Journal of Statistical Planning and Inference 178 (2016) 70–83

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Assessing statistical significance in variance components linkage analysis: A theoretical justification

Gengxin Li [a], Yuehua Cui [b,c,*]

[a] Department of Mathematics and Statistics, Wright State University, Dayton, OH 45435, United States
[b] Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, United States
[c] School of Economics, Wuhan University of Technology, Wuhan, China

ARTICLE INFO

Article history:
Received 1 January 2016
Received in revised form 12 April 2016
Accepted 7 June 2016
Available online 16 June 2016

Keywords:
Identity-by-decent
Likelihood ratio test
Linkage analysis
Mixture chi-square
Quantitative trait loci

ABSTRACT

Variance components analysis has been a standard means in family-based genetic data analysis. The variance components technique treats genetic effects as random, and tests whether variance components are zero using the likelihood ratio (LR) test. In the literature, the asymptotic distribution of the LR is claimed to follow a mixture chi-square distribution, where the mixture proportions are calculated based on the binomial coefficients, a special case in Self and Liang (1987). This threshold calculation, however, often yields conservative test results as discussed in a number of studies, especially in multi-trait analyses. In this work, we show that the LR statistic asymptotically follows a mixture chi-square distribution where the mixture proportions depend on the estimated Fisher information matrix in both univariate and multivariate trait analyses. We provide a general approximation form for the distribution of the LR under the null hypothesis of no genetic effects. We illustrate our idea with three variance components models in genetic linkage analysis. The performance of the new threshold calculation method is demonstrated via simulation studies, and its application is further illustrated via a real data analysis.

1. Introduction

Genes are the functional units responsible for inheriting biological variations in phenotypes from parents. Inheritance of these characteristics of quantitative traits is attributed to multiple genes working in a coordinated manner. The detectable regions of the genome that contain or are closely linked to causal genes are defined as quantitative trait loci (QTL). The association between QTL and closely linked genes contributing to phenotypic variations is termed as genetic linkage. In human linkage analysis, a variance components (VCs) model is a powerful tool for QTL mapping. In a VC analysis, genetic effects are often partitioned as additive, dominance and polygene effects whereby each one is treated as random (Amos, 1994; Xu and Atchley, 1995). Consider two alleles (A and a) at a gene locus, the additive genetic effect measures "the quantitative change in a trait that is associated with substituting one allele (one genotype) with that of another allele" (from Wikipedia) within a population. Specifically, "the additive effect is half of the difference between the mean of all cases that are homozygous for one version of the allele (a/a) compared to the mean of all cases that are homozygous for the other allele (A/A)" (from Wikipedia). The dominance effect is a non-linear genetic effect which measures the effect where one allele masks the contribution of a second allele at the same locus. For example, if the effect of allele A masks the effect

* Corresponding author at: Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824, United States.
E-mail addresses: gengxin.li@wright.edu (G. Li), cui@stt.msu.edu (Y. Cui).

http://dx.doi.org/10.1016/j.jspi.2016.06.002
0378-3758/

of allele a, then allele A is called a dominant allele and allele a is called a recessive allele (Lynch and Walsh, 1998). In case of over-dominance, the effect of heterozygote genotype Aa is larger than the effect of homozygous genotypes AA and aa. In plants, this is called the hybrid vigor. The polygene accounts for the effect of genes or QTLs not located on the same region as the tested genes (Lynch and Walsh, 1998). Thus, the polygene effect is treated as the background gene effect. In a VC linage analysis, the interest is to test whether the variance component of a genetic effect (i.e., additive and/or dominance effect) is significantly different from zero. Likelihood ratio (LR) test is often applied for such purpose (Amos, 1994). Due to irregular conditions (i.e., parameter boundary problem), the asymptotic distribution of the LR does not follow a regular chi-square distribution, rather a mixture χ^2 distribution, where the mixture proportions are calculated with standard binomial coefficients, a special case discussed in Self and Liang (1987).

A number of studies have demonstrated the asymptotic distribution of LR under irregular conditions, see for example, Chernoff (1954), Self and Liang (1987) and Shapiro (1988). Chernoff (1954) showed that the limiting distribution of the LR has a mixture chi-square distribution when parameters of interest are on one side of a hyperplane, or in the first quadrant within an $\mathbb{R}^2$ space. Self and Liang (1987) extended the Chernoff's result to boundary cases. For a multivariate normal distribution, Kudô (1963) showed the geometric nature of the LR with respect to the slipping means and its mixture chi-square distribution. Kudô and Choi (1975) later on extended the argument to the inference of a one-sided test. Afterwards, Self and Liang (1987) and Shapiro (1988) demonstrated that the boundary problem is equivalent to that of a restricted mean under a multivariate Gaussian distribution. Shapiro (1985) provided one simple proof of the large sample distribution of the LR under the boundary condition for any convex cone. The author further developed a unified theory corresponding to the inequality constrained testing in multivariate normal population for general cases (Shapiro, 1988).

In a univariate linkage analysis with a variance components model, the boundary problem occurs with variance components testing. The distribution of LR in case 9 in Self and Liang (1987) has been commonly applied for a threshold determination (e.g., Amos, 1994 and Hanson et al., 2001). This result is based on the assumption that the unknown parameters are independent, leading to a diagonal variance–covariance matrix for unknown parameters. In reality, the above assumption could be easily violated. This matter consequently leads to conservative hypothesis tests (Allison et al., 1999).

In a bivariate linkage analysis, Amos et al. (2001) proposed an approach to approximate the null distribution of the LR. Again, their derivation assumes a diagonal Fisher information matrix. Additionally, they assumed that the genetic correlation between two traits is perfectly correlated either positively ($\rho = 1$) or negatively ($\rho = -1$) in their derivation, which is unrealistic in reality. Recently Morris et al. (2009) defined a constrained likelihood ratio test (CLRT). They applied Geyer's regularity (1994) to show the asymptotic distribution of the CLRT, but with uncertainty on whether the global M-maximizer can be attained. Because of this limitation, a simulation-based method was developed. However, the computational burden limits its applicability.

In this work, we revisit the LR statistic in testing variance components in linkage analysis under three genetic models (Almasy and Blangero, 2010; Xu and Atchley, 1995; Wang and Zeng, 2009; Nagy et al., 2014), and show that it follows a mixture chi-square distribution. According to the distribution, we provide a new calculation of the mixture proportions based on the estimated Fisher information matrix. The simulation results show the improved performance of the approximation based on the new mixture proportion calculation method. We applied the method to a Genetic Analysis Workshop (GAW) 18 dataset in which two genetic traits, systolic blood pressure (SBP) and diastolic blood pressure (DBP) were measured (Almasy et al., 2014). One can analyze each trait separately and estimate genetic (additive and dominance) effect size. Since the two traits are correlated, we may benefit by a joint analysis of multiple traits using a multi-trait VC model. The rest of this paper is organized as follows. Section 2 introduces three classical VC models with both univariate and multivariate trait analysis. The main result is illustrated in Section 3. Section 4 shows the performance of the new approximation via simulation examples. We demonstrate the utility of the method via a real data analysis in Section 5.

2. Motivating models

The variance components model in a genetic linkage study is typically composed of both fixed and random effects, with non-genetic effects treated as fixed and genetic effects treated as random (Goldgar, 1990; Amos, 1994). The total genetic effect is typically decomposed into additive, dominance and polygenic effects and all are treated as random (Amos, 1994). Below we briefly introduce three VC models and investigate the limiting distribution of the LR statistics under these model setups.

2.1. Model I

Assuming that K families are collected and the phenotype for the kth family is denoted by $\mathbf{y}_k$ with n_k offsprings. For example, $\mathbf{y}_k$ can be a vector of SBP or DBP measures for members in the kth family. Under the variance components model mapping framework, the total genetic effect is partitioned into several components expressed as

$$\mathbf{y}_k = \mu 1_{n_k} + \mathbf{a}_k + \mathbf{d}_k + \mathbf{g}_k + \mathbf{e}_k \tag{1}$$

where μ is the overall mean; $\mathbf{a}_k \sim N(\mathbf{0}, \mathbf{\Pi_k}\sigma^2_{\mathbf{a}})$ is the random additive effect of the major QTL; $\mathbf{d}_k \sim N(\mathbf{0}, \mathbf{\Delta_k}\sigma^2_{\mathbf{d}})$ is the dominance effect of QTL; $\mathbf{g}_k \sim N(\mathbf{0}, \mathbf{\Phi_k}\sigma^2_{\mathbf{g}})$ is the polygenic effect that reflects the effects of unlinked QTLs; $\mathbf{e}_k \sim N(\mathbf{0}, \mathbf{I_k}\sigma^2_{\mathbf{e}})$