遗传 HEREDITAS (Beijing) 7 (4): 12—14 1985

多数量性状遗传分析的数据结构

刘垂玕　　刘来福

（安徽农学院，合肥）（北京师范大学）

应用多元统计技术作多数量性状遗传分析已有一些报道，例如 Ramand 等[9]、Bhatt[4] 以及 Hussaini 等[6]在七十年代的有关研究，引起了广泛的注意。在我国，刘来福[1]、杨德等[2]也都提出了自己的研究报告。

Mather 等[7]认为“遗传学的首要原理是：表现型为个体的基因型和个体发育和生活所在的环境的共同结果”。其实质，可以表达为熟知的公式 $P = G + E$ (1)。这里，P 是表现型值，G 是基因型值，E 是环境值。因此，在进行数量性状的遗传分析时，一般都要尽可能地排除环境的效应。按照传统的方法，通常是利用方差-协方差分析技术来达到这一目的的。这样所得到的结果，常被称为遗传方差、遗传协方差、遗传相关系数。由它们构成的矩阵，常被称为遗传协差阵、遗传相关阵。

必须指出，这种传统估计方法产生于对单个数量性状作研究的较早时期。五十年代就有研究报告[10,8]指出，由常规的方差-协方差分析而得到的遗传相关系数误差很大。并且，常常可以在一些研究报告中见到遗传相关系数大于 1 的情况。

传统的遗传协差阵及由其标准化而得的遗传相关阵，会给多元统计方法应用于多数量性状的遗传分析带来具有根本性的困难。

在遗传相关阵的基础上对多数量性状作主成份分析时，有可能得到负的特征根，这不仅难于给出生物学解释，并且使得本来可以接着进行的遗传距离及其对应的聚类分析、因子分析等等都丧失了数学基础。在遗传相关阵的基础上对两组数量性状作典型相关分析时，关联阵有可能得出负的特征根，更是毫无意义的。

下面仅以 28 个水稻品种，4 个区组，每小区调查 10 株，每株观测 15 个性状的随机化完全区组试验资料（摘自芮重庆、赵安常与本文作者的通信）为例，考虑到篇幅，抽出了前 4 个性状，其遗传协差阵为

$$\begin{bmatrix} 0.32654 & -25.6100 & 0.32495 & -4.02559 \\ -25.6100 & 1197.34 & -14.8209 & 129.833 \\ 0.32495 & -14.8209 & 0.38906 & -2.63350 \\ -4.02559 & 129.833 & -2.63350 & 54.4442 \end{bmatrix},$$

其对应的遗传相关阵为

$$\begin{bmatrix} 1.00000 & -1.29698 & 0.91294 & -0.95609 \\ -1.29698 & 1.00000 & -0.68669 & 0.50851 \\ 0.91294 & -0.68669 & 1.00000 & -0.57220 \\ -0.95609 & 0.50851 & -0.57220 & 1.00000 \end{bmatrix}。$$

遗传相关阵的 4 个特征根为 $\lambda_1 = 3.5212$，$\lambda_2 = 0.5215$，$\lambda_3 = 0.3678$，$\lambda_4 = -0.4105$。这里 $\lambda_4 = -0.4105$，它的“遗传贡献”若按绝对值计算是不可忽略的；但 $\sqrt{\lambda_4}$ 在实数范围内无意义，使得对应的主成份计算成为不可能。若将这 4 个性状中的前两个性状对于后两个性状作典范相关分析，则可以得到关联阵为

$$\begin{bmatrix} -0.29809 & 0.09396 \\ -1.0881 & 0.61327 \end{bmatrix},$$

关联阵的特征根为 $\lambda_1 = 0.48227$，$\lambda_2 = -0.33417$。这里 λ_2 为负值，无对应的典范相关系数，更无法作显著性检验。

数学上容易证明：设对于任何 i 及 j 都有 $r_{ii} = 1$ 且 $r_{ij} = r_{ji}$，如果在 n 阶矩阵 $R = (r_{ij})$ 中有任一 $|r_{ij}| \geqslant 1$，$i \neq j$，则矩阵 R 是非正定

Liu Chuiyu et al.: A Structure of Data for Genetical Analysis on Quantitative Characters

的。显而易见，传统方法容易导致遗传相关阵非正定，这时对应的遗传协差阵必然也是非正定的；然而，协差阵的正定性却是多元统计分析的基本条件之一[3]。

从遗传育种的广泛实践来看，采用多元统计技术来进行多数量性状遗传分析，已经成为发展的一个趋势。有鉴于此，寻求一个可以导致遗传分析与统计数学理论和谐一致的解决方案，成为一项紧迫的课题。

在遗传学上，基因型值指的是具有同一基因型的总体的期望[5]，而用样本均值来估计总体期望则是统计学的最佳选择。基于这一认识，本文作者们建议采用以表现型值的均值作为基因型值的估计值的方法来解决上述课题。下面仅以有 p 个品种，c 个性状，b 个区组，每小区调查 m 个个体的随机化完全区组试验为例来加以说明。

1. 如果把随机化完全区组试验的数学模型[1]定为 $X_{ijkl} = g_{ij} + b_{ik} + (g \times b)_{ijk} + e_{ijkl}$ (2)

其中 i、j、k、l 分别为亲本、性状、区组、个体的序号。由于模型(2)具有假设

$$\sum_k b_{ik} = \sum_k (g \times b)_{ijk} = \sum_k \sum_l e_{ijkl} = 0 \qquad (3)$$

从而可得第 i 个品种的第 j 个性状的基因型值的估计值

$$\hat{g}_{ij} = \bar{x}_{ij\cdot\cdot} = \frac{1}{mb} \sum_k \sum_l X_{ijkl} \qquad (4)$$

由于基因型值被定义为表现型值的总体期望，通过公式(4)用样本均值去估计总体期望显然是合理的。

2. 把所得到的基因型值的估计值 $\hat{g}_{ij}$ 按亲本计算出性状的方差($s = t$ 时)及协方差($s \neq t$ 时)，$COV(\hat{g}_s, \hat{g}_t) = \sum_{n=1}^{p} (\hat{g}_{ns} - \bar{g} \cdot s)(\hat{g}_{nt} - \bar{g} \cdot t)/(p-1)$ (5)

这显然是亲本间基因型值真实平均变异量的估计。必须指出，假设(3)只有对于总体才总是真实的；因此，按公式(5)计算出来的方差及协方差中一般仍含有某些环境分量。按传统方法计算出来的遗传方差及遗传协方差，虽然通过减去"误差项均方"而企图排除环境的影响，然而只有在总体的情况下(这时，误差项均方将成为环境方差)以及基因型与环境无交互作用的前提下(实际上这种交互作用总是或多或少地存在的)才能完全达到目的，也就是说，一般情况下遗传方差及遗传协方差也不能完全摆脱环境的影响。

仅仅是出于避免混淆，本文作者们建议把 $COV(\hat{g}_s, \hat{g}_t)$ 在 $s = t$ 时称为基因型值方差，在 $s \neq t$ 时称为基因型值协方差，从而相应地有基因型值协差阵及基因型值相关阵。

应当着重地指出，基因型值相关阵的元素，基因型值相关系数，实际上就是由品种基因型值的估计值直接计算出来的一个普通相关系数，并且这一特性不会因试验设计的不同而变化。如果用 r 表示某个基因型值相关系数，则 $r/\sqrt{(1-r^2)/(n-2)}$ 服从自由度为 $n-2$ 的 t 分布，r 的显著性检验与普通相关系数完全一致，并且基因型值相关系数的误差远较遗传相关系数的误差为小(况且遗传相关系数的估算方法会因试验设计不同而变化，它的误差估计是一个相当复杂的问题)。

仍考察前例，其基因型值协差阵为

$$\begin{bmatrix} 1.18859 & -20.3083 & 0.25879 & -2.11376 \\ -20.3083 & 1404.24 & -15.7553 & 142.885 \\ 0.25879 & -15.7553 & 0.39331 & -2.51639 \\ -2.11376 & 142.885 & -2.51639 & 62.7628 \end{bmatrix},$$

对应的基因型值相关阵为

$$\begin{bmatrix} 1.00000 & -0.49709^{**} & 0.37850^{*} & -0.24473 \\ -0.49709^{**} & 1.00000 & -0.67041^{**} & 0.48130^{*} \\ 0.37850^{*} & -0.67041^{**} & 1.00000 & -0.50648^{**} \\ -0.24473 & 0.48130^{*} & -0.50648^{**} & 1.00000 \end{bmatrix},$$

这里 * 表示 0.05 水平上的显著，** 表示 0.01 水平上的显著。

基因型值相关阵的特征根均为正值：$\lambda_1 = 2.4148$，$\lambda_2 = 0.7723$，$\lambda_3 = 0.5007$，$\lambda_4 = 0.3122$。仍以前两个性状对于后两个性状作典范相关分

析，则有关联阵

$$\begin{bmatrix} 0.02110 & -0.03588 \\ -0.25337 & 0.45864 \end{bmatrix},$$

关联阵的特征根为 $\lambda_1 = 0.47727, \lambda_2 = 0.00247$；从而可得对应的两个典范相关系

$$r_1 = 0.69085^{**},\ r_2 = 0.04970。$$

按照公式(5)而得出的基因型值协差阵是一个真实的样本协差阵[3]，这也是采用多元统计技术的一个基础。对于遗传协差阵，即使是正定的，也不可能保证多元统计技术的执行。例如，我们在实际工作中就曾在遗传协差阵正定的情况下作典范相关分析，而得到"典范相关系数"高达 5.4876，根本无法给出生物学解释。基因型值协差阵具有一切样本协差阵的优良性质，可以适应多元统计技术在多数量性状遗传分析中实施的需要。

如果用 C_p 表示随机化完全区组试验的品种项均协方，则按模型(2)，它的期望为

$$E(C_p) = mbCOV_g + mCOV_{g\times b} + COV_e \qquad (6)$$

然而，按照公式(4)(5)，就有 $COV(\hat{g}_s, \hat{g}_t) = \frac{1}{mb} C_p$ (7)

从而它的期望为

$$E(COV(\hat{g}_s, \hat{g}_t)) = \frac{1}{mb} E(C_p) = COV_g + \frac{1}{b} COV_{g\times b} + \frac{1}{mb} COV_e \qquad (8)$$

这里，COV_g 恰为遗传协方差(在 $s = t$ 时就是遗传方差)的期望。从(8)式显而易见，基因型值方差(协方差)与对应的遗传方差(协方差)是渐近的。从而，基因型值协差阵与对应的遗传协差阵，基因型值相关阵与对应的遗传相关阵，也都是渐近的。

两种渐近的数据结构之所以会在多数量性状遗传分析上产生巨大的差异，究其原因有二：其一是数学上的，就一般矩阵而言，当其元素在某一范围内变动时，其特征根的变动尚无法予以控制；其二是生物学上的，生物学的试验规模受客观条件限制，使得我们无法完全把环境的效应排除干净，例如，即使是作随机化完全区组试验设计，一般情况下，区组项的自由度 b-1 总是比较小的，就所估计的统计量而言，小的自由度对应着大的置信区间，这样就削弱了两种数据结构的渐近性。

参 考 文 献

[1] 刘来福：1979。遗传学报，6(3)：349—355。

[2] 杨 德、戴君惕：1982。同上，9(3)：188—195。

[3] 张尧庭、方开泰：1982。多元统计分析引论，科学出版社，65—193 页。

[4] Bhatt, G. M.: *Aust. Jour. Agric. Res.*, 21(1970): 1—7; 24(1973): 457—464.

[5] Bulmer, M. G.: 1980. *The Mathematical Theory of Quantitative Genetics*, Clarendon Press, Oxford, p. 18.

[6] Hussaini, S. H. et al.: 1977. *Crop Science*, 17: 257—263.

[7] Mather, K. et al.: 1981. 生统遗传学导论（冯午等译），农业出版社，第 5 页。

[8] Mode, C. J. et al.: 1959. *Biometrics*, 15: 518—537.

[9] Ramand, J. et al.: 1970. *Ind. Jour. of Genetics and Plant Breeding*, 30: 1—10.

[10] Reeve, E. C. R.: 1955. *Biometrics*, 11: 357—374.