MODE for detecting and estimating genetic causal variants

V. S. Sundar[1,2], Chun-Chieh Fan[1], Dominic Holland[1,3], Anders M. Dale[1,2,3,4]

[1] Center for Multimodal Imaging and Genetics, [2] Department of Radiology,

[3] Department of Neuroscience, [4] Department of Psychiatry,

University of California San Diego, La Jolla, USA

April 18, 2018

Abstract

Determining the genetic causal variants and estimating their effect sizes are considered to be correlated but independent problems. Fine-mapping studies often rely on the ability to integrate useful functional annotation information into genome wide association univariate/multivariate analysis. In the present study, by modeling the probability of a SNP being causal and its effect size as a set of correlated Gaussian/non-Gaussian random variables, we design an optimization routine for simultaneous fine-mapping and effect size estimation. The algorithm is released as an open source C package MODE.

Availability and Implementation: http://sites.google.com/site/sundarvelkur/mode

Contact: amdale@ucsd.edu, svelkur@ucsd.edu

1 Introduction

Detection and estimation of the genetic causal variants associated with a particular phenotypic trait is typically accomplished by reinforcing Genome Wide Association Studies (GWAS) findings with fine mapping analysis. However, for highly polygenic phenotypes, more often than not, biologically causal SNPs do not reach genome-wide significance [1, 2, 3, 4, 5, 6]. In this work, we aim to simultaneously estimate the probability of a SNP being causal and its effect size by developing a well designed optimization routine, that allows for incorporation of functional annotation data. This could potentially aid in accurate detection and estimation of causal loci.

2 Problem statement

Modeling the genotype-phenotype relation through a linear model [7]

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where N is the number of subjects, n is the number of genetic markers, $\boldsymbol{y}$ is a $N \times 1$ phenotype vector, $\mathbf{X}$ related is the $N \times n$ genotype matrix, and $\boldsymbol{\varepsilon}$ is a $N \times 1$ is the vector of noise terms modeled as $N(0, \Sigma_{\boldsymbol{\varepsilon}})$, we aim to estimate the regression coefficients $\boldsymbol{\beta}$ such that $\hat{\varepsilon} = (\hat{\boldsymbol{y}} - \boldsymbol{y})^T(\hat{\boldsymbol{y}} - \boldsymbol{y})$ is minimum. The elements of $\mathbf{X}$ are typically coded as 0,1 or 2. It is well documented that due to the correlated and sparse nature of the SNPs, the univariate regression results in erroneous estimates [8, 9, 10]. Following [11], we minimize $F = L + C$, with

$$\begin{aligned} L &= -\log\Big[(2\pi)^{-n/2}|\tilde{\Sigma}_{\boldsymbol{\varepsilon}}|^{-1/2}\Big] + \frac{1}{2}(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T\tilde{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}(\boldsymbol{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ C &= \sum_{j=1}^{n} c_j(\hat{\beta}_j) \end{aligned} \tag{2}$$

where the cost associated with the j^{th} SNP is given as

$$c_j(\hat{\beta}_j) = -\log\Big[\hat{\pi}_{1j}p_{1j}(\hat{\beta}_j) + (1 - \hat{\pi}_{1j})p_{0j}(\hat{\beta}_j)\Big] \tag{3}$$

Here $\hat{\boldsymbol{\pi}}_1 = [\hat{\pi}_{11}, \hat{\pi}_{12}, \cdots, \hat{\pi}_{1n}]^T$ is the $n \times 1$ vector of non-null prior probabilities of the SNPs. $p_{1j}(\bullet)$ and $p_{0j}(\bullet)$ denote the pdf of causal and null SNPs, respectively. The causal variants and their effect sizes are obtained by minimizing F with respect to $\boldsymbol{\pi}_1$ and $\boldsymbol{\beta}$. Due to Linkage Disequilibrium (LD) and other covariates, the effect sizes and the prior probabilities could be correlation. We take into account these correlations while solving for $\boldsymbol{\pi}_1$ and $\boldsymbol{\beta}$. Minimization of the two-term objective function (likelihood and cost functions) is carried out efficiently using the conjugate gradient method.

3 Method

The function to be minimized is a combination of an error minimizing term - the negative log-likelihood function, and a cost term which imparts the necessary sparse characteristics to the effect sizes. We model the probability of a SNP being causal as uniformly distributed

between a and b, i.e. $\pi_{1j} \sim U[a_j, b_j]$, and the effect sizes as a Gaussian distribution, $\beta_j \sim N(\mu, \sigma)$. Probabilistic shrinkage is achieved by modeling the null SNPs using a Laplace pdf with zero mean and σ_0 standard deviation. Incorporation of Linkage Disequilibrium and function annotation information is through the distributions of prior probabilities, effect sizes, and the correlations among them. The $2n \times 2n$ correlation matrix structure is given as

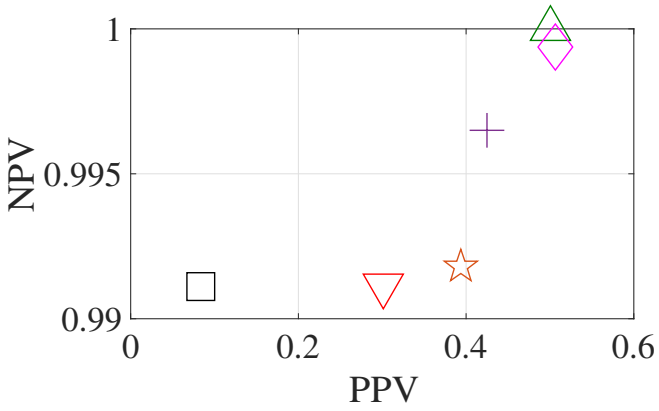
$$\boldsymbol{\rho} = \begin{bmatrix} \boldsymbol{\rho}_{\pi\pi} & \boldsymbol{\rho}_{\pi\beta} \\ \boldsymbol{\rho}_{\beta\pi} & \boldsymbol{\rho}_{\beta\beta} \end{bmatrix}$$

If no information about the correlation structure is known, the $\boldsymbol{\rho}$ matrix is taken to be the identity matrix. The correlated non-Gaussian random variables are first transformed into standard normal random variables using the Nataf's transformation [12], as implemented in [13]. Central difference scheme is used to obtain the gradients of the cost function with respect to $\boldsymbol{\beta}$ and $\boldsymbol{\pi}_1$. Mathematical details regarding the gradients and hessian (with respect to $\boldsymbol{\beta}$) can be found in [11]. A similar approach can be used in obtaining the derivatives with respect to $\boldsymbol{\pi}_1$. MODE source code and binary executables can be downloaded from `http://sites.google.com/sites/sundarvelkur/mode`. MODE borrows functions from ART [13], an open source package for simulation of correlated non-Gaussian random variables.

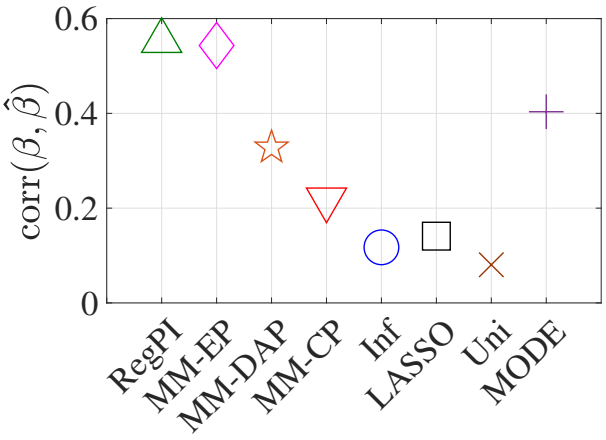
4 Results

4.1 Simulation studies

The phenotype vector with a heritability 0.5 is simulated for 100000 individuals using Eq. (1) utilizing the genotype matrix obtained using Hapgen2 [14] and 1000 Genomes [15]. We consider the first 20000 SNPs of chromosomes 1 to 22 with minor allele frequency greater than 0.01. The number of causal variants are taken to be 50% of all the SNPs belonging to the functional annotation {Exon, 3'UTR, 5'UTR}, with effect sizes distributed as $N(0, 1)$. We consider three replication sets and three different effect size vectors, resulting in 18 cases to estimate the mean positive predictive value (PPV), negative predictive value (NPV), and correlation between the estimated and true effect size. MODE results, along with few other techniques [11] are shown in Figure 1.



(a)



(b)

Figure 1: (a) correlation between the estimated and true effect sizes, (b) PPV and NPV. RegPI: Regularized pseudo inverse (green triangle); MM-EP: Mixture model with enriched priors (magenta diamond); MM-DAP: Mixture model with DAP priors (orange star); MM-CP: Mixture model with constant priors (red inverted triangle); Infinitesimal: Normal prior (no mixture) (blue circle); LASSO (black square); Univariate (brown cross); MODE (purple plus). Refer to [11] for the details of the methods.

5 Discussions

It is observed that in comparison with the MM-CP and MM-DAP [6], MODE estimates have better PPV and NPV characteristics. The improvement in the correlation is due to better NPV of MODE. The relationship between the causal nature of the SNPs and its effect size, and correlations among the causal probabilities and effect sizes are typically unknown. Specification of these quantities require heuristics or additional information, possibly from gene-network analysis, which could identify potential causal SNPs and their relationship with adjacent SNPs. The algorithm is computational tractable unlike MCMC based bayesian methods which requires heavy computational resources and time. MODE locates the causal SNPs and estimates its effect size efficiently with acceptable accuracy.

Funding

This work was supported by the National Institutes of Health (ABCD-USA Consortium, 5U24DA041123).

References

[1] A. J. Schork, W. K. Thompson, P. Pham, A. Torkamani, J. C. Roddey, P. F. Sullivan, J. R. Kelsoe, M. C. O'Donovan, H. Furberg, N. J. Schork, et al., All snps are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated snps, PLoS Genet 9 (4) (2013) e1003449.

[2] R. W. Zablocki, A. J. Schork, R. A. Levine, O. A. Andreassen, A. M. Dale, W. K. Thompson, Covariate-modulated local false discovery rate for genome-wide association studies, Bioinformatics 30 (15) (2014) 2098–2104.

[3] G. Kichaev, W.-Y. Yang, S. Lindstrom, F. Hormozdiari, E. Eskin, A. L. Price, P. Kraft, B. Pasaniuc, Integrating functional data to prioritize causal variants in statistical fine-mapping studies, PLoS Genet 10 (10) (2014) e1004722.

[4] G. Kichaev, B. Pasaniuc, Leveraging functional-annotation data in trans-ethnic fine-mapping studies, The American Journal of Human Genetics 97 (2) (2015) 260–271.

[5] S. L. Spain, J. C. Barrett, Strategies for fine-mapping complex traits, Human molecular genetics 24 (R1) (2015) R111–R119.

[6] X. Wen, Y. Lee, F. Luca, R. Pique-Regi, Efficient integrative multi-snp association analysis via deterministic approximation of posteriors, The American Journal of Human Genetics 98 (6) (2016) 1114–1129.

[7] T. H. E. Meuwissen, B. J. Hayes, M. E. Goddard, Prediction of total genetic value using genome-wide dense marker maps, Genetics 157 (4) (2001) 1819–1829.

[8] S. Kim, K.-A. Sohn, E. P. Xing, A multivariate regression approach to association analysis of a quantitative trait network, Bioinformatics 25 (12) (2009) i204–i212.

[9] X. Zhou, P. Carbonetto, M. Stephens, Polygenic modeling with bayesian sparse linear mixed models, PLoS genetics 9 (2) (2013) e1003264.

[10] G. de los Campos, J. M. Hickey, R. Pong-Wong, H. D. Daetwyler, M. P. Calus, Whole-genome regression and prediction methods applied to plant and animal breeding, Genetics 193 (2) (2013) 327–345.

[11] V. S. Sundar, C.-C. Fan, D. Holland, A. M. Dale, Determining genetic causal variants through multivariate regression using mixture model penalty, Frontiers in Genetics 9 (2018) 77.

[12] J. S. Liu, Monte Carlo strategies in scientific computing, Springer Science & Business Media, 2008.

[13] V. S. Sundar, Art for safety assessment, available from `http://sites.google.com/site/sundarvelkur/art/` (2018).

[14] Z. Su, J. Marchini, P. Donnelly, Hapgen2: simulation of multiple disease snps, Bioinformatics 27 (16) (2011) 2304–2305.

[15] . G. P. Consortium, et al., An integrated map of genetic variation from 1,092 human genomes, Nature 491 (7422) (2012) 56.