Integrating Multiple Correlated Phenotypes for Gene and Environment Interactions Analysis by Maximizing Heritability

November 26, 2015

Abstract

Keywords:

Principal component of heritability, Co-heritability, GWAS, Multivariate analysis

1 Introduction

2 Material and Method

2.1 Integration of Phenotypes

Let m be the unknown number of independent causal loci, indexed by k, n be the number of individuals, indexed by i, and T be the number of phenotypes, indexed by t. In the absence of any covariates or major gene effects, each phenotype is assumed to have the standard polygenic model (**?**), given by

$$\begin{aligned} y_{ti} &= \mu_t + \sum_{k=1}^{m} a_{tk} x_{ki} + e_{ti} + \epsilon_{ti} \\ &= \mu_t + g_{ti} + e_{ti} + \epsilon_{ti}, \end{aligned} \tag{1}$$

where y_{ti} is the tth phenotypic value for the ith individual; μ_t is the mean of the phenotype; x_{ki} is the standardized minor allele count at locus k of individual i, a_{tk} is the additive allelic effect of locus k on phenotype t, $g_{ti} = \sum_{k=1}^{m} a_{tk} x_{ki}$ is the total additive genetic effect of individual i's phenotype t, and the ϵ_{ti} are the residual effects. We treat a_{tk} as random variables independent of the x_{ki}s and of each other, with zero means and common variances and covariances, so that

$$\begin{aligned} \mathrm{E}\left(g_{ti}\right) &= 0 \\ \mathrm{Var}(g_{ti}) &= \sigma_{at}^2 \\ \mathrm{Cov}(g_{ti}, g_{t'i}) &= \sigma_{tt'} \\ &= \sigma_{at}\sigma_{at'}\rho_{tt'}, \end{aligned}$$

where $\sigma_{at}^2 = \mathrm{Var}(\sum_{k=1}^{m} a_{tk} x_{ki})$ is the total additive genetic variance and $\sigma_{tt'}$ is the covariance between the additive effects for phenotypes t and t', average over the k causal loci. This $\sigma_{tt'}$ can be viewed as the average pleiotropy. Finally, assuming the genetic and environmental effects are independent we have

$$\boldsymbol{V}_p = \mathrm{Var}(\boldsymbol{y}_i) = \mathrm{Var}(\boldsymbol{g}_i + \boldsymbol{\epsilon}_i) = \boldsymbol{V}_g + \boldsymbol{V}_e, \tag{2}$$

where $\mathbf{y}_i$, $\mathbf{g}_i$, and $\boldsymbol{\epsilon}_i$ are the length T vectors of phenotypes, genetic and environment components for the ith individual, and

$$\begin{aligned} \boldsymbol{V}_g &= \mathrm{Var}(g_{ti}) = \begin{pmatrix} \sigma_{a_1}^2 & \dots & \sigma_{a_1 T} \\ \dots & \dots & \dots \\ \sigma_{a_{T1}} & \dots & \sigma_{a_T}^2 \end{pmatrix} \\ \boldsymbol{V}_e &= \mathrm{Var}(\epsilon_{ti}) = \begin{pmatrix} \sigma_{e_1}^2 & \dots & \sigma_{e_1 T} \\ \dots & \dots & \dots \\ \sigma_{e_{T1}} & \dots & \sigma_{e_T}^2 \end{pmatrix}. \end{aligned} \tag{3}$$

Note that this model also implies

$$\begin{aligned} \mathrm{Cov}(y_{ti}, y_{ti'}) &= G_{ii'}\sigma^2_{at} \\ \mathrm{Cov}(y_{ti}, y_{t'i'}) &= G_{ii'}\sigma_{att'} \end{aligned}$$

where the $G_{ii'}$s are the genetic relationship coefficients for individuals i and i'. Elements of the $n \times n$ genetic relationship matrix, $\boldsymbol{G}$, can be determined from pedigree information (**?**) or estimated from GWAS data (**?**). This multivariate polygenic model is discussed in **?** and **?**.

Narrow sense heritability of the tth phenotype is defined as the proportion of the additive genetic variance among the total phenotypic variance, i.e.,

$$h^2_t = \frac{\sigma^2_{a_t}}{\sigma^2_{a_t} + \sigma^2_{e_t}}.$$

To integrate multiple phenotypes, our goal is to find a vector of coefficients $\boldsymbol{l}$ such that $\boldsymbol{Yl}$ has the maximum heritability among all such linear combinations of the phenotypes, where $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$ is a $n \times T$ matrix of the collection of all T phenotypes. The heritability of any linear combination of phenotypes $\boldsymbol{Yl}$, can be expressed as the Rayleigh quotient (**?**),

$$h^2_l = \frac{l'\boldsymbol{V}_g l}{l'\boldsymbol{V}_p l}. \tag{4}$$

Henceforth we denote $\boldsymbol{Yl}$ with $\boldsymbol{l}$ chosen to maximize heritability as the set of MaxH phenotypes. The same optimization problem (4) has also been encountered in Fisher's linear discriminant analysis (LDA) for classification (**?**). Detailed explanation for optimizing equation (4) can be found in Supplementary Material and the notes (**?**). Briefly, one needs to eigendecompose the matrix $\boldsymbol{V}_g^{\frac{1}{2}}\boldsymbol{V}_p^{-1}\boldsymbol{V}_g^{\frac{1}{2}}$ and the desired optimization solution is to find the biggest eigenvalue, i.e., maximized heritability h^2_l and the corresponding eigenvector $\boldsymbol{w}$.

The above calculation assumes the parameters in $\boldsymbol{V}_p$ and $\boldsymbol{V}_g$ are known; in reality we need to estimate them. Historically $\boldsymbol{V}_p$ and $\boldsymbol{V}_g$ were estimated using data on pedigrees with known genetic relationships, i.e., $\boldsymbol{G}$. More recent work shows how to approximate $\boldsymbol{G}$ and estimate $\boldsymbol{V}_p$ and $\boldsymbol{V}_g$ from GWAS data on population based samples (**?**). With $\boldsymbol{G}$ treated as known, $(\boldsymbol{V}_g, \boldsymbol{V}_p)$ can be estimated using Maximum Likelihood (ML), Restricted ML (REML) or Method of Moments (MOM) approaches. When the sample size is large, the maximization is not trivial and the computation is costly. We used ML for the application example, and recommend that ML or REML be used in practice. For efficiency of computation, we used the much simpler MOM approach to estimate $\boldsymbol{V}_g$ and $\boldsymbol{V}_p$ in the simulations (**?**). We summarize the steps needed to compute the MaxH phenotype in the Supplementary Material.

2.2 Association Testing and Power Approximation

Thus far, we have focused on maximizing heritability in order to integrate multiple phenotypes. Now we consider testing and power for individual SNPs using MaxH phenotype. To test the hypothesis of no association for a single variant, we include a major gene effect and use the "mixed model" (**?**)

$$y_{ti} = \mu_t + b_t x_{0i} + g_{ti} + \epsilon_{ti} \tag{5}$$

where x_{0i} is the standardized additive coding for the SNP we wish to test and $\boldsymbol{b} = (b_1, \ldots, b_T)$ is the vector of genetic effects for the T phenotypes. Letting $\boldsymbol{Y}_l = (y_{li}) = \boldsymbol{Y}\boldsymbol{l}$ denote the n-vector of MaxH phenotypes, for each element, we have

$$y_{li} = \boldsymbol{l}'\boldsymbol{y_i} = \mu_l + b_l x_{0i} + g_{li} + \epsilon_{li}$$

where $\boldsymbol{y}_i = (y_{1i}, y_{2i}, \ldots, y_{ti})'$ is individual i's T phenotypic measurments, $b_l = \boldsymbol{l}'\boldsymbol{b}$, $\mu_l = \boldsymbol{l}'(\mu_1, \ldots, \mu_T)$, $g_{li} = \boldsymbol{l}'\boldsymbol{g_i}$, and $\epsilon_{li} = \boldsymbol{l}'\boldsymbol{\epsilon}_i$. Hence

$$\begin{aligned} \mathrm{E}\,(y_{li}) &= \mu_l + b_l \\ \mathrm{Var}(y_{li}) &= \boldsymbol{l}'\boldsymbol{V}_p\boldsymbol{l}. \end{aligned}$$

To test $\mathrm{H}_0 : b_l = \boldsymbol{l}'\boldsymbol{b} = 0$, a Wald test is given by

$$W = \frac{\hat{b}_l}{\mathrm{SE}(\hat{b}_l)} \tag{6}$$

where $\hat{b}_l$ is the ordinary least squares (OLS) estimator of b_l and SE is its standard error under the regression model (**?**). In the calculation of $\mathrm{SE}(\hat{b})$ we have neglected the correlation of subjects' phenotypes generated by the polygenic background, since in a population based sample, the genetic relationships are small in practice. But the correlations are considered when generating MaxH phenotypes. Simulation example shows that the type I error rate is protected.

The power of any test to reject $\mathrm{H}_0 : b_l = \boldsymbol{l}'\boldsymbol{b} = 0$ depends not only on the test statistic, but also on how $\boldsymbol{b} = (b_1, \ldots, b_T)$ is chosen. The vector $\boldsymbol{b}$ can be chosen arbitrarily, but if the polygenic model is correct, in a GWAS setting with polygenic effects, it is natural to consider testing SNPs whose genetic effects are consistent with the polygenic model, i.e., $\boldsymbol{b} \sim c\mathcal{N}(0, \boldsymbol{V}_g)$, where c is a scale parameter chosen to determine the heritability of the major gene effect. When including a major gene effect, the overall genetic variance of a linear combination becomes $b_l^2 + \boldsymbol{l}'\,\mathrm{Var}(g_i)\boldsymbol{l}$. In order to maintain a fixed overall heritability (Equation (4)), we choose the major gene effect to satisfy,

$b_l^2 = c^2 \boldsymbol{l}' \boldsymbol{V}_g \boldsymbol{l}$, where $\boldsymbol{l}' \boldsymbol{V}_g \boldsymbol{l}$ is again the total genetic variance including the major gene effect; this implies that b_l explains a fraction c^2 of the total heritability.

The Wald test statistic W^2 in equation (6) follows a chi-square distribution with 1 degree of freedom, i.e., $\chi^2(\delta, 1)$ with non-centrality parameter (NCP)

$$\delta^2 = n \frac{c^2 h_l^2}{1 - c^2 h_l^2}. \tag{7}$$

As heritability h_l^2 increases, the NCP and the power of the test increases, as does the asymptotic power. Power gain is heavily dependent on the gain of heritability. For the MaxH phenotype, the structure of the genotypic and phenotypic variance-covariance matrix and the number of phenotypes combined determines the heritability. In practice $\boldsymbol{V}_p$ and $\boldsymbol{V}_g$ must be estimated, and sampling error may decrease power if too many phenotypes are added. This is considered later, as well as when $\boldsymbol{b}$ comes from arbitrary distributions.

Acknowledgments*