varianceExplainTheory

December 28, 2017

A linear mixed model in model organism association mapping is typically expressed as $y = x + Zu + e$ where y is an n Œ 1 vector of observed phenotypes, and x is an n Œ q matrix of fixed effects including mean, SNPs, and other confounding variables. is a q Œ 1 vector representing coefficients of the fixed effects. Z is an n Œ t incidence matrix mapping each observed phenotype to one of t inbred strains. And in general Z is None, and will be ignored at some positons in the fellowing document. u is the random effect of the mixed model with $Var(u) = \frac{2}{g}K$, where K is the t Œ t kinship matrix, and e is an n Œ n matrix of residual effect such that $Var(e) = \frac{2}{e}$. The overall phenotypic variance–covariance matrix can be represented as $V = \frac{2}{g}K + \frac{2}{e}I$ This part, in italic font, explains likelihood distribition and the symbols have no relationship with context. For liner model y = x. The distribution of y $N(X, ^2I)$ So the likelihood $L(y; , ^2) = \frac{1}{(2^2)^{\frac{n}{2}}}e^{-\frac{(y-x)(y-x)}{2^2}}$

For the linear mixed model: y $N(x, \frac{2}{g}K + \frac{2}{e}I) = N(x, \frac{2}{g}H)$ $H = \frac{-2}{g}V = K + I$ (there is a typo in the emma paper) $= \frac{\frac{2}{e}}{\frac{2}{g}}$ $likelihood(y; , _g,) = \frac{1}{(2\frac{2}{g}H)^{\frac{n}{2}}}e^{-\frac{(y-x)H^{-1}(y-x)}{\frac{2}{g}}}$ ML (maximum likelihood): $l_F(y; , _g,) = \frac{1}{2}(-nlog(2\frac{2}{g}) - log|H| - \frac{1}{\frac{2}{g}}(y-x)H^{-1}(y-x))$ REML: $l_R(y; _g,) = l_F(y; \hat{}, _g,) + \frac{1}{2}(qlog(2\frac{2}{g}) + log|xx| - log|xH^{-1}x|)$ Gradient of the LMM log likelihood w.r.t. $l_R(y; _g,) = \frac{d - \frac{1}{2\frac{2}{g}}(y-x)^T(K+I)^{-1}(y-x)}{d} = \frac{1}{2}\frac{1}{g}(-x^T(k+I)^{-1}y + x^T(k+I)^{-1}x)$ set gradient to zero: $x^TH^{-1}x = x^TH^{-1}y$ could be estimated as$\hat{} = (xH^{-1}x)^{-1}xH^{-1}y$

Note that this solution is analogous to the ML solution of the linear regression $(x^tx)^{-1}x^Ty$

For ML: $\frac{2}{g} = \frac{R}{n}$ Then: $l_F(y; \hat{}, \hat{}\frac{2}{g}, \hat{}) = \frac{1}{2}(-nlog\frac{2R}{n} - log|H| - n)$

For REML: $\frac{2}{g} = \frac{R}{n-q}$ $R = (y-x)H^{-1}(y-x)$ $H = K + I = U_F(_i + , ..., _n +)U_F$

$K = U_FU_F$ with eigen decomposition so $log|H| = \sum_{i=1}^{n} log(_i +)$ And $R = (y-x)H^{-1}(y-x) = y(I - x(xH^{-1}x)^{-1}xH^{-1})H^{-1}(I - x(xH^{-1}x)^{-1}xH^{-1})y = yPH^{-1}Py$ P is defined as $P = I - x(xH^{-1}x)^{-1}xH^{-1}$ $S = I X(XX)^{1}X $ $SHS = S(K + I)S$ do eigen decomposition And $(SHS)(PH^{-1}P)(SHS) = SHS\ (PH^{-1}P)(SHS)(PH^{-1}P) = (PH^{-1}P)\ PS = P$ (there is a typo in the emma paper) and $SP = S\ SHS = [U_R, W_R]diag(_1 + , ..., _{n-q} + , 0, ..., 0)[U_R, W_R]$
$= U_Rdiag(_1 + , ..., _{n-q}, 0, ..., 0)U_R$

U_R is an n Œ (n q) eigenvector matrix corresponding to the nonzero eigenvalues. W_R is an n Œ q eigenvector matrix corresponding to zero eigenvalues. Here the eigen decomposition of H and SHS do not dependent on any unknow parameter and could be done directly

so $PH^{-1}P = (SHS)^+ = U_Rdiag((_s +)^{-1})U_R$ here (.)^+ denotes the pseudo-inverse of a matrix Let $U_Ry = [_1, _2, ..., _{nq}]$ and $R = yPH^{-1}Py = (U_Ry)diag((_s +)^{-1})(U_Ry) = \sum_{s=1}^{n-q}\frac{\frac{2}{s}}{_s +}$

Then for ML: $l_F(y; \hat{}, \hat{}\frac{2}{g}, \hat{}) = \frac{1}{2}nlog\frac{n}{2} - n - nlog\sum_{s=1}^{n-q}\frac{\frac{2}{s}}{_s +} - \sum_{i=1}^{n}log(_i +)$

$(SHS)(SHS)^+ = (SHS)(PH^{-1}P) = SHPH^{-1}P = SP = S$ On the other hand $(SHS)(SHS)^+ = (U_R diag(_s +)U_R)(U_R diag((_s +)^{-1})U_R) = U_R U_R$ so $U_R U_R = S = I$ taking account $\hat{\frac{2}{g}} = \frac{R}{n-q}$ Then for REML: $l_R(y; \hat{\frac{2}{g}}, \hat{}) = \frac{1}{2}(n-q)log\frac{n-q}{2} - (n-q) - (n-q)log\sum_{s=1}^{n-q}\frac{\frac{2}{s}}{s+} - \sum_{s=1}^{n-q} log(_s +)$ The derivatives of these functions: ML: $f_F = \frac{n}{2}\frac{\sum_s \frac{\frac{2}{s}}{(s+)^2}}{\frac{2}{s+}} - \frac{1}{2}\sum_i \frac{1}{i+}$ REML: $f_F = \frac{n-q}{2}\frac{\sum_s \frac{\frac{2}{s}}{(s+)^2}}{\frac{2}{s+}} - \frac{1}{2}\sum_i \frac{1}{i+}$ if we could find B such that $BB = H = \frac{V}{\frac{2}{g}} = K + I$ we can substitete $y^* = B^{-1}y$, $x^* = B^{-1}x$ and $^* = B^{-1}(Zu +)$ (now * includes both random effects and errors) to get $y^* = x^* + ^*$ $Var(^*) = Var(B^{-1}(Zu +)) = B^{-1}V(B^{-1}) = \frac{2}{g}B^{-1}H(B^{-1}) = \frac{2}{g}B^{-1}BB(B^{-1}) = \frac{2}{g}I$ The value of the residual sum of squares (RSS) from solving the transformed equation $y^* = X^*\beta + \epsilon^*$ is the Mahalanobis RSS for the original equation $y = x\beta + Zu + \epsilon$.

Taking advantage of the eigen decomposition of H performed in the EMMA algorithm, the computation of a valid B^{-1} can be simplified to $B^{-1} = diag(1/\sqrt{\xi_1 + \delta}, ..., 1/\sqrt{\xi_n + \delta})U_F' \; B^{-1}$ is H_sqrt_inv in the code

could be estimate by solving $y^* = x^* + ^*$

The F test here is performed on y^* and x^* In the code h0_rss is the Residual sum of squares without the effecte of fixed variable under testing (H0) mahalanobis_rss is the Residual sum of squares with the effecte of fixed variable under testing (H1)

If here x is the intercept and all the significant genotypic variants, after being calculated Then residuals = y - x Variance explained by significant genotypic variants is (var(y)-residuals.T * residuals) / var(y)