MULTIVARIATE VARIANCE COMPONENTS IN TWIN STUDIES

JAN DE LEEUW AND AGATHA LEE

Abstract. We outline the computation of both FIML and REML estimates in homogeneous sibgroups of a multivariate quantitative characteristic, in the case in which there are no covariates. A program in R is provided.

1. Single Population

Suppose $\underline{x}$ and $\underline{y}$ are m-dimensional normal random vectors [1] with

$$\mathbf{E}(\underline{x}) = \mathbf{E}(\underline{y}) = \mu, \tag{1a}$$

$$\mathbf{E}(\underline{x}-\mu)(\underline{x}-\mu)' = \mathbf{E}(\underline{y}-\mu)(\underline{y}-\mu)' = \Sigma, \tag{1b}$$

$$\mathbf{E}(\underline{x}-\mu)(\underline{y}-\mu)' = \mathbf{E}(\underline{y}-\mu)(\underline{x}-\mu)' = \Omega. \tag{1c}$$

This implies $\Sigma \gtrsim \Omega$ in the Loewner sense, i.e. $\Sigma - \Omega$ must be positive semi-definite. The intraclass correlation matrix [Rao, 1945] is the matrix $\Gamma = \Sigma^{-\frac{1}{2}}\Omega\Sigma^{-\frac{1}{2}}$.

Date: Saturday 10th October, 2009 — 14h 24min — Typeset in Kepler.

Key words and phrases. Multivariate Variance Components, Maximum Likelihood, Twin Studies.

[1] When calculating within a model we underline [Hemelrijk, 1966] the hypothetical random variables whose realizations we observe. The realizations themselves use the same symbol as that for the random variable, but without the underlining. In addition we use overlining for means of vectors of observed quantities, and tildes for the elements of the vector in deviations from the mean.

Define the new random vectors $\underline{u} = \frac{1}{2}(\underline{x} - \underline{y})$ and $\underline{v} = \frac{1}{2}(\underline{x} + \underline{y})$. Then

$$\mathbf{E}(\underline{u}) = 0, \tag{2a}$$

$$\mathbf{E}(\underline{v}) = \mu, \tag{2b}$$

$$\mathbf{E}(\underline{u}\underline{u}') = \frac{1}{2}(\Sigma - \Omega), \tag{2c}$$

$$\mathbf{E}(\underline{u}\underline{v}') = 0, \tag{2d}$$

$$\mathbf{E}(\underline{v} - \mu)(\underline{v} - \mu)' = \frac{1}{2}(\Sigma + \Omega). \tag{2e}$$

Rao et al. [1987]; Srivastava et al. [1988] If we have n independent realizations (u_i, v_i) of $(\underline{u}, \underline{v})$, then the deviance [2] is

$$\begin{aligned}\mathcal{D}(\Delta, \Xi, \mu) = n \log \mathbf{det}(\Delta) + n \log \mathbf{det}(\Xi) + \\ + \sum_{i=1}^n u_i' \Delta^{-1} u_i + \sum_{i=1}^n (v_i - \mu)' \Xi^{-1} (v_i - \mu).\end{aligned} \tag{3}$$

where $\Delta = \frac{1}{2}(\Sigma - \Omega)$ and $\Xi = \frac{1}{2}(\Sigma + \Omega)$.

Thus the maximum likelihood estimate of μ is

$$\hat{\mu} = \overline{v} = \frac{1}{2}(\overline{x} + \overline{y}),$$

which is the mean vector of all $2n$ vectors.

Define $G = \frac{1}{n}\sum_{i=1}^n u_i u_i'$ and $H = \frac{1}{n}\sum_{i=1}^n \tilde{v}_i \tilde{v}_i'$. The concentrated negative log-likelihood is

$$\begin{aligned}\mathcal{D}_\star(\Delta, \Xi) \overset{\Delta}{=} \min_\mu \mathcal{D}(\Delta, \Xi, \mu) = n[\log \mathbf{det}(\Delta) + \log \mathbf{det}(\Xi) + \\ + \mathbf{tr}\ \Delta^{-1} G + \mathbf{tr}\ \Xi^{-1} H].\end{aligned} \tag{4}$$

Maximum likelihood estimates of the variance components can be computed as

$$\hat{\Sigma} = \hat{\Xi} + \hat{\Delta} = H + G,$$

$$\hat{\Omega} = \hat{\Xi} - \hat{\Delta} = H - G.$$

[2] The deviance is twice the negative log-likelihood, except for irrelevant constants.

Note that $\hat{\Sigma} - \hat{\Omega} = 2G \gtrsim 0$, but it is not guaranteed that $\hat{\Omega} \gtrsim 0$. Maximum likelihood estimates that guarantee positive semi-definiteness are discussed in the Appendix.

Since $\mathbf{E}(\underline{G}) = \frac{1}{2}(\Sigma - \Omega)$ and $\mathbf{E}(\underline{H}) = \frac{1}{2}\frac{n-1}{n}(\Sigma + \Omega)$ we can compute unbiased estimates using

$$\hat{\hat{\Sigma}} = \frac{n}{n-1}H + G,$$
$$\hat{\hat{\Omega}} = \frac{n}{n-1}H - G.$$

These unbiased estimates are also the restricted maximum likelihood or REML estimates.

2. Heritability

The ACE model for monozygotic and dizygotic twins decomposes variation into additive genetic, common environmental, and unique environmental variation. In the simplest case we have for the two types of twins

$$\Sigma_M = \Sigma_D = \Theta_A + \Theta_C + \Theta_E,$$
$$\Omega_D = \frac{1}{2}\Theta_A + \Theta_C,$$
$$\Omega_M = \Theta_A + \Theta_C.$$

Solving gives the unbiased estimates

$$\hat{\Theta}_A = 2(\Omega_M - \Omega_D) = 2(\frac{n}{n-1}(H_M - H_D) - (G_M - G_D)),$$
$$\hat{\Theta}_C = 2\Omega_D - \Omega_M = \frac{n}{n-1}(2H_D - H_M) - (2G_D - G_M),$$
$$\hat{\Theta}_E = \Sigma_M - \Omega_M = 2G_M.$$

Appendix A. A Matrix Problem

Consider the problem of minimizing

$$\log\mathbf{det}(\Delta)+\log\mathbf{det}(\Xi)+\mathbf{tr}\ \Delta^{-1}G+\mathbf{tr}\ \Xi^{-1}H$$

over the two matrices Δ and Ξ, with $\Xi\gtrsim\Delta>0$. Clearly if $H\gtrsim G>0$ the solution is $\hat{\Xi}=H$ and $\hat{\Delta}=G$.

Assume $H>0$. Then there is a nonsingular S such that $G=S\Phi S'$ and $H=SS'$ with $\Phi\gtrsim 0$ diagonal. Define $\overline{\Delta}=S'\Delta^{-1}S$ and $\overline{\Xi}=S'\Xi^{-1}S$. Suppose $\tilde{\xi}_i$ and $\tilde{\delta}_i$ are the diagonal elements and λ_i and θ_i are the eigenvalues of $\tilde{\Xi}$ and $\tilde{\Delta}$. Then

$$\sum_{i=1}^n-\log\lambda_i-\log\theta_i+\xi_i\phi_i+\lambda_i$$

References

J. Hemelrijk. Underlining Random Variables. *Statistica Neerlandica*, 20:1–7, 1966.

C.R. Rao. Familial Correlations or the Multivariate Generalization of the Intraclass Correlation. *Current Science*, 14:66–67, 1945.

D.C. Rao, G.P. Vogler, M. McGue, and J.M. Russell. Maximum Likelihood Estimation of Familial Correlations from Multivariate Quantitative Data on Pedigrees: A General Method and Examples. *American Journal of Human Genetics*, 41:1104–1116, 1987.

M.S. Srivastava, K.J. Keen, and R.S. Katapa. Estimation of Interclass and Intraclass Correlations in Multivariate Familial Data. *Biometrics*, 44:141–150, 1988.

(Jan de Leeuw) Department of Statistics, University of California Los Angeles

E-mail address, Jan de Leeuw: deleeuw@stat.ucla.edu

(Agatha Lee) Laboratory of Neuro Imaging, University of California, Los Angeles

E-mail address, Agatha Lee: alee.loni@gmail.com