On the Use of Z-Scores for Fine-Mapping with Related Individuals

Jicai Jiang[*1]

[1]Department of Animal Science, North Carolina State University, Raleigh, NC, USA

[*]Address correspondence to JJ (jicai_jiang@ncsu.edu)

October 9, 2021

Abstract

Using summary statistics from genome-wide association studies (GWAS) has been widely used for fine-mapping complex traits in humans. The statistical framework was largely developed for unrelated samples. Though it is possible to apply the framework to fine-mapping with related individuals, extensive modifications are needed. Unfortunately, this has often been ignored in summary-statistics-based fine-mapping with related individuals. In this paper, we show in theory and simulation what modifications are necessary to extend the use of summary statistics to related individuals. The analysis also demonstrates that though existing summary-statistics-based fine-mapping methods can be adapted for related individuals, they appear to have no computational advantage over individual-data-based methods.

Introduction

This paper has three sections. First, I go over z-scores-based fine-mapping methods for unrelated individuals. Second, I derive a z-scores-based method for related individuals under the framework of BFMAP [1] that I developed earlier. Third, I show in simulation that it may be risky to use existing z-scores-based methods (e.g., GCTA-COJO [2] and CAVIARBF [3]) for fine-mapping with related individuals.

Results

Fine-mapping with unrelated individuals

Under the polygenic model of inheritance, we have the linear model below for a sample of unrelated individuals:

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e} \text{ with } \mathbf{b} \sim N(0, \mathbf{I}\varphi\sigma_e^2) \text{ and } \mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2), \tag{1}$$

where **y** is an *n*-vector of phenotypes for *n* individuals, **X** is an *n*-by-*m* matrix of standardized matrix for *n* individuals and *m* SNPs, **b** is an *m*-vector of joint SNP effects, **e** is an *n*-vector of residual effects, and φ and σ_e^2 are hyperparameters related to variance components. φ represents the ratio of per-SNP genetic variance to residual variance (σ_e^2) and we can set φ=0.01 to specify a moderate prior variance for SNP effects. In fine-mapping, **X** often represents SNPs in a genomic region rather than whole-wide ones.

Model (1) can be expressed as

$$\mathbf{y} \big| \varphi, \sigma_e^2, \mathbf{X} \sim N(0, \sigma_e^2(\mathbf{X}\mathbf{X}'\varphi + \mathbf{I})), \tag{2}$$

which is further transformed to $\left(\mathbf{X}'\mathbf{y} \big/ \sqrt{n\sigma_e^2}\right) \Big| \varphi, \sigma_e^2, \mathbf{X} \sim N(0, \mathbf{X}'\mathbf{X}\mathbf{X}'\mathbf{X}\varphi/n + \mathbf{X}'\mathbf{X}/n)$. Knowing $\mathbf{z} = \mathbf{X}'\mathbf{y} \big/ \sqrt{n\sigma_e^2}$ and assuming $\mathbf{\Sigma}_x = \mathbf{X}'\mathbf{X}/n$, we have

$$\mathbf{z} \sim N(0, \mathbf{\Sigma}_x\mathbf{\Sigma}_x\varphi n + \mathbf{\Sigma}_x), \tag{3}$$

where **z** is an *m*-vector of z-scores from single-marker association tests and $\mathbf{\Sigma}_x$ is an *m*-by-*m* matrix of Pearson's correlations between *m* SNPs. Given (3), we can compute the Bayes factor of a model including any SNPs over the reduced model including no SNPs:

$$BF = \left| n\varphi \mathbf{\Sigma}_x + \mathbf{I} \right|^{-\frac{1}{2}} \exp\left(\frac{1}{2} \mathbf{z}' \left(\mathbf{I} n^{-1} \varphi^{-1} + \mathbf{\Sigma}_x \right)^{-1} \mathbf{z} \right). \quad (4)$$

Equation (4) provides a way of computing Bayes factors for multi-SNP models using summary statistics from GWAS and SNP correlations from a reference sample. This may be easier than the computation using individual data based on model (2), particularly in human genetic studies where we can use the 1000 Genomes Project reference panel [4] or other resources for calculating SNP correlations; however, the same as model (2), the z-scores-based method is applicable to only samples of unrelated individuals.

Fine-mapping is basically a problem on model selection. There are totally 2^m possible models for *m* SNPs in a candidate genomic region. It is optimal but often infeasible to enumerate all the models; instead, we can use stepwise selection [1, 2] or stochastic shotgun search [1, 5] to reduce the model search burden. Though some fine-mapping methods are not based Bayes factors (e.g., *P*-value-based GCTA-COJO), they are consistent with those using Bayes factors. For example, if GCTA-COJO produces a small conditional *P* value for a SNP, we can observe a much larger log(*BF*) of a model including the SNP than the reduced one.

Fine-mapping with related individuals

In this section, we will develop a z-scores-based fine-mapping method under the same framework as BFMAP for related individuals [1]. BFMAP is a fine-mapping method using individual genotype and phenotype data based on the linear mixed model below

$$\mathbf{y} = \mathbf{Xb} + \mathbf{g} + \mathbf{e} \text{ with } \mathbf{b} \sim N(0, \mathbf{I}\varphi\sigma_e^2),\ \mathbf{g} \sim N(0, \mathbf{G}\eta\sigma_e^2), \text{ and } \mathbf{e} \sim N(0, \mathbf{I}\sigma_e^2). \quad (5)$$

Model (5) has an additional random-effect term, **g**, to model the relatedness between individuals and/or polygenic effects compared to model (1). **G** is a genomic relationship matrix [6], and η is the ratio of genetic variance to residual variance (σ_e^2). Other terms are defined the same as in (1).

Assuming $\mathbf{H} = \mathbf{G}\eta + \mathbf{I}$, we transform (5) to

$$\mathbf{X}'\mathbf{H}^{-1}\mathbf{y}|\varphi, \sigma_e^2, \mathbf{X}, \mathbf{H} \sim N\left(0, \sigma_e^2(\mathbf{X}'\mathbf{H}^{-1}\mathbf{X}\mathbf{X}'\mathbf{H}^{-1}\mathbf{X}\varphi + \mathbf{X}'\mathbf{H}^{-1}\mathbf{X})\right). \tag{6}$$

Defining a diagonal matrix $\mathbf{D} = diag(\mathbf{X}'\mathbf{H}^{-1}\mathbf{X})$ and an m-by-m matrix $\widetilde{\mathbf{\Sigma}}_x = \mathbf{D}^{-1/2}\mathbf{X}'\mathbf{H}^{-1}\mathbf{X}\mathbf{D}^{-1/2}$, we get from (6)

$$\mathbf{z} \sim N\left(0, \widetilde{\mathbf{\Sigma}}_x\mathbf{D}\widetilde{\mathbf{\Sigma}}_x\varphi + \widetilde{\mathbf{\Sigma}}_x\right), \tag{7}$$

where $\mathbf{z} = (\mathbf{D}\sigma_e^2)^{-1/2}\mathbf{X}'\mathbf{H}^{-1}\mathbf{y}$ (that is, an m-vector of z-scores from linear-mixed model associations). We can readily have

$$BF = \left|\mathbf{D}\widetilde{\mathbf{\Sigma}}_x\varphi + \mathbf{I}\right|^{-\frac{1}{2}} exp\left(\frac{1}{2}\mathbf{z}'\left(\mathbf{D}^{-1}\varphi^{-1} + \widetilde{\mathbf{\Sigma}}_x\right)^{-1}\mathbf{z}\right). \tag{8}$$

Equation (8) has a similar form to (4). Genotype correlations between SNPs in (4) for unrelated individuals, $\mathbf{\Sigma}_x = \mathbf{X}'\mathbf{X}/n$, become $\widetilde{\mathbf{\Sigma}}_x = \mathbf{D}^{-1/2}\mathbf{X}'\mathbf{H}^{-1}\mathbf{X}\mathbf{D}^{-1/2}$ in (8) for related individuals. $\widetilde{\mathbf{\Sigma}}_x$ can accordingly be regarded as relatedness-adjusted genotype correlations, which may significantly differ from $\mathbf{\Sigma}_x$. Existing methods, e.g., GCTA-COJO [2] and CAVIARBF [3], use $\mathbf{\Sigma}_x = \mathbf{X}'\mathbf{X}/n$ to calculate correlations even if one provides a sample of related individuals as reference. As a result, the fine-mapping may be misleading.

***Computational burden*.** As shown in our derivation above, the z-scores-based method for related individuals requires the computation of relatedness-adjusted genotype correlations, which is as intensive as individual-data-based BFMAP [1].

Simulations

We show in simulations that GCTA-COJO and CAVIARBF may miss true causal mutations or report false positive ones compared to BFMAP in fine-mapping with related individuals. Mixed-model associations were computed, and original genotypes were used as reference for GCTA-COJO and CAVIARBF.

Pig data. Our first simulation analysis is based on a pig data set recently published [7]. We simulated two quantitative trait nucleotides (QTNs) that are in high LD with each other and a polygenic term amounting to a heritability of 0.5. Scripts and commands for reproducing the analysis are provided in the URLs. In summary, GCTA-COJO always missed one QTN while BFMAP often mapped both.

Dairy cattle genomes. We simulated 500K SNPs and 5K individuals using GENOSIM [8]. Three levels of relatedness were considered: 1) 5K unrelated individuals, 2) 2,500 independent full-sib pairs, and 3) 10 half-sib families each including 500 half-sibs. We randomly selected a SNP on Chr1 as the causal variant and set its proportion of variance explained to 0.05. A polygenic term was simulated so that the heritability reached 0.5. In summary, CAVIARBF might report false positive causal mutations when high relatedness was available while BFMAP rarely did so (Figure 1).

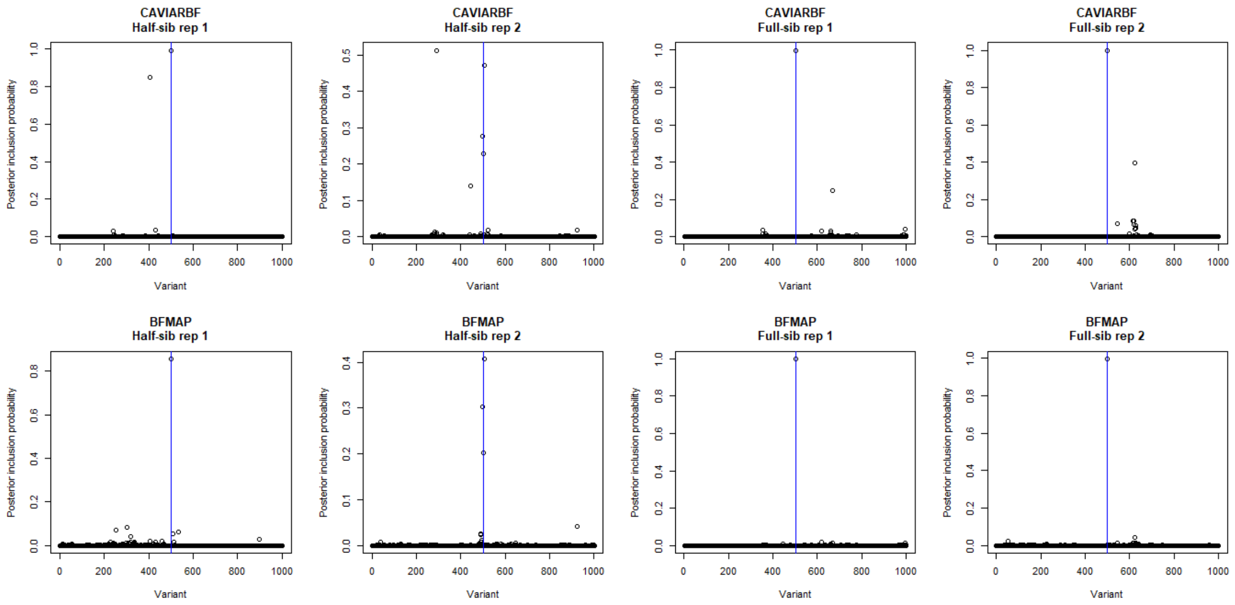

Figure 1. Posterior inclusion probability computed by CAVIARBF and BFMAP for half-sib and full-sib samples. Two simulated replicates were presented for each level of relatedness. Blue lines indicate true causal variants.

Discussion

I have derived a z-scores-based fine-mapping method for related individuals under the framework of BFMAP. The method uses a relatedness-adjusted genotype correlations between SNPs. This key computation is not available in existing z-scores-based methods; as a result, it is risky to apply them to related individuals. We show in simulation that GCTA-COJO and CAVIARBF may miss true causal mutations or report false positive ones compared to BFMAP in fine-mapping with related individuals. Our theoretical analysis also demonstrates that though existing summary-statistics-based fine-mapping methods can be adapted for related individuals, they appear to have no computational advantage over individual-data-based methods. In a nutshell, this study supports the use of BFMAP for fine-mapping with related individuals.

URLs

1. GCTA software: https://cnsgenomics.com/software/gcta/
2. BFMAP software: https://jiang18.github.io/bfmap/
3. CAVIARBF software: https://bitbucket.org/Wenan/caviarbf
4. GENOSIM software: https://aipl.arsusda.gov/software/genosim/
5. Scripts and commands for reproducing the pig simulation analysis: https://github.com/jiang18/bfmap/blob/master/vs_cojo.md

References

1. Jiang, J., et al., *Functional annotation and Bayesian fine-mapping reveals candidate genes for important agronomic traits in Holstein bulls.* Commun Biol, 2019. **2**: p. 212.
2. Yang, J., et al., *Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits.* Nat Genet, 2012. **44**(4): p. 369-75, S1-3.
3. Chen, W., et al., *Fine Mapping Causal Variants with an Approximate Bayesian Method Using Marginal Test Statistics.* Genetics, 2015. **200**(3): p. 719-36.
4. Genomes Project, C., et al., *A global reference for human genetic variation.* Nature, 2015. **526**(7571): p. 68-74.
5. Benner, C., et al., *FINEMAP: efficient variable selection using summary data from genome-wide association studies.* Bioinformatics, 2016. **32**(10): p. 1493-501.
6. VanRaden, P.M., *Efficient methods to compute genomic predictions.* J Dairy Sci, 2008. **91**(11): p. 4414-23.
7. Yang, R., et al., *Accelerated deciphering of the genetic architecture of agricultural economic traits in pigs using a low-coverage whole-genome sequencing strategy.* Gigascience, 2021. **10**(7).
8. VanRaden, P.M., C. Sun, and J.R. O'Connell, *Fast imputation using medium or low-coverage sequence data.* BMC Genet, 2015. **16**: p. 82.