

最小二乘分析和主成份分析 在黄牛体型特征差异性分析方面的应用*

武 彬

(西北农业大学畜牧系)

摘 要

本研究借助于APPLE—Ⅱ型电子计算机,运用最小二乘分析和主成份分析法处理了秦川牛、晋南牛、南阳牛和延边牛的十三项体尺和体重资料,以主成份值构成了复合性状,用来分析我国黄牛在体型上存在的差异性。研究结果与现实情况基本相符,为目前研究黄牛品种间体型差异性分析提出了一个比较合理的方法。

关键词 最小二乘分析;主成分分析;黄牛体型特征;体尺;体重

前 言

家畜育种工作,多为多目标育种,即希望获得具有多个优良性状的品种。在对不同品种进行比较时,往往要借助于许多指标,如各项体尺、体重以及生产性能等。但是,由于这些目标多为数量性状,有的遗传力很低,而且性状之间又彼此相关,所以对单个性状的研究并不能反映出多个性状的遗传规律及遗传差异。同时,在比较品种间各项指标的差异时,往往受各种因素(如性别、牧场等)的影响,资料不能直接合并,这为我们的研究带来了困难。如能从整体角度出发,借助于多元统计分析理论,消除各因素的影响,构成复合性状或综合性状,在此基础上进行品种内遗传规律的研究和种畜的选择以及研究品种间的真实差异,就可以提高多目标育种的选择效果,提高比较品种差异的准确性。

本研究的目的是用最小二乘分析法(Least Square Analysis)来校正各种因素的影响,然后用主成份分析法(Principal Component Analysis)消除性状之间的相关,形成新的综合指标,在相同的方差下比较我国黄牛品种在体型上的差异性,试图为我国牛品种资源的调查研究探索一个科学的途径和方法。

* 本研究承邱怀教授指导,陕西省乾县秦川牛场,山西省运城地区农牧局,河南省南阳地区黄牛等单位的大力支持和协作,在此一并表示感谢。

本文于1985年6月19日收到。

材 料 来 源

1984年2~9月份分别在陕西省乾县秦川牛场、河南省南阳地区黄牛场、山西省运城地区黄牛场和运城地区种公牛站,测定和收集了五岁以上的成年牛的以下体尺、体重资料:(1)头长、(2)额宽、(3)体高、(4)体斜长、(5)胸围、(6)管围、(7)胸深、(8)胸宽、(9)十字部高、(10)腰角宽、(11)坐骨端宽、(12)前肢高、(13)尻斜长和(14)体重。其中秦川牛83头(公牛13头)、晋南牛65头(公牛10头)、南阳牛70头(公牛11头),同时也引用了延边牛的部分资料。根据以上资料计算以下体尺指数:

- (1) 头长指数 = 头长/体高 $\times 100\%$;
- (2) 额宽指数 = 最大额宽/头长 $\times 100\%$;
- (3) 体型指数 = 体长/体高 $\times 100\%$;
- (4) 胸围指数 = 胸围/体高 $\times 100\%$;
- (5) 骨量指数 = 管围/体高 $\times 100\%$;
- (6) 体躯指数 = 胸围/体长 $\times 100\%$;
- (7) 肢长指数 = 前肢高/体高 $\times 100\%$;
- (8) 胸宽指数 = 胸宽/胸深 $\times 100\%$;
- (9) 胸髁指数 = 胸宽/腰角宽 $\times 100\%$;
- (10) 胸深指数 = 胸深/体高 $\times 100\%$;
- (11) 尻高指数 = 十字部高/体高 $\times 100\%$;
- (12) 尻宽指数 = 坐骨端宽/腰角宽 $\times 100\%$;
- (13) 狭尻指数 = 腰角宽/坐骨端宽 $\times 100\%$;
- (14) 尻长指数 = 尻斜长/腰角宽 $\times 100\%$ 。

原 理 与 方 法

(一) 资料的校正与合并

1. 方差分析最小二乘法的数学模型

$$y_{ijk} = \mu + a_i + b_j + e_{ijk} \quad (1)$$

其中, y_{ijk} 为A因素取第*i*个水平、B因素取第*j*水平的观测值;

μ ——均值;

a_i ——A因素取第*i*个水平的效应;

b_j ——B因素取第*j*个水平的效应;

e_{ijk} ——A取第*i*个水平, B取第*j*个水平观测时的随机误差, 且 $e_{ijk} \sim N(0, \sigma^2)$;

k ——重复数。

本试验中, A为性别因素, 即 a_1 为母畜效应, a_2 为公畜效应; B为品种因素, 即 b_1 为

秦川牛效应, b_2 为晋南牛效应, b_3 为南阳牛效应。

2. 资料合并

用公式 $X_i = \frac{X}{\mu + a_i}$ 校正性别差异后, 对资料进行合并。

(二) 数量性状差异的度量——距离系数

所谓二品种之间的距离系数, 通常指两品种间的欧氏距离。假定第 i 个品种可以用 n 个性状描述, 即用 $(x_{i1}, x_{i2}, x_{i3}, \dots, x_{in})$ 表示, 同样第 j 个品种可用 $(x_{j1}, x_{j2}, x_{j3}, \dots, x_{jn})$ 表示, 则品种 i 与品种 j 之间的欧氏距离平方为:

$$D_{ij}^2 = \sum_{k=1}^n (x_{ik} - x_{jk})^2 \quad (2)$$

由此可见, D_{ij} 表示二品种之间的相异程度。然而, 只有当各性状相互独立, 且具有等方差时, 才可用欧氏距离来定义两品种或群体的距离系数。可是这些性状之间并不互相独立, 各性状的方差也不一定相等, 为此就要对原来的数据矩阵做如下的线性变换:

$$\tilde{\xi}_i = \sum_{k=1}^n L_{ki} \bar{X}_{k..} \quad (i = 1, 2, \dots, n) \quad (3)$$

式中 L_{ki} 为变换系数, 它可以构成一个 $n \times n$ 维系数矩阵, $\tilde{\xi}_i$ 是由 $\bar{X}_{k..}$ 变换的新变量, 是原有性状的综合指标, 它们之间相互独立, 且具有相同的方差。这时可以用新变量计算品种间的欧氏距离, 这一距离反映了品种间的真实差异。

按 (3) 式要求, 由矩阵代数得知, 要得到系数矩阵 $[L_{ki}]$, 只要解特征方阵:

$$(R - \lambda I) L = 0 \quad (4)$$

即可, 式中 R 为 $\bar{X}_{k..}$ 所构成的相关系数矩阵。特征方阵的解 λ 为:

$$\lambda = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix} \quad (5)$$

而且 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$, 称为矩阵 R 的特征根, 而对应的特征向量矩阵为:

$$L = \begin{pmatrix} l_{11} & l_{12} & \dots & l_{1n} \\ l_{21} & l_{22} & \dots & l_{2n} \\ \dots & \dots & \dots & \dots \\ l_{n1} & l_{n2} & \dots & l_{nn} \end{pmatrix} \quad (6)$$

其中第 i 列向量 $(l_{1i}, l_{2i}, \dots, l_{ni})$ 称为对应于特征根 λ_i 的特征向量。由方程 (6) 的特征向量可以求出主成份 ξ_i , 即为所求的新变量。为了使主成份 ξ_i 的方差为 1, 可令每一特征向量 i 除以对应的特征根的平方根 $(\sqrt{\lambda_i})$ 即可, 即

$$\tilde{\xi}_i = \frac{1}{\lambda_i} \sum_{k=1}^n L_{ki} \bar{X}_{\cdot k} \quad (7)$$

于是, 品种间的距离系数定义为:

$$D_{ij}^2 = \sum_{k=1}^n (\tilde{\xi}_{ik} - \tilde{\xi}_{jk})^2 \quad (8)$$

为了简化讨论及提高讨论问题的精确性, 一般仅选取前 n' 个特征根较大的使累积贡献率

$$\lambda = \sum_{j=1}^{n'} \lambda_j / \sum_{j=1}^n \lambda_j \geq 85\%$$

的特征向量所对应的变量作为计算距离系数。

在多元统计分析中, λ_1 所对应的变量

$$\tilde{\xi}_1 = \sum_{k=1}^n L_{k1} \bar{X}_{\cdot k}$$

称为第一主成份, 第一主成份显然对品种的影响最大; 而 λ_2 所对应的变量

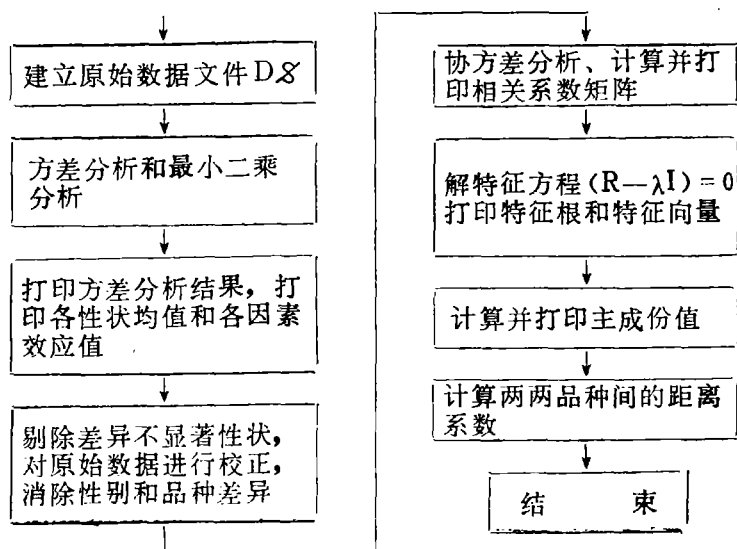
$$\tilde{\xi}_2 = \sum_{k=1}^n L_{k2} \bar{X}_{\cdot k}$$

称为第二主成份, 对品种的影响次之; 其余依次类推。

计算出两两品种间的距离系数之后, 即可对品种进行比较和聚类分析了。

(三) 编制计算机程序和上机运算

采用 BASIC 语言在 APPLE-Ⅱ 型计算机上计算, 程序框图如下:



结果分析与讨论

(一) 方差分析和最小二乘分析

对所测的14个性状以及所计算的14个体尺指数进行品种间和性别间的方差分析, 结果如表1所示。

表1 黄牛各项体尺及其指数、体重的方差分析

方差来源	品种间	性别间	方差来源	品种间	性别间
自由度	2	1	自由度	2	1
头长	5.71**	62.62**	头长指数	244.45**	5.50**
额宽	12.14**	115.61**	额宽指数	0.95	11.39**
体高	108.50**	266.70**	体型指数	25.28**	3.02
体斜长	30.83**	188.38**	胸围指数	59.08**	85.42**
胸围	11.46**	427.99**	骨量指数	37.02**	74.17**
管围	1.83	461.22**	体躯指数	3.32*	49.45**
前肢高	107.39**	51.11**	肢长指数	7.82**	7.76**
胸深	18.90**	286.92**	胸宽指数	7.82**	49.42**
胸宽	6.21**	230.43**	胸腕指数	0.41	113.52**
十字部高	124.51**	201.15**	胸深指数	30.27**	10.72**
腰角宽	12.50**	83.58**	尻高指数	1.19	2.52
尻斜长	14.82**	116.78**	尻宽指数	112.38**	16.03**
坐骨端宽	133.60**	99.09**	狭尻指数	121.21**	19.13**
体重	5.81**	497.04**	尻长指数	10.28**	6.73**

注: 表中的值为F值; **表示差异极显著; *表示差异显著。

从表1可知, 14个性状在性别之间均存在着极显著的差异($P < 0.01$), 唯有管围在品种间未表现出明显差异($P > 0.05$); 同时, 体型指数和尻高指数在性别间未表现出明显差异($P > 0.05$), 其它指数的差异均达到极显著水准。这表明这些差异是由品种和性别提供的, 为了研究品种间的差异, 采用最小二乘法校正了各品种在性别方面的差异, 然后再作进一步分析。

(二) 主成份分析

计算选留性状之间的相关系数如表2所示, 其结果给出了一个十三阶相关矩阵, 计算这一相关矩阵的特征根(λ_i)和特征向量(l_i)及其累积贡献率(Λ)如表3所示。从表3取其累积贡献率大于85%的7个特征根和相应的特征向量, 计算出主成份值(ξ_i)如表4所示。用同样的方法计算体尺指数的6个主成份值(见表4)。

表 2 黄牛各项体尺、体重的相关系数

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
头 长 x_1	1												
额 宽 x_2	0.2424	1											
体 高 x_3	0.4440	0.1969	1										
体 斜 长 x_4	0.3909	0.1643	0.5297	1									
胸 围 x_5	0.3944	0.2511	0.6567	0.5333	1								
前 肢 高 x_6	0.2470	0.0633	0.7494	0.3080	2956	1							
胸 深 x_7	0.3714	0.2456	0.6256	0.4541	0.6612	0.0089	1						
胸 宽 x_8	0.2811	0.0734	0.4044	0.3478	0.6344	0.1913	0.4031	1					
十字部高 x_9	0.4055	0.1098	0.7187	0.3774	0.3719	0.6253	0.3622	0.2866	1				
腰 角 高 x_{10}	0.2867	0.1770	0.5704	0.5015	0.6217	0.2913	0.5329	0.5346	0.3662	1			
尻 斜 长 x_{11}	0.3037	0.0680	0.4683	0.4183	0.4979	0.2436	0.4206	0.3970	0.3754	0.5413	1		
坐骨端宽 x_{12}	0.3397	0.1002	0.4241	0.3946	0.4735	0.2176	0.3725	0.4397	0.2754	0.4538	0.3538	1	
体 重 x_{13}	0.4571	0.2454	0.6380	0.6352	0.8749	0.3214	0.6068	0.6791	0.3968	0.6955	0.5449	0.5037	1

表 3 由黄牛体尺、体重相关矩阵解出的特征根和特征向量

λ_i	$\wedge$	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}	X_{12}	X_{13}
λ_1	6.1295	.4714	.2300	.1118	.3432	.2825	.3444	.2046	.2872	.2690	.3104	.2637	.2466	.3590
λ_2	1.4673	.5844	.0761	-.0972	.3282	-.0204	-.1974	.6375	-.2496	-.2454	.4807	-.1461	-.0653	-.1882
λ_3	1.0607	.6660	.3494	.8168	.0343	.0318	-.0052	-.0892	.1724	-.2774	.0235	-.1465	-.2214	-.0368
λ_4	.7454	.7233	-.7142	.2931	.1457	-.1408	.1668	.1522	.0931	.0476	-.0132	.2431	.0501	-.4834
λ_5	.7124	.7781	-.1017	.3003	-.0714	-.1924	.0858	.2897	-.4556	.4244	-.1512	-.0021	-.4094	.1119
λ_6	.6167	.8255	-.1681	.2504	-.1590	.4597	-.2055	.1018	-.3875	-.3338	-.2031	.2256	.4587	-.0282
λ_7	.5863	.8706	.3118	.1679	-.1624	-.5199	.0664	-.0154	-.2554	.2890	.1056	.0388	.5991	-.0193
λ_8	.5727	.9147	-.2529	.1132	.1463	-.4737	-.1347	-.0923	.3224	-.2330	.1674	.0953	.1431	.6046
λ_9	.3774	.9437	.1339	-.0078	-.1221	.0469	-.3950	-.1456	-.0193	.1302	.2495	.7571	-.3178	-.1491
λ_{10}	.3742	.9725	-.2886	.1668	-.1814	.3412	-.1977	-.2461	.0216	.4125	.5485	-.3816	.0778	.0555
λ_{11}	.2317	.9903	-.0622	-.0563	-.3184	-.1412	.4198	-.2250	-.2930	-.4142	.4848	.0833	-.1036	.0541
λ_{12}	.0986	.9979	.0638	.0060	-.0218	.1319	.6075	-.0385	-.1259	.0675	.0201	.1202	.0134	-.0123
λ_{13}	.0272	1.0000	.0174	-.0087	-.724	.0015	.0446	.5327	.4329	-.0109	.0146	.0118	.0083	.0160

表4 由体尺及其指数、体重得到的主成份值

品 种		秦川牛	晋南牛	南阳牛	延边牛
项目	主成份				
体尺、 体重	ξ_1	3.4678	3.3718	3.6873	0.3204
	ξ_2	0.1882	0.1619	0.2249	0.1941
	ξ_3	0.4996	0.4774	0.5171	0.5498
	ξ_4	-0.1047	-0.0440	-0.1378	0.1161
	ξ_5	0.2630	0.1790	0.2871	0.1474
	ξ_6	0.2417	0.1898	0.2887	0.2859
	ξ_7	0.3654	0.4002	0.3712	0.3622
体尺 指数	ξ_1	1.4769	1.4885	1.4433	1.5653
	ξ_2	0.6850	0.7119	1.6643	0.6301
	ξ_3	2.0825	2.0337	2.0239	1.9830
	ξ_4	0.9260	0.8962	1.0226	0.9705
	ξ_5	-0.2258	-0.3240	-0.1765	-0.3059
	ξ_6	0.1370	0.1219	0.1401	0.0989

(三) 黄牛品种间体型差异的分析

分别以体尺、体重和体尺指数估算品种之间的距离系数平方 (D^2) 如表5所示。

表5 黄牛品种间体型差异的距离系数平方值

品 种	秦川牛	晋南牛	南阳牛	延边牛
秦川牛		0.0140	0.0168	0.0312
晋南牛	0.0507		0.0427	0.0214
南阳牛	0.0538	0.1362		0.0398
延边牛	0.0874	0.0512	0.2554	

注：上三角表示由体尺指数计算的距离系数；
下三角表示由体尺、体重计算的距离系数。

从表5可知，无论就体型大小还是体尺指数而言，秦川牛与晋南牛差异最小；其次是秦川牛与南阳牛；延边牛与南阳牛的差异最大。这与作者依血液型的基因频率所估算的遗传距离结果基本相符。总的看来，秦川、晋南和南阳黄牛品种之间的差异都小于延边牛的差异，这可能与品种的起源和生态类型有关。

总之，本文运用多元统计分析理论的方差分析的最小二乘法 and 主成份分析法处理了许多目标的性状，得到了目前研究畜禽品种间在体型差异性分析方面较合适的方法，在我国畜禽品种分类方面进行了首次尝试，研究结果基本与品种间现实存在的差异相符。但是，本试验限于目前各牛场记载的资料，尚未剖分出基因型值，仅从表型值进行

了分析,其结果可能与真实的遗传距离有所偏差。随着育种制度的健全,在可剖分出基因型值之时,估算出各数量性状的遗传距离,借此参数可进行品种间遗传差异的分析和杂交优势的预测。

参 考 文 献

- [1] 延边黄牛育种课题组: 延边黄牛品种资源调查报告,《延边农学院学报》,1982年第1期,第38—47页。
- [2] 刘来福: 作物数量遗传距离的测定,《遗传学报》,1979年第3期,第349—355页。
- [3] 徐克学、李德中: 科学的分类方法—数量分类法,《生物科学参考资料》,1983年第17卷,第16—25页。
- [4] 毛盛贤: 冬小麦数量性状遗传差异及其在作物育种上的应用,《遗传》,1979年第5期,第26—30页。
- [5] 华元渝等: 主成份分析法在鲢鱼种生长体征指标分析中的应用,《生态学报》,1982年第3期,第267—278页。
- [6] Kendall, M. 著,中国科学院计算机中心概率统计组译:《多元分析》,科学出版社,1975年,第16—37页。

The Application of the Least Square and Principal Component Analysis to the Differentiae on Body Conformation among Some Yellow Cattle Breeds

Wu Bin

(Northwestern Agricultural University)

Abstract

Data of 13 items of body measurements and body weight of Qinchuan, Jinnan cattle, Nanyang and Yanbian yellow cattle breeds are processed by means of the least square and principal component analysis with the aid of Apple II type computer. The complex traits consist of the principal component values, which are used to analyse the differentiae on body conformation of these cattle breeds. The research results are basically agreeable to the actual conditions so that a more rational approach has been advanced to study the differentiae on body conformation among these yellow cattle breeds.

Key words, the least square analysis, principal component analysis, features of body conformation of yellow cattle breeds, body measurements, body weight