Effect of Trace Estimation and Genotype Architecture on Variance Component Estimates

Michelle Johnson, Noelle Wheeler,
Ali Pazokit, & Sriram Sankararaman

UCLA genetics & genomics

Abstract

Variance components analysis has emerged as a powerful tool in complex trait genetics. Applying this method to large-scale genetic datasets can reveal important insights into genetic architecture, but previous methods of fitting variance components do not scale to these datasets. To address this, we used the scalable Method-of-Moments (MoM) estimator. The key computational bottleneck in the MoM estimator is computing the trace of a large transformed matrix A. Explicitly computing A requires $O(N^3)$ which is not feasible in these cases.

In this project, we assessed the effect of different scalable stochastic trace estimators on variance component estimation. To compare these estimators, we looked at the bias and variance of the MoM estimator as a function of distribution and number of random vectors. We found that a Rademacher distribution yielded the smallest standard error, which decreased with an increasing number of random vectors. This decrease was stronger in Hutch++ than in Hutchinson estimator.

Heritability and Variance Components

$$h^2 = \frac{\hat{\sigma}^2_{genetic}}{\hat{\sigma}^2_{genetic} + \hat{\sigma}^2_{environment}}$$

Eq. 1. Heritability is determined by genetic variance component σ^2_g and residual variance component σ^2_e

$$\mathbf{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}$$

$$\epsilon \sim \mathcal{D}\left(\mathbf{0}, \sigma^2_e \boldsymbol{I}_N\right), \quad \beta \sim \mathcal{D}\left(\mathbf{0}, \frac{\sigma^2_g}{M}\boldsymbol{I}_M\right)$$

Eq. 2. Modeling phenotype off of variance components and genetic data. y (the phenotype data) is determined by X (the genetic data) and vectors β and ϵ, which are pulled from distributions using the variance components.

$$\begin{bmatrix} T & N \\ N & N \end{bmatrix} \begin{bmatrix} \sigma^2_G \\ \sigma^2_e \end{bmatrix} = \begin{bmatrix} c \\ \mathbf{y}^T\mathbf{y} \end{bmatrix}$$

$$T = \frac{tr(\mathbf{X}\mathbf{X}^T\mathbf{X}\mathbf{X}^T)}{M^2} \quad c = \frac{\mathbf{y}^T\mathbf{X}\mathbf{X}^T\mathbf{y}}{M}$$

Eq. 3. The Method of Moments Estimator. The computational bottleneck for solving this equation is finding the trace of XX^TXX^T.

Testing trace estimators

Hutchinson	$\frac{1}{B}\sum_{i=1}^{B} z^T XX^T XX^T z$	Hutch'	$\frac{1}{B}\sum_{i=1}^{B} \left(z_{2i}^T XX^T z_{2i+1}\right)^2$

Hutch++ (1).

Given $\mathbf{A} \in \mathbb{R}^{N\times N}$, where $\mathbf{A} = XX^TXX^T$:

$$z \sim \mathcal{D}\left(\mathbf{0},\ I_N\right)$$

1. Sample $\mathbf{S} \in \mathbb{R}^{N\times B}$ and $\mathbf{G} \in \mathbb{R}^{N\times B}$ with i.i.d.
2. Compute orthonormal basis $\mathbf{Q} \in \mathbb{R}^{N\times B}$ for **AS** (via QR decomposition)
3. **return** $\text{Hutch++}(\mathbf{A}) = tr(\mathbf{Q}^T\mathbf{A}\mathbf{Q}) + \frac{1}{B}tr(\mathbf{G}^T(\mathbf{I} - \mathbf{Q}\mathbf{Q}^T)\mathbf{A}(\mathbf{I} - \mathbf{Q}\mathbf{Q}^T)\mathbf{G})$

Method comparison - *Hutch*' has high standard error

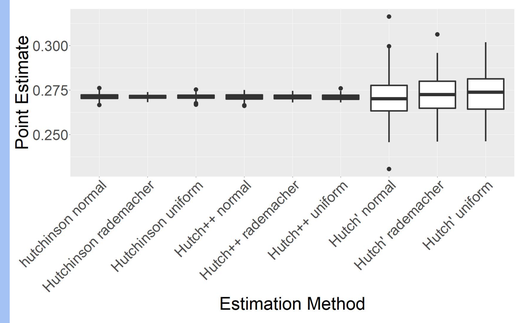
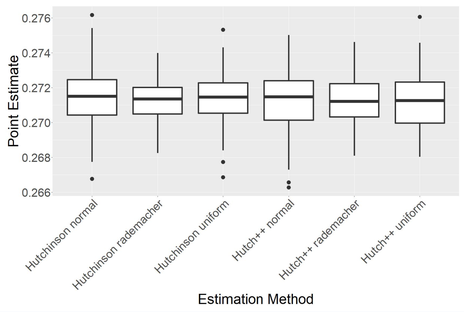


Fig. 1. σ^2_g over three methods. Comparing method and distributions on a simulated genotype matrix containing 10000 individuals and 20000 SNP's. Here estimates were run on the same phenotype, the random number of vectors B =100 and true σ^2_g=0.25

Rademacher yields smallest standard error

Fig. 2. σ^2_g estimate over two methods. Comparing methods and distributions on a simulated genotype matrix containing 10000 individuals and 20000 SNP's. Here estimates were run on the same phenotype, B =100 and true σ^2_g=0.25



Hutch++ scales better with number of random vectors

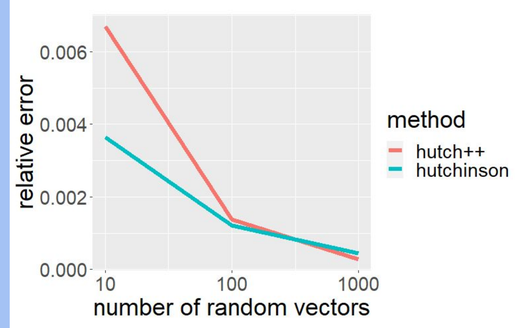


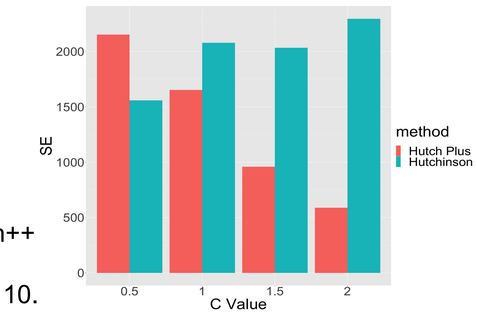
Fig. 3. Relative error of σ^2_g estimates. Comparing relative error, using the rademacher distribution for both methods, on a simulated genotype matrix containing 10000 individuals and 20000 SNP's. σ^2_g =0.25

Hutch++ is better with fast eigenvalue decay

$$A = Q^T\Lambda Q$$
$$\Lambda_{ii} = i^{-c}$$

Eq. 4. A Genotype matrix A is generated using an orthonormal basis Q and diagonal matrix Λ, such that A has desired eigenvalue decay

Fig. 4. SE of trace estimates with set eigenvalues. Comparing Hutchinson and Hutch++ on a 5000 by 5000 genotype matrix simulated using eq. 4. The number of random vectors B = 10.



MoM gets worse and REML gets better with faster decay

Eq. 5. A Genotype matrix A is generated using orthonormal basis Q and diagonal matrix Λ, such that XX^TXX^T has desired eigenvalue decay

$$A = Q^T\Lambda Q$$
$$\Lambda_{ii} = \sqrt[4]{i^{-c}}$$

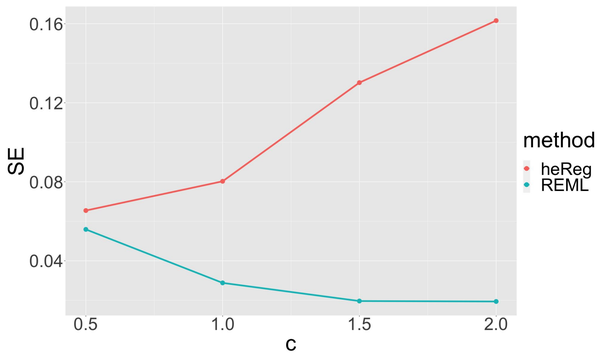
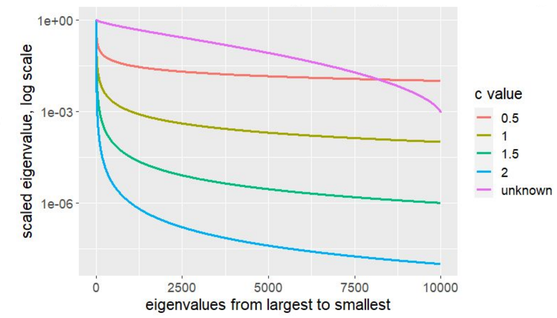


Fig. 5. MoM vs. REML with varying eigenvalues. Variance estimates as given by GCTA software, run on matrices of size 10000 by 10000 generated using eq. 5. heReg represents a MoM estimator and REML uses maximum likelihood

Simulated data has slow eigenvalue decay

Fig. 6. Unknown eigenvalue decay of genotype matrix vs. set eigenvalues. Eigenvalue decay for different given c values, and the eigenvalue decay of a simulated genotype XX^TXX^T, labeled here as "unknown".



Takeaways

- Rademacher distribution yields the best estimates
- Hutch++ outperforms Hutchinson overall, and does much better with faster eigenvalue decay
- MoM has high SE and REML has small SE when XX^TXX^T has high eigenvalue decay
- Hutch++ does not outperform Hutchinson on simulated genotype data, which has a low eigenvalue decay

References and acknowledgements: Thank you to Ali Pazoki & Sriram Sankararaman for overseeing this work, and the BIG summer program for this opportunity.

(1). Meyer, R. A., Musco, C., Musco, C., & Woodruff, D. P. (2021). Hutch++: Optimal stochastic trace estimation. In *Symposium on Simplicity in Algorithms (SOSA)* (pp. 142-155). Society for Industrial and Applied Mathematics.