistics. We simulated 1000 genotypes for 2000 subjects away from the null hypothesis and computed both the OLS $\hat{\beta}_g$ and the described approximation. Figure 2 demonstrates that this approximation is very close to the truth, even when the genetic effect is large. We see that genotype with

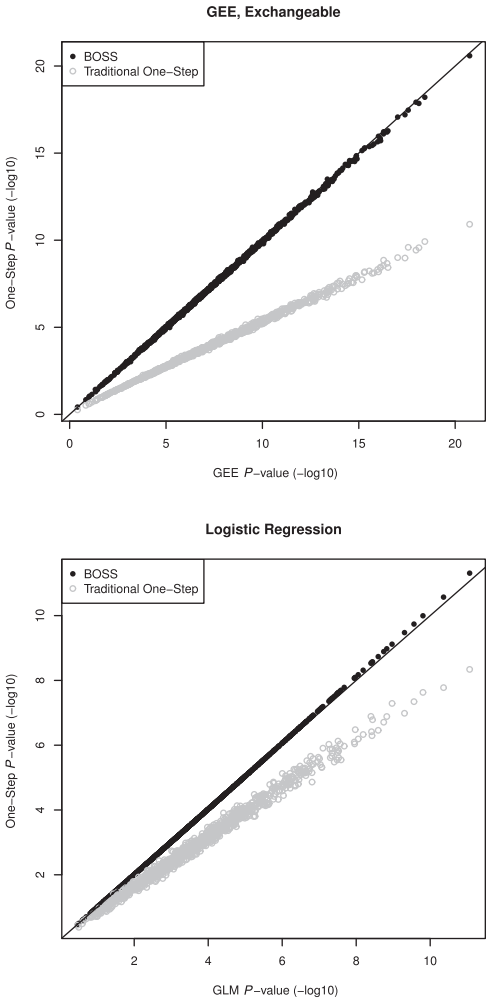


**Fig. 1.** Comparison of *P*-values obtained with BOSS and those obtained with the fully converged model. Results are also displayed for the conventional one-step approximations

$p \approx 10^{-8}$ differs from the standard OLS coefficient by about 1%. Since $\text{var}[Y|\mathbf{X}] \geq \text{var}[Y|\mathbf{X}+G]$, we know that the approximation is an overestimate, but we see that it is by a very small amount.

## 5 CONCLUSION

In this article, we provided general computational methods that capitalize on the structure of genome-wide association studies to increase speed. BOSS is composed of two distinct procedures, but each takes advantage of two important features of genome-wide association studies (i) few variables change from analysis to analysis and (ii) genotype at each locus explains a small component of variation in phenotype. Recognizing this, we demonstrated that one could perform nearly exact estimation and testing in a wide variety of statistical models fit with IWLS including GLM, GEE and mixed models, for a lower computational cost than currently implemented least squares. This had also been noticed in mixed models, but our contribution allows the analyst to use a much

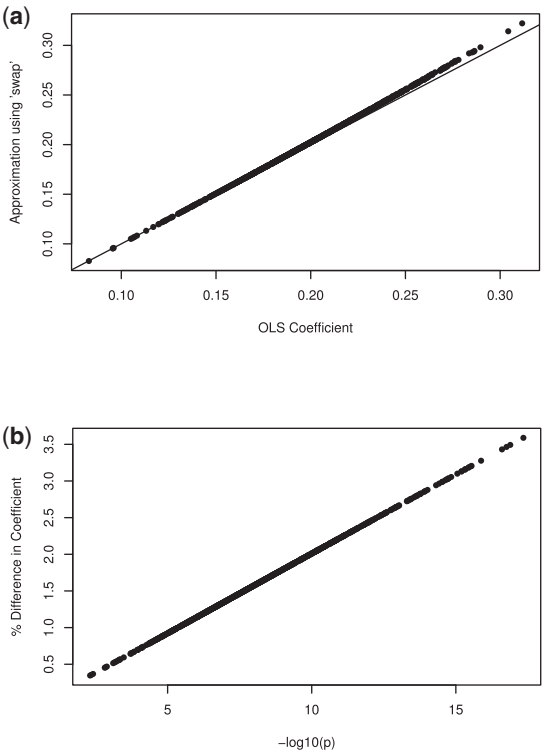


**Fig. 2.** Comparison of approximate coefficients to the OLS estimates using 'swap' method. (**a**) The approximate against the OLS estimate directly. (**b**) The error in the approximation, as a percent of the OLS coefficient, against the *P*-value. Herein, the approximate coefficient, with its associated variance, yields identical inference to the OLS estimate

broader set of tools, and for single marker models we provide a method which is substantially faster (Meyer and Tier, 2011). In particular, it is now feasible to take advantage of repeated measures with GEE and doubly robust standard errors on a genome-wide scale.

*Funding*: National Institutes of Health/National Heart, Lung, and Blood Institute training [T32 HL07183-34, in part] and research [R01 HL074745].

*Conflict of Interest*: none declared.

## REFERENCES

Aulchenko,Y. *et al.* (2007) Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*, **177**, 577.

Bates,D. and Maechler,M. (2009) *lme4: Linear Mixed-effects Models using S4 Classes*. R package version 0.999375-31, http://CRAN.R-project.org/package=lme4

Browning,B. and Browning,S. (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am. J. Hum. Genet.*, **84**, 210–223.

Diggle,P. (2002) *Analysis of Longitudinal Data*, Vol. 25. Oxford University Press, USA.

Dongarra,J. (1979) *LINPACK: Users' Guide*. Number 8. Society for Industrial Mathematics, Philadelphia, Pennsylvania, USA.

Efron,B. *et al.* (2004) Least angle regression. *Ann. Stat.*, **32**, 407–499.

Golub,G. and Van Loan,C. (1996) *Matrix Computations*, Vol. 3. Johns Hopkins Univ Pr., Baltimore, Maryland, USA.

Howie,B. *et al.* (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics*, **5**, e1000529.

Janssen,A. (2000) Global power functions of goodness of fit tests *Ann. Stat.*, **28**, 239–253.

Kang,H. *et al.* (2008) Efficient control of population structure in model organism association mapping. *Genetics*, **178**, 1709.

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

Lehmann,E. and Casella,G. (1998) *Theory of Point Estimation*, Vol 31. Springer Verlag, New York, New York, USA.

Liang,K. and Zeger,S. (1993) Regression analysis for correlated data. *Annual review of public health*, **14**, 43–68.

Lipsitz,S. *et al.* (1994) Performance of generalized estimating equations in practical situations. *Biometrics*, **50**, 270–278.

McCullagh,P. and Nelder,J.A. (1989) *Generalized Linear Models*, 2nd edn., Chapman and Hall/CRC, Boca Raton, Florida, USA.

Meyer,K. and Tier,B. (2011) 'snp snappy': a strategy for fast genome wide association studies fitting a full mixed model. *Genetics*, **190**, 275–277.

Peng,J. *et al.* (2009) Partial correlation estimation by joint sparse regression models. *J. Am. Stat. Assoc.*, **104**, 735–746.

Purcell,S. *et al.* (2007) Plink: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.

R Development Core Team (2009) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Scheet,P. and Stephens,M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am. J. Hum. Genet.*, **78**, 629–644.

Voorman,A. *et al.* (2011) Behavior of qq-plots and genomic control in studies of gene-environment interaction. *PloS one*, **6**, e19416.

Yan,J. (2002) geepack: Yet another package for generalized estimating equations. *R-News*, **2/3**, 12–14.

Yan,J. and Fine,J.P. (2004) Estimating equations for association structures. *Stat. Med.*, **23**, 859–880.

Zhong,H. and Prentice,R. (2010) Correcting winner's curse in odds ratios from genomewide association findings for major complex human diseases. *Genet. Epidemiol.*, **34**, 78–91.