

# Heritability Estimation by Non-negative Least Squares for fMRI Data

Xu Chen<sup>\*</sup>, Thomas Nichols<sup>\*</sup>, Essi Viding<sup>†</sup> and Alice Jones<sup>‡</sup>

<sup>\*</sup> Department of Statistics, University of Warwick, UK,

<sup>†</sup> Department of Psychology, University College London, London, UK,

<sup>‡</sup> Department of Psychology, University of London, Goldsmiths College, London, UK

## Introduction

Many studies have demonstrated that brain imaging measures are under considerable genetic control [2]. While there are a variety of tools to estimate heritability, currently few are implemented in imaging setting. Heritability is defined to be the proportion of phenotypic variance explained by additive genetic effects in classical ACE model for twin studies, which decomposes the phenotypic variance into three parts [6]: additive genetic variance (A), common environmental variance (C) and unique environmental variance (E). Narrow-sense (additive only) heritability is defined as  $h^2 = A/(A + C + E)$ .

Here we propose two new approaches for voxel-wise heritability measurement, 2-stage restricted maximum likelihood (2-stage ReML) and non-negative least squares (NNLS). NNLS is more computationally efficient and, since it is non-iterative, can never fail to converge. While we are motivated by fMRI, our method is applicable to any type of imaging data.

## Methods

The earliest and simplest way to estimate voxel-wise heritability is to use Falconer's formula [1]:  $h^2 = 2(r_{MZ} - r_{DZ})$ , where  $r_{MZ}$  and  $r_{DZ}$  are the sample correlations of monozygotic (MZ) twins and dizygotic (DZ) twins respectively.

Another approach to voxel-wise heritability mapping is with restricted maximum likelihood (ReML) [4] (equivalent to SEM methods). ReML has much lower bias and variance relative to Falconer's method [7]. However, ReML can often fail to converge, so we propose an alternative "2-stage ReML" method to tackle this problem. It starts by using Fisher scoring algorithm to optimize ReML log-likelihood over  $A$ ,  $C$  and  $E$ . If convergence fails, we switch to another parameterization " $A^*C^*E^*$ ", where  $A^* = A + \frac{E}{4}$ ,  $C^* = C + \frac{E}{4}$ , and  $E^* = \frac{E}{2}$ , and then do the optimization over  $A^*$ ,  $C^*$  and  $E^*$ . This 2-stage approach should avoid convergence problems on voxels where  $A$  is very small.

Statistical Parametric Mapping (SPM) has a ReML function for variance component estimation. This ReML method uses a Bayesian version of ReML where a log Gaussian prior is used to assure the non-negativity. For a given set of specified hyperparameters for the prior, the calculation of variance parameters and heritability is straightforward by SPM's ReML function.

Grimes and Harvey [3] proposed a method for heritability estimation based on least squares. The standard approach to twin modeling uses the covariance matrices for MZ and DZ twin pairs. In the covariance matrices, the phenotypic variance of each individual is  $A + C + E$  regardless of MZ or DZ type; the covariance of MZ twin pair is  $A + C$ , and the covariance of DZ twin pair is  $\frac{A}{2} + C$ . Consider the squared differences of twin pairs; their expectations are

$$E[(MZ_1 - MZ_2)^2] = \text{var}(MZ_1 - MZ_2) = 2E,$$

$$E[(DZ_1 - DZ_2)^2] = \text{var}(DZ_1 - DZ_2) = A + 2E.$$

And, for pairs of unrelated individuals (i.e. unpaired-twins or singletons),

$$E[(I_1 - I_2)^2] = \text{var}(I_1 - I_2) = 2A + 2C + 2E.$$

Thus, the expressions of  $\hat{A}$ ,  $\hat{C}$  and  $\hat{E}$  can be quickly derived by solving a linear regression model as  $D = Z\sigma$ , where  $D$  represents a vector of all possible squared differences of the raw data,  $Z$  is specified by the above equations, and  $\sigma = [A, C, E]^T$ . In order to enforce non-negative requirement for

variance components, NNLS algorithm [5] is proposed for the least squares estimation of  $\sigma$ .

## Simulation Results

Due to the poor performance of Falconer's method, we will only compare three methods (2-stage ReML, SPM with suggested hyperparameters and NNLS) in Monte Carlo simulations. 15 true value sets for  $[A, C, E]^T$  satisfying  $A + C + E = 1$ ,  $E \geq \frac{1}{3}$ , and 3 different sample sizes with both MZ and DZ twins have been chosen. For each setting, 10,000 Monte Carlo simulations are used. MSE by three methods are compared first. Then restricted maximum likelihood ratio test (rLRT) is done to check the validity and to compute the statistical power.

### • Mean Squared Error Comparisons.

Figure 1 shows the MSE comparison of 2-stage ReML, SPM and NNLS. For the two larger sample sizes ("30+30" and "50+50"), these three methods tend to depict identical MSE. While in the "10+10" case, NNLS has smaller MSE in almost all 15 true value settings.

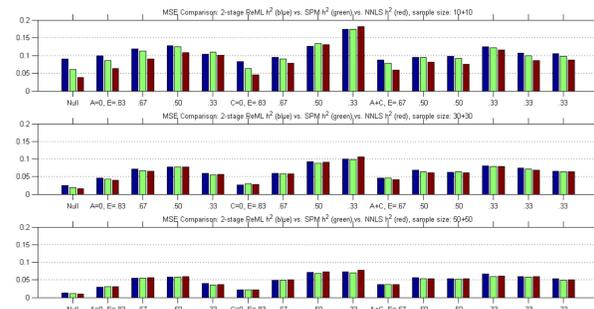


Figure 1: MSE Comparison of Heritability Estimates by Three Methods: 2-stage ReML (blue), SPM (green) and NNLS (red).

### • Validity Checking.

Figure 2 shows that the estimated false positive rates of both SPM and NNLS lie in or below the 95% binomial proportion confidence interval. The conclusion is that both of these methods are valid, and control false positive risk under  $H_0$ : No heritability.

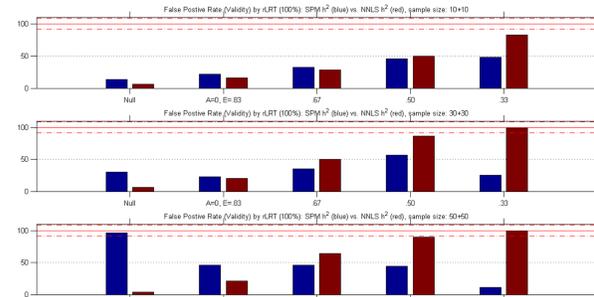


Figure 2: False Positive Rates (100% w.r.t. the nominal size  $\alpha_0 = 0.05$ ) by rLRT of SPM (blue) and NNLS (red).

### • Statistical Power.

Figure 3 shows the comparison of the statistical power of SPM and NNLS. When the sample size is fairly small of "10 + 10", both methods return similar low power. But, it is clear that in nearly all settings, the power of NNLS is

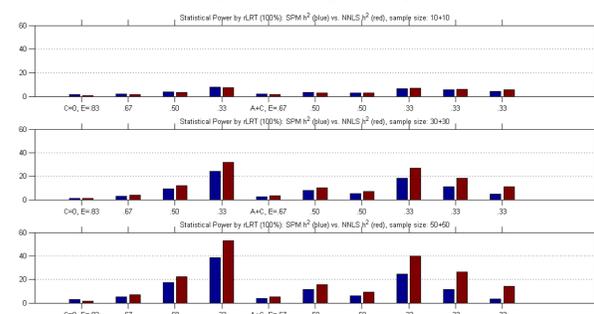


Figure 3: Statistical Power (100%) by rLRT of SPM (blue) and NNLS (red).

bigger than that of SPM for comparatively large sample.

## Real Data Analysis

111 subjects, including 32 MZ twins, 50 DZ twins and 29 singletons, were males (aged 10-12) from Twins Early Development Study (TEDS), some with behavioral problems by SDQ assessment. All participants performed a matching IAPS emotional pictures task. NNLS is the chosen approach for estimation, and rLRT is for activation detection. We use FDR to account for the multiple testing problem in the amygdala only as amygdala is a brain area typically implicated in emotional processing tasks.

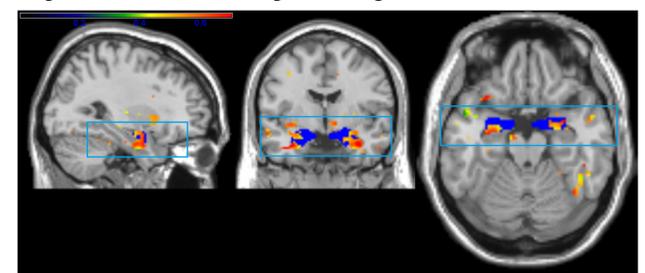


Figure 4: IAPS Emotional Pictures Task Based fMRI Heritability (colored) and Amygdala area (blue). For illustration only, all voxels at  $p < 0.05$  are shown.

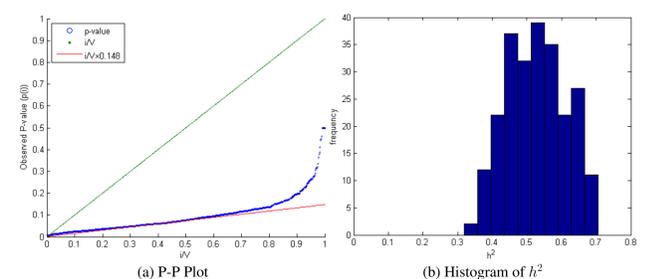


Figure 5: P-P Plot of the 555 Amygdala Voxels (a); Histogram of Heritability Estimates for the 14.8% FDR Significant Amygdala Voxels (b).

Figure 4 shows all  $p < 0.05$  significant voxels, as well as the amygdala mask used for inference. Figure 5(a) shows the distribution of heritability p-values in the amygdala, which are mostly small; the best FDR significance attainable is 14.8%, which leads to 240 voxels being detected of the 555. Applying NNLS to the univariate data by averaging across voxel-wise fMRI data within the amygdala gives heritability estimate as 0.53 and the corresponding p-value as 0.02.

## Conclusion

• For all the three methods considered above, NNLS is very fast, taking around 2 mins to run on one real dataset, and the most time efficient approach. Since it is not an iterative approach, and only deals with the simple linear regression model, which will never encounter convergence problem like ReML and SPM.

• In small sample ("10+10"), NNLS performs better than other two methods first because of the lower MSE. In power detection, it also presents larger statistical power than SPM with the verified false positive rate.

• Although the statistical power of NNLS is still low in an absolute sense, we have successfully identified heritable voxels in the amygdala, and our method still can be improved later if we consider employing smoothing method like Gaussian kernel smoothing.

## References

- [1] Falconer, D.S. and Mackay, T.F.C., *Introduction to quantitative genetics*, 1996.
- [2] Glahn, D.C., et al., *HBM*, 28(6): 488-501, 2007.
- [3] Grimes, L.W. and Harvey, W.R., *J. Animal Sci.*, 50(4): 634-644, 1980.
- [4] Harville, D.A., *JASA*, 72(358): 320-338, 1977.
- [5] Lawson, C.L. and Hanson, R.J., *Solving Least Squares Problems*, 1987.
- [6] Lee, A.D., et al., *IEEE ISBI*, 2010.
- [7] Nichols, T.E., et al., "Improving Heritability Estimates with Restricted Maximum Likelihood (ReML)", *OHBM*, Poster, 2009.