Meta-analysis of multiple populations along with GxE

Alencar Xavier and Shizhong Xu

Genome-wide association

Let the dataset have n observations, f subpopulations, m markers and e environments. The method starts from the genome-wide association analysis in the j^{th} environment, following the alternative model

$$y_j = \mu_j + Z\hat{\gamma}_i + \xi_j + \epsilon_j$$

Where y is the vector corresponding to the response variable, μ is the intercept, Z is a $n \times f$ incidence matrix indicating the haplotype of the maker under evaluation, $\hat{\gamma}_i$ is a vector of the allele effects of the i^{th} marker of f subpopulations, ξ is a vector of length n corresponding to the polygenic term and ϵ is a vector of residuals with length n.

Meta-analysis

For each i marker, the meta-analysis is based upon the concept of sufficient statistics, assuming that environments are independent and all information in each environment can be expressed by the allele effects $\hat{\gamma}_i$ and the observed residual matrix R_i obtained from the association analysis.

The meta-analysis attempt to verify whether the genetic (G) and environmental (E) components of γ differ from zero and, in addition, to verify the existance of $G \times E$ component. In this step, the set of $\hat{\gamma}$ from the association analyses becomes the response variable, a vector with length $e \times f$. For the i^{th} marker, variance components are obtained from the following random model:

$$\gamma_i = \mu_i + Z\alpha_i + W\beta_i + H\delta_i + R$$

Where μ_i is the intercept, Z is a $ef \times f$ incidence matrix indicating the allele source, α_i is the genetic effect associated to the marker, W is a $ef \times e$ incidence matrix indicating the environmental factor, β_i is the coefficient associated to each environment, H is the incidence matrix of genotype by environment interaction, δ_i is the coefficient associated to the $G \times E$ term, and R is the known residual covariance matrix, a block diagonal matrix $ef \times ef$.

AMMI term

The $G \times E$ term might saturate the model once each regression coefficient γ is observed as an unreplicated combination of genotype and environment. The saturation does not occur because the residuals are not independent and the structure is known. Yet, there exist an alternative reparameterization of this term: the additive main effect and multiplicative interaction (AMMI) term.

The AMMI term works as follows. Suppose that the analysis are being performed in a dataset with $f = 5$ subpopulations and $e = 4$ environments. Once γ has been estimated from the association analysis (step 1) and variance components of the genetic and environmental have been estimated with meta-analysis (step 2), then one can build the following E matrix of residuals that also contains the higher-order interaction term:

	E1	E2	E3	E4
G1	ε_{11}	ε_{21}	ε_{31}	ε_{41}
G2	ε_{12}	ε_{22}	ε_{32}	ε_{42}
G3	ε_{13}	ε_{23}	ε_{33}	ε_{43}
G4	ε_{14}	ε_{24}	ε_{34}	ε_{44}
G5	ε_{15}	ε_{25}	ε_{35}	ε_{45}

The AMMI term is extracted from the singlar-value decomposition (SVD). The SVD procedure is commonly used for the extraction of signals from non-square matrices. The decomposition is as follows:

$$E = UDV$$

Where, U is a $e \times e$ matrix, D is a $e \times f$ retangular diagonal matrix, and V is a $f \times f$ matrix. In analogy to the Eigendecomposition, U and V represent Eigenvectors while D are Eigenvalues. Likewise, a small fraction of principal components contain the most amount of information to recontruct the original matrix. Suppose one recontructs E using the first $p = 2$ principal components:

	E1	E2	E3	E4
G1	q_{11}	q_{21}	q_{31}	q_{41}
G2	q_{12}	q_{22}	q_{32}	q_{42}
G3	q_{13}	q_{23}	q_{33}	q_{43}
G4	q_{14}	q_{24}	q_{34}	q_{44}
G5	q_{15}	q_{25}	q_{35}	q_{45}

The matrix above can be rearranged as a vector, and be included into the model of meta-analysis replacing the current $G \times E$ term, thus:

$$\gamma_i = \mu_i + Z\alpha_i + W\beta_i + Q\tau_i + R$$

Hypothesis testing

The log-likelihood of the model is, therefore,

$$L(\mu, \sigma_\alpha^2, \sigma_\beta^2, \sigma_\tau^2) = -0.5(log|V| + (y-\mu)^T V^{-1}(y-\mu))$$

where the variance is expressed as

$$\sigma_y^2 = V = ZZ^T\sigma_\alpha^2 + WW^T\sigma_\beta^2 + QQ^T\sigma_\tau^2 + R$$

and the log-likelihood of the model above is tested agaist $L(\mu = \sigma_\alpha^2 = \sigma_\beta^2 = \sigma_\tau^2 = 0)$, providing the evidence that at least one of the coefficients (intercept and variance components) is not null. Thus

$$LRT = -2(L_{\mu,\sigma_\alpha^2,\sigma_\beta^2,\sigma_\tau^2} - L_{0,0,0,0}).$$

And the second interesting hypothesis to be tested is the importance of the $G \times E$ term alone. Where the likelihood of $L(\mu, \sigma_\alpha^2, \sigma_\beta^2, \sigma_\tau^2)$ is tested against $L(\mu, \sigma_\alpha^2, \sigma_\beta^2, 0)$.

Woodbury's matrix identities

The computational burden associated to the analysis above is originated from the determinant and inversion of the covariance matrix V, a square matrix with ef rows and columns. Let $X = [Z\sigma_\alpha||W\sigma_\beta||Q\sigma_\tau]$, such that $V = XX^T + R$. Using the Woodbury's matrix identities, we have

$$V^{-1} = R^{-1} - R^{-1}X(X^T R^{-1}X + I)^{-1}X^T R^{-1}$$

and

$$|V| = |X^T R^{-1}X + I||R|$$

where the square matrix to be inverted has dimension $e + f + p$. For the example, in the analysis of a dataset with $e = 18$ environments, $f = 41$ subpopulations and using $p = 2$ principal components for the $G \times E$ term, we invert a square matrix with dimension $18 + 41 + 2 = 61$ rows and columns intead of a matrix with $18 \times 41 = 738$ rows and columns.