高寒草甸生态系统国际学术讨论会论文集 1988 年 10 月
THE PROCEEDINGS OF THE INTERNATIONAL
SYMPOSIUM OF ALPINE MEADOW ECOSYSTEM Oct., 1988

角百灵雏鸟生长的研究
II. 角百灵雏鸟生长度量指标的主成分分析*

周 立
（中国科学院西北高原生物研究所）

雏鸟的生长是一个复杂的过程，通常用体重、体长和各外部器官等众多的生长度量指标来描述。然而，每一个特定的指标只能描述雏鸟生长的某一侧面，例如体重和体长只能分别刻划雏鸟身体重量和长度的生长，并且这些指标的增长往往不是同步等速的。如何利用这众多的生长度量指标综合地、定量地估计雏鸟整体于各日龄生长的快慢和整个个体生长状况，关键在于在这些生长度量指标的基础上寻找一个称之为生长指数的综合度量指标。生长指数应尽可能多地携带各指标提供的信息。它将多维信息浓缩为一维信息，在损失少量信息的条件下，大幅度降低信息维数，使人们容易认识研究对象的数量规律。由于研究对象的各指标之间，常常存在着某种内在联系，并非独立，即存在着相关性，故通常采用主成分分析方法去科学地寻找综合度量指标。

本文探讨青藏高原鸟类优势种角百灵（*Eremophila alpestris elwesi*）雏鸟各生长度量指标之间相互关系；利用主成分分析方法确定生长综合度量指标生长指数，并依据生长指数分析雏鸟生长规律；根据各指标对生长指数的贡献大小，确定各指标对雏鸟整体生长状况的影响(作用)大小及其相对重要性。

一、材料来源

角百灵雏鸟生长的各度量指标：体重、全体长、标准体长、翅、嘴峰、跗蹠、中趾和尾各日龄数据，取自郑生武(1984)在海北高寒草甸生态系统定位站地区的调查报告，列于表1。

二、主成分分析

用随机变量 $Y_1, Y_2, \cdots, Y_8$ 表示上述 8 个生长度量指标。将每个随机变量逐日取

* 本项工作承蒙夏武平教授、王祖望教授热情指导和帮助，特此一并致谢。

表1 角百灵雏鸟的生长

Table 1 The growth of *Eremophila alpestris elwesi* nestlings

日龄(天) age (day)	体重(克) body weight (g)	全体长(毫米) total body length (mm)	标准体长(毫米) standard body length (mm)	翅(毫米) wing (mm)	嘴峰(毫米) beak (mm)	跗蹠(毫米) Tarsometatarsus (mm)	中趾(毫米) middle toe (mm)	尾(毫米) tail (mm)
0	2.5	34.6	34.6	6.2	5.0	5.8	4.1	
1	4.1	39.7	39.7	7.5	5.7	8.0	4.7	
2	6.4	44.1	44.1	8.6	5.8	9.9	5.6	
3	9.6	49.4	49.4	9.9	6.1	12.5	7.0	
4	12.0	53.3	52.3	13.4	6.5	14.3	8.0	1.0
5	15.4	60.0	58.0	18.2	7.1	17.0	8.9	2.1
6	18.1	63.7	60.2	23.4	8.3	18.0	10.7	3.7
7	21.5	69.8	63.4	28.9	9.1	20.3	11.5	6.3
8	23.9	75.1	66.4	35.8	9.5	21.6	11.5	8.8
9	26.5	80.3	68.7	42.4	9.5	21.6	11.5	11.6
10	23.1	85.0	71.3	47.0	9.5	21.6	11.5	13.7
11	21.2	87.8	72.8	50.9	9.5	21.6	11.5	15.7

样数据按列排列，得到一个 12×8 阶矩阵，记为矩阵 $Y=(y_{ij})$，y_{ij} 表示变量 Y_j 在第 $i-1$ 天的取样值 $(i=1, 2, \cdots, 12, j=1, 2, \cdots, 8)$。由于主成分依赖于各度量指标所使用的尺度(量纲)，故将各变量标准化，以消除量纲的影响，即将矩阵Y按列标准化

$$x_{ij}=\frac{y_{ij}-\bar{y}_j}{\sqrt{s_{jj}}} \tag{1}$$

$$i=1, 2, \cdots, 12;\ j=1, 2, \cdots, 8$$

其中

$$\bar{y}_j=\frac{\sum_{k=1}^{12} y_{kj}}{12}$$

$$s_{jj}=\frac{1}{11}\sum_{k=1}^{12}(y_{kj}-\bar{y}_j)^2$$

$$j=1, 2, \cdots, 8$$

表2 各指标的均值和标准离差

Table 2 The mean value and standard deviation for each index.

指标 Index	均值 Mean	标准离差 Standard deviation	变异系数 Variation coefficient
体重 Body weight	15.36	8.27	0.54
全体长 Total body length	61.90	18.02	0.29
标准体长 Standard body length	56.74	12.71	0.22
翅 Wing	24.35	16.31	0.67
嘴峰 Beak	7.63	1.77	0.23
跗蹠 Tarsometatarsus	16.02	5.79	0.36
中趾 Middle toe	8.88	2.91	0.33
尾 Tail	5.24	5.84	1.11

令

$$X_j = (x_{1j}, x_{2j}, \cdots, x_{12j})^T$$
$$j = 1, 2, \cdots, 8$$

则 $X_1, X_2, \cdots, X_8$ 是标准化后的 $Y_1, Y_2, \cdots, Y_8$，记矩阵 $X = (X_1, X_2, \cdots, X_8) = (x_{ij})$．显然有期望 $E(X_j) = 0$，方差 $\mathrm{Var}(X_j) = 1\ (j = 1, 2, \cdots, 8)$．$\bar{y}_j$、$s_{jj}$ 分别是随机变量 $Y_j(j = 1, 2, \cdots, 8)$ 的期望和方差的无偏估计．$Y_1, Y_2, \cdots, Y_8$ 的均值($\bar{y}_j$)、标准离差 ($\sqrt{s_{jj}}$) 及变异系数 $\left(\frac{\sqrt{s_{jj}}}{\bar{y}_j}\right)$ 列于表 2．由表 2 可以看出，各指标的标准离差在 1.77—18.02 之间，以全体长最大，嘴峰最小．尾的变异系数最大，翅次之，标准体长变异系数最小．

令

$$s_{ij} = \frac{1}{11}\sum_{k=1}^{12}(y_{ki} - \bar{y}_i)(y_{kj} - \bar{y}_j) \tag{2}$$

$$r_{ij} = \frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}} \tag{3}$$
$$i, j = 1, 2, \cdots, 8$$

则矩阵 $S = (s_{ij})$、$R = (r_{ij})$ 分别是初始随机变量 $Y_1, Y_2, \cdots, Y_8$ 的协方差矩阵和相关矩阵的无偏估计．容易看出，相关矩阵 R 也是随机变量 $X_1, X_2, \cdots, X_8$ 协方差矩阵的最好估计．随机向量 Y 的相关阵即随机向量 X 的协方差阵 R 列于表 3．由表 3 可见，各变量 Y_i 间均显著相关．除尾 (Y_8) 与中趾 (Y_7)、尾 (Y_8) 与跗蹠 (Y_6) 和体重 (Y_1) 相关系数分别为 0.82、0.84 外，其它各变量间的相关系数均在 0.89 以上．

表 3　相关矩阵 R

Table 3　Correlation Matrix R

指标 Index	1	2	3	4	5	6	7	8
1	1.00	0.95	0.97	0.90	0.98	0.99	0.98	0.84
2	0.95	1.00	0.99	0.98	0.97	0.97	0.95	0.95
3	0.97	0.99	1.00	0.94	0.97	0.99	0.98	0.89
4	0.90	0.98	0.94	1.00	0.94	0.90	0.89	0.99
5	0.98	0.97	0.97	0.94	1.00	0.97	0.97	0.90
6	0.99	0.97	0.99	0.90	0.97	1.00	0.99	0.84
7	0.98	0.95	0.98	0.89	0.97	0.99	1.00	0.82
8	0.84	0.95	0.89	0.99	0.90	0.94	0.82	1.00

把随机变量 X_i 看作 8 维空间的坐标轴，则随机向量 $X = (X_1, X_2, \cdots, X_8)$ 的每一个样本值 $P_k = (x_{k1}, x_{k2}, \cdots, x_{k8})$ $(k = 1, 2, \cdots, 12)$ 就是 8 维空间中的一个点．求随机变量 $X_1, X_2, \cdots, X_8$ 的主成分 $z_1, z_2, \cdots, z_8$，实际上是将坐标轴绕原点旋转，即对随机向量 $X = (X_1, X_2, \cdots, X_8)$ 做一适当的正交变换

$$Z = XL \tag{4}$$

其中

$$Z = (z_1, z_2, \cdots, z_8)$$

$$L=(l_1, l_2, \cdots, l_8)=(l_{ij})$$
$$l_i=(l_{1i}, l_{2i}, \cdots, l_{8i})^T$$
$$i=1, 2, \cdots, 8$$

L是正交矩阵

$$LL^{-1}=LL^T=I \tag{5}$$

符号T表示转置,使得点 $P_1, P_2, \cdots, P_{12}$ 到新的坐标轴 z_1 的距离平方和最小. 根据商高定理,也就是使新的坐标轴 z_1 上的方差最大. 在 z_1 的正交子空间中, 这些点在坐标轴 z_2 上的方差最大,余此类推. 由主成分分析方法 (Green, 1976) 可知, $l_i(i=1,2,\cdots,8)$ 是方程

$$(R-\lambda I)l_i=0 \tag{6}$$

的解. 由矩阵L的性质,l_i是非0向量,因此(6)式等价于

$$|R-\lambda I|=0$$

即l_i是随机向量X的协方差矩阵R的特征向量. 显然R是对称半正定矩阵, 因此可设其特征值为 $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_8 \geqslant 0$, $V_1, V_2 \cdots, V_8$ 为相应的正交规则化特征向量, 若不计符号,必有

$$l_i=V_i \quad (i=1, 2, \cdots, 8)$$

所以, 通过求矩阵R的特征向量确定的 $L=(l_1, l_2, \cdots, l_8)$ 是上述要求的正交变换矩阵. 于是

$$RL=L\Lambda \tag{7}$$

这里

$$\Lambda=\begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_8 \end{bmatrix}$$

由 $l_i(i=1, 2, \cdots, 8)$ 的正交性

$$L^TRL=\Lambda \tag{8}$$

主成分 z_i 的方差

$$\begin{aligned} \mathrm{Var}(z_i) &= E(z_i^2)=E(z_i^Tz_i) \\ &= l_i^TE(X^TX)l_i \\ &= l_i^TRl_i=\lambda_i \end{aligned} \tag{9}$$

主成分 z_i 和 $z_j(i \neq j)$ 不相关

$$\begin{aligned} E(z_iz_j) &= l_i^TE(X^TX)l_j \\ &= l_i^TRl_j=\lambda_jl_i^Tl_j \\ &= \begin{cases} 0; & i \neq j \\ \lambda_j; & i=j \end{cases} \end{aligned} \tag{10}$$

由于

$$\sum_{i=1}^{8}\lambda_i=\sum_{i=1}^{8}\mathrm{var}(X_i)$$

从方差反映信息的角度来看,第 i 个主成分 z_i 的方差 λ_i 占原指标 $X_i(i=1,2,\cdots,8)$

总方差的百分比 $\lambda_i \Big/ \sum_{j=1}^{8} \lambda_j$，表示主成分 z_i 反映原指标信息的多少程度，称为主成分 z_i 的贡献率．而前 m 个主成分的总贡献率 $\sum_{i=1}^{m} \lambda_i \Big/ \sum_{i=1}^{8} \lambda_i$ 称为 $z_1, z_2, \cdots, z_m$ 的累计贡献率．

根据协方差矩阵 R 的特点，我们采用 Jacobi 方法计算特征值和正交规则化特征向量．矩阵 R（表 3）的特征值及相应主成分的贡献率列于表 4．

表 4 特征值和贡献率

Table 4 Eigenstruture values and contribution rates.

i	1	2	3	4	5	6	7	8
特征值 λ_i Eigenstruture value λ_i	7.6035	0.3268	0.0461	0.0171	5.0367×10^{-3}	1.1342×10^{-3}	3.3760×10^{-4}	2.5101×10^{-5}
z_i 贡献率 z_i Contribution rate	0.950	0.041	5.76×10^{-3}	2.14×10^{-3}	0.63×10^{-3}	0.14×10^{-3}	0.42×10^{-4}	0.31×10^{-5}

第 1 主成分 z_1 的贡献率为 95%，即 z_1 反映了原指标 95% 的信息．

对于一个指标 X_i 而言，主成分 z_j 与 X_i 的相关系数

$$\rho(z_j, X_i) = \sqrt{\lambda_j}\, l_{ij} \tag{11}$$

称为主成分 z_j 对指标 X_i 的因子负荷量 (factor loading)．由于 X_i 是 $z_1, z_2, \cdots, z_8$ 的线性组合，因此

$$\sum_{k=1}^{8} \rho^2(z_k, X_i) = \sum_{k=1}^{8} \lambda_k l_{ik}^2 = \mathrm{var}(X_i) = 1$$

称

$$\nu_i = \frac{\rho^2(z_k, X_i)}{\mathrm{var}(X_i)} = \rho^2(z_k, X_i)$$

$$= \lambda_k l_{ik}^2 \tag{12}$$

为主成分 z_k 对指标 X_i 的贡献率，它表示 z_k 携带 X_i 信息的百分比．主成分 z_1 对 X_i 的因子荷量和贡献率列于表 5．表 5 显示主成分 z_1 除对指标 X_8 的贡献率为 86% 外，对其他指标的贡献率均在 95% 以上，即 z_1 携带了这些指标的 95% 以上的信息．

表 5 主成分 z_1 对指标 x_i 的因子负荷量和贡献率

Table 5 The factor loading and contribution rate of the principal component z_1 for indexes x_i.

i	1	2	3	4	5	6	7	8
因子负荷量 Factor loading $\rho(z_1, x_i)$	0.976	0.995	0.991	0.995	0.989	0.981	0.973	0.926
贡献率 ν_i Contribution rate ν_i	0.95	0.99	0.98	0.99	0.98	0.96	0.95	0.86

由于第 1 主成分 z_1 反映了全部指标 $X_1, X_2, \cdots, X_8$ 所包含信息的 95%，而其余的 7 个主成分只载有 5% 的信息，所以只取一个主成分 z_1 作为刻划雏鸟生长的 8 个指标复

杂数集的一个概括，几乎反映了原指标的全部信息．这样就将载有这些信息的有效维数从 8 降到 1，使得对雏鸟整个个体生长的分析变得直观、容易．经计算，λ_1 的正则化特征向量 $l_1 = (0.35, 0.36, 0.36, 0.35, 0.36, 0.36, 0.35, 0.34)^T$，于是

$$z_1 = Xl_1 = (X_1, X_2, \cdots, X_8)l_1$$

$$= 0.35(X_1 + X_4 + X_7) + 0.36(X_2 + X_3 + X_5 + X_6) + 0.34X_8 \quad (13)$$

我们定义 z_1 为雏鸟的生长指数，作为雏鸟整体生长状况的综合度量指标． z_1 是原指标 $X_i(i = 1, 2, \cdots, 8)$ 的线性组合，在 z_1 的表达式中各指标的系数（或各指标的权）的大小，表示该指标对雏鸟生长指数的贡献大小，亦即对雏鸟整体生长状况的影响（作用）大小．各指标的系数在 0.34—0.36 之间，可见标准化后的原指标对雏鸟整体生长的作用几乎是相等的，其相对重要性几乎是相同的．根据原指标 $Y_1, Y_2, \cdots, Y_8$ 于各日龄的取样值(表 1)，由(13)式计算的角百灵雏鸟于各日龄生长指数值列于表 6．表 6 显示，在雏期的前 9 天里每天都有较大的生长指数日增量，其中以第6、7天雏鸟生长的最快．第 10，11 天

表 6　各日龄雏鸟的生长指数

Table 6　The growth index for each day age of nestlings.

日龄 Day age	0	1	2	3	4	5	6	7	8	9	10	11
生长指数 Growth index	−4.16	−3.47	−2.88	−2.07	−1.35	−0.43	0.62	1.64	2.39	2.98	3.23	3.44
日增量 Daily increase	—	0.69	0.59	0.81	0.72	0.92	1.05	1.02	0.75	0.59	0.25	0.21

生长速度显著减慢，生长指数日增量比前 9 天的最小值的一半还小，标志第 10 天以后雏鸟的生长发育进入缓慢阶段．

由多元回归理论可知，标准化后的原指标 $X_1, X_2, \cdots, X_8$ 对主成分 $z_1, z_2, \cdots, z_8$ 可以建立回归方程，回归方程恰好为

$$X = ZL^T \quad (14)$$

若只取一个主成分 z_1，则回归方程为

$$X = Z_1 l_1^T$$

即

$$X_1 = l_{11} z_1$$

$$X_2 = l_{21} z_1 \quad (15)$$

$$\cdots$$

$$X_8 = l_{81} z_1$$

相关系数的平方

$$\rho^2(X_i, z_1) = v_i \quad (16)$$

注意表 5，除 $v_8 = 0.86$ 外，余下的 v_i 值均在 0.95 以上，说明(15)式中大部分表达式是较精确的，这意味着大部分指标近似程度不同地成比例变化．在表 5 中，z_1 与全体长 (X_2)、翅长 (X_4) 的相关系数均为 0.995，(15)式中的

$$X_2 = 0.36 z_1$$

$$X_4 = 0.35 z_1$$

几乎精确成立，说明指标 X_2 或 X_4 的值可以几乎精确地确定生长指数 z_1 的值．因此可以用全体长或翅长的观测值相对地衡量雏鸟生长的快慢和整体生长状况．翅长的变异系数

远大于全体长(表2),翅长随日龄相对幅度变化较大，用翅长观测值相对度量雏鸟生长较全体长为好.

三、小　结

1. 角百灵雏鸟的各生长度量指标之间显著相关,相关系数均$\geqslant$0.82.

2. 各生长度量指标的第一主成分为

$$z_1 = 0.35(X_1 + X_4 + X_7) + 0.36(X_2 + X_3 + X_5 + X_6) + 0.34X_8$$

$X_1, X_2, \cdots, X_8$ 分别是标准化后的体重、全体长、标准体长、翅、嘴峰、跗蹠、中趾和尾. z_1 反映全部生长指标携带信息的95%,定义为生长指数,作为雏鸟生长的综合度量指标.各生长度量指标 $X_i(i=1, 2, \cdots, 8)$ 对生长指数贡献几乎相等.

3. 雏期的前9天雏鸟生长较快,尤以第6—7天生长最快,第10天以后雏鸟的发育进入缓慢生长阶段.

4. $X_2 = 0.36z_1$ 和 $X_4 = 0.35z_1$ 几乎精确成立，因此可以用翅长或全体长相对度量雏鸟的生长状况.

参　考　文　献

郑生武,1984,角百灵雏鸟生长的初步观察,高原生物学集刊(2): 75—80。

Green, P. E. and J. D. Carroll, 1976, Mathematical tools for applied multivariate analysis. Academic Press. New York. San Francisco. London.

STUDIES ON THE GROWTH OF NESTLINGS IN THE HORNED LARK (*EREMOPHILA ALPESTRIC ELWEST*) II. THE PRINCIPAL COMPONENT ANALYSIS OF MEASURING INDEXES FOR NESTLINGS'S GROWTH

Zhou Li

(*Northwest Plateau Institute of Biology, Academia Sinica*)

The growth of nestilings is a complicated process, which depdnds on both genetic and external factors. Studies on the growth of nestlings have many measuring indexes, such as body weight (x_1), total body length (x_2), standard body length (x_3), wring (x_4), beak (x_5), tarsometatarsus (x_6), middle toe (x_7) and tail (x_8) (Tab. 1—2), however none of these indexes can completely describe the body growth of nestilings, but a side of the growth.

In order to use the multi-indexes quantitatively to measure body growth of nestilings, we need to search a comprehensive index which is a linear combination of the multi-indexes. Because there is a close correlation among the multi-indexes (Tab. 3), we use the method of principal component analysis to find the comprehenive index.

Write

$$X = (x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8)$$

X is a random vector. We construct a ortho-transform matrix L for X and write

$$Z = XL$$

where

$$Z = (z_1, z_2, z_3, z_4, z_5, z_6, z_7, z_8)$$
$$L = (l_1, l_2, l_3, l_4, l_5, l_6, l_7, l_8)$$
$$LL^T = I$$
$$l_i = \text{column vector } (i = 1, 2, \cdots, 8)$$

so that

$$RL = LB$$

where

R = the correlation matrix among random variables $x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8$.

B = a diagonal 8-order matrix

diag $B = (b_1, b_2, b_3, b_4, b_5, b_6, b_7, b_8)$

b_i = the eigenstruture value of the matrix $R(b_1 \geqslant b_2 \geqslant b_3 \geqslant b_4 \geqslant b_5 \geqslant b_6 \geqslant b_7 \geqslant b_8)$

therefor

$$Z = XL$$

is the principal component vector of the random vector X.

Based on multi-indexes data, we obtain the first principal component

$$z_1 = 0.35(x_1 + x_4 + x_7) + 0.36(x_2 + x_3 + x_5 + x_6) + 0.34x_8$$

which is called a body growth index. Its contribution rate is above 95% i.e. z_1 keeps above 95% information of original indexes (Tab. 4—5). Numerical values of z_1 with the passage of day age of nestlings can overall describe body growth of nestlings (Tab. 6), losing less than 5% information and decreasing information dimensions from 8 to 1.

According to values of z_1 (Tab. 6), body growth of nestlings in the horned lark is rapider during the first 9 days of the nestling period, specially, body growth of the 6th and 7th day of the nestling period is the rapidest.

In addition, we can know that the relative importance of the measuring indexes for measuring body growth of nestlings almost is equal one another from the expression of z_1.