数理统计与应用概率

Mathematical Statistics and Applied Probability

1988

第三卷 第一期

学术期刊出版社

《数理统计与应用概率》

1988 年 3 月　　（第 3 卷第一期）

编　　辑：　《数理统计与应用概率》编辑委员会
印　　刷：　北京工业大学印刷厂
主办单位：　全国工科院校应用概率统计委员会
北京工业大学应用数学系
长沙铁道学院科研所
出版单位：　学术期刊出版社
发行单位：　新华书店北京发行所
（本刊登记：　湖南省期刊登记证 160 号）

ISBN 7-80045-020-1/0.2　　**定价：1.50元**

目　录

应用成果

理论与方法

综　述

Contents

主成分估计与回归诊断在数量化计算中的应用

——也谈河南省马尾松人工林数量化立地指数表的编制

陶靖轩 （信阳师院数学系应用室）

李秀兰 （信阳师院图书馆）

摘 要

本文针对原河南省马尾松人工林数量化表的编制，利用主成分估计和回归诊断，删除了复共线性关系并诊断了模型中的异常点，强权点；纠正了原计算中若干错误，提出了更进一步精确的编制方法，并编制了新表。它为今后数量化Ⅰ型的计算提出了值得注意的几个问题。

引 言

马尾松是我省主要造林树种，为确定马尾松生长的立地类型及评价马尾松林地和宜林地的立地质量，以便科学地经营管理，适地、适树发展马尾松林业生产。一九八五年，我省马尾松协作组在广泛收集了信阳、南阳、驻马店等地区126块标准地资料的基础上，利用数量化Ⅰ型的方法，编制了河南省马尾松人工林数量化立地指数表。具体做法是：先依马尾松的生物学特征及生态学习性，结合本省自然条件特点，凭籍专业知识及实际经验选择了对马尾松生长影响较大的七个环境因子称作项目，如土壤质地、A层土厚、坡度等；将各标准地数据填入表1，制成标准地一览表：

表1 标准地一览表

标准地号	年龄	优势高 m	立地指数	A层土厚 (cm)	土壤厚度	坡向	坡度	坡位	土壤质地	石砾含量
1	12	5.6	9	24	44	W	4°	上	砂	多
2	12	5.8	9	32.2	50.2	EN	23°	上	中	少
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

而每一项目又可细分为若干类目(如土厚就可按不同厚度分为五个类目)，然后参照表1对这些定性因子引进示函数（即0—1数量化方法）实现定量推断；

收稿日期：1987年3月22日。

$$\delta_i(j,k)=\begin{cases}1 & \text{当第 } i \text{ 个样本第 } j \text{ 个项目的定性指标为第 } k \text{ 类目}\\ 0 & \text{否则}\end{cases}$$

（其中 $i=1,\cdots,n$； $j=1,\cdots,m$； $k=1,\cdots,r_j$； r_j为 j项目的类目数；n—样本数，m—项目数，p—类目总数 且 $\sum_{j=1}^{m} r_j=p$）

从而将各标准地的立地因子代入上式，然后将其函数值填入反应表，即可制成数量化反应表

表2　数量化反应表

标准地号	立地指数	X_1 土壤质地					X_2 A层土厚(cm)					X_3 坡度				X_4 坡向				X_5 石砾含量(%)				X_6 土厚(CM)			X_7 坡位		
		轻	中	重	砂	粘	0	1—5	6—15	16—24	25—	∧ 5°	6°—15°	16°—25°	26°—	S·WS	W·ES	E·WN	N·EN	0	1—5	6—25	26多	∧ 30	31—60	61—	上	中	下
1	9	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0	0	0	1	0	1	0	0
2	9	1	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

其次依此反应表中数据作成反应矩阵，利用LSE方法，对正规方程组 $X'Xb=X'Y$ 中参向量 b 进行估计。其中 $b=(b_{11}\cdots b_{1r_1},b_{21}\cdots b_{2r_2},\cdots b_{m1}\cdots b_{mr_m})'$，$X=(x_{ij})_{n\times p}$，$X'=(x_{ji})_{p\times n}$，$Y=(y_1,\cdots y_n)'$

X称为反应矩阵，Y即各 y_i作成列向量（具体为立地指数级）

计算出 $\hat{b}$ 之后再依各项目贡献之大小，依次剔除较次要项目，建立七个予测数模

$\hat{Y}=\sum_{j=1}^{m}\sum_{k=1}^{r_j}\delta(j,k)\,\hat{b}_{jk}$　（$m=1,\cdots,7$），并编制数量化表，使用者可按照一定精度要求任意选取予测方程及查对此表。

从第一次计算结果来看，28个参数估值为$\hat{b}_{11}=8.349$，$\hat{b}_{12}=3.280$，$\hat{b}_{12}=-8.811$，$\hat{b}_{14}=-10.655$，$\hat{b}_{15}=-9.598$，$\hat{b}_{21}=4.416$，$\hat{b}_{22}=4.896$，$\hat{b}_{23}=5.641$，$\hat{b}_{24}=1.336$，$\hat{b}_{25}=1.200$，$\hat{b}_{31}=0.591$，$\hat{b}_{32}=0.597$，$\hat{b}_{33}=1.687$，$\hat{b}_{34}=2.026$。$\hat{b}_{41}=2.263$，$\hat{b}_{42}=8.914$，$\hat{b}_{43}=9.682$，$\hat{b}_{44}=8.742$，$\hat{b}_{51}=8.668$，$\hat{b}_{52}=0.626$，$\hat{b}_{53}=-1.050$，$\hat{b}_{54}=-1.430$ $\hat{b}_{61}=-1.175$，$\hat{b}_{62}=-1.624$，$\hat{b}_{63}=4.429$，$\hat{b}_{71}=3.656$。$\hat{b}_{72}=3.414$，$\hat{b}_{73}=3.3$；复相关系数 $r=0.594$。

$F=9.181$，$\sigma=1.891$，虽然F检验效果显著（$F=9.181>2.02=F_{0.01}(22,100)$），但实际工作者认为复相关系数 r 太小，提出至少应在 0.7 左右；再者 126 块样地有32块残差相对值即$(y_i-\hat{y}_i)/y_i$超过20%；为改善这一状况，当时又作了第二次计算；首先算出当整套数据去掉一个样点时修正复相关系数新值，$R'_i=\sqrt{1-(1-R_i^2)\frac{n}{n-p}}$（在计算中曾假定因变量及各自变量的均值保持不变），这样从直观上找到了那些能使去掉后复相关系数增大最快的若干点。然后从数据里去掉了十个这样的“超常点”（样地）。第二次计算结果是：

$\hat{b}_{11}=2.087$, $\hat{b}_{12}=3.106$, $\hat{b}_{13}=2.711$, $\hat{b}_{14}=-0.134$, $\hat{b}_{15}=1.422$, $\hat{b}_{21}=-2.463$, $\hat{b}_{22}=-1.796$, $\hat{b}_{23}=-0.702$, $\hat{b}_{24}=3.083$, $\hat{b}_{25}=3.392$, $\hat{b}_{31}=1.914$, $\hat{b}_{32}=1.987$, $\hat{b}_{33}=-2.772$, $\hat{b}_{34}=-2.074$, $\hat{b}_{41}=-1.799$, $\hat{b}_{42}=4.021$, $\hat{b}_{43}=5.082$, $\hat{b}_{44}=2.949$, $\hat{b}_{51}=3.448$, $\hat{b}_{52}=9.948$, $\hat{b}_{53}=8.080$, $\hat{b}_{54}=8.491$, $\hat{b}_{61}=8.175$, $\hat{b}_{62}=7.860$, $\hat{b}_{63}=-1.445$, $\hat{b}_{71}=-2.509$, $\hat{b}_{72}=-2.709$, $\hat{b}_{73}=-2.844$复相关系数$r=0.7040$, $F=15.164$。$\sigma=1.704$，误差中值超过20%的有21块样地；从复相关系数看已基本达到要求，F 值检验也效果显著$F=15.164>2.045=F_{0.01}(28,96)$；于是就以此结果作为编表根据进行编制。但仔细考察计算结果却发现这样做对原结果改善不大，比如考虑到去掉了十块样地，这十块样地的误差中值均超过20%，故总共仍有31块样地误差中值超过20%，再从28个参数估计的符号来看有11个负值，原先仅 9 个，估计结果甚至更差一些。本文就从这里着手对进一步精确的编表方法进行探讨。

一、独立方程的补充与复共线性关系的删除

大家知道，利用0—1数量化方法，通常要先从反应矩阵中删去各项目（除第一项目外）的一个类目。这是因为在数量化回归中，由于从定性转化为定量表示时，各因素的若干水平总是相关的，$\sum_{k=1}^{r_j}\delta_i(j,k)=1$ $(i=1,\cdots,n;\ j=1,\cdots,m)$；使得$X$的列向量之间存在明显的线性相关。致使正规方程$X'X\beta=X'Y$的 p个方程并不相互独立，根据数量化回归的知识可知相互独立的方程至多只有 $p-(m-1)$个（见[1]）。由线性代数理论，这种方程组的解不唯一。因此要使 p个方程独立，至少需补充 $m-1$个独立方程：$X_{25}=0$, $X_{34}=0$, $X_{44}=0$, $X_{54}=0$, $X_{63}=0$, $X_{73}=0$, 实际上就是在计算时输入的数据划去 $m-1$ 列。原表的计算（不论第一、第二次）由于没有这一过程，所以估计的结果有理由认为是不可靠的；有鉴于此，在本次计算中我们先补充了 $m-1$个（即$7-1=6$)方程，即令$X_{25}=0,\cdots,X_{73}=0$,可以证明（见[1]）任意补充 $m-1$个独立的 x_{jk}的线性方程构成的p个独立线性方程并解得唯一解时，作出y 的予测结果不变。（予测结果的唯一性）。

在划出 $m-1$列后剩下的 22 列数据仍可能存在线性关系或近似线性关系——即多重共线性关系，而如果反应矩阵X的数据事先经过标准化，则估计的均方误差$\mathrm{MSE}(\hat{\beta})=\sigma^2\cdot\sum_{i=1}^{p}\frac{1}{\lambda_i}$（$\lambda_i$是$X'X$的规范化特征根，$i=1,\cdots,p$），在$X$的列向量间存在较强复共线性关系的时候，矩阵$X'X$的一部分特征根就可能变得很小，从而$\mathrm{MSE}(\hat{\beta})$变得很大，依判断估计优劣的第一个准则，这种估计有可能是相当恶劣的。为了对这一情况进行观测并作出及时处理，我们将解一般线性方程组的 LSE 方法改为主成分估计法，以便观察$X'X$的特征根从而直接删去多重共线性关系。通过计算，$\lambda_1=2.935$, $\lambda_2=2.371$, $\lambda_3=1.879$, $\lambda_4=1.793$, $\lambda_5=1.564$, $\lambda_6=1.458$, $\lambda_7=1.260$, $\lambda_8=1.217$, $\lambda_9=1.134$, $\lambda_{10}=1.054$, $\lambda_{11}=0.896$, $\lambda_{12}=0.823$, $\lambda_{13}=0.772$, $\lambda_{14}=0.683$, $\lambda_{15}=0.575$, $\lambda_{16}=0.504$, $\lambda_{17}=0.344$, $\lambda_{18}=0.328$, $\lambda_{19}=0.237$, $\lambda_{20}=0.124$, $\lambda_{21}=0.0495$, $\lambda_{22}=0.000000881$；由于 $\lambda_{22}\ll0.01$于是决定

再划去一列（实际上就是再补充一个独立方程$X_{15}=0$，为的是能求出$(X'X)^{-1}$以便进行回归诊断），并取20个主成分（因为λ_{21}也极小）。估计结果为：$\hat{b}_0=9.483, \hat{b}_{11}=1.287$，$\hat{b}_{12}=1.353$，$\hat{b}_{13}=0.861$，$\hat{b}_{14}=-1.029$，$\hat{b}_{21}=-1.168$，$\hat{b}_{22}=-0.678$，$\hat{b}_{23}=0.765$，$\hat{b}_{24}=0.609$，$\hat{b}_{31}=0.023$，$\hat{b}_{32}=-0.625$，$\hat{b}_{33}=-0.276$，$\hat{b}_{41}=0.475$，$\hat{b}_{42}=1.159$，$\hat{b}_{43}=0.195$，$\hat{b}_{51}=1.798$，$\hat{b}_{52}=0.116$，$\hat{b}_{53}=-0.279$，$\hat{b}_{61}=-0.053$，$\hat{b}_{62}=1.174$，$\hat{b}_{71}=0.396$，$\hat{b}_{72}=0.178$　复相关系数$r=0.593$，$\sigma=2.016$。

二、异常点诊断

原算法中将复相关系数增大最快的十个点作为超常点予以剔除，这当然也不是没有道理的。但这十个点是否就是异常点呢？如果仅仅为提高复相关而剔除。则可能把有些正常点（符合$y_i=X_i'\beta+\varepsilon_i$）剔去而将异常点（$y_i=X_i'\beta+\delta_i+\varepsilon_i$中$\delta_i\neq0$）保留下来，而这些点将对最后生成的模型产生破坏性的影响。为此我们参照回归诊断理论（见[2]）在计算程序中增加了以下统计量：(1) $H=X(X'X)^{-1}X'$，h_{ii}；当$h_{ii}>2m/n$，则第i个点为强权点。(2) $r_i=e_i/\hat{\sigma}\sqrt{1-h_{ii}}$（内学生化残差），当$|r_i|$较大（比如$|r_i|>2$），可怀疑为异常点。(3) $F(i)=(n-m-1)r_i^2/(n-m-r_i^2)\sim F_\alpha(1,n-m-1)$，当$F(i)>F_\alpha(1,n-m-1)$时可以认为是异常点。(4) Cook 距离：$D_i=r_i^2\cdot h_{ii}/n(1-h_{ii})$，当$D_i$较大时可判断为强影响点。(5) 误差中值$e_i/y_i$。(6) 残差$e_i$（$i=1,\cdots,n$）。将以上统计量利用$12_6$组数据，补充7个独立方程并取20个主成分进行计算，诊断统计量估值列出表3；从中可以看出$F(i)$列中$i=8,26,63,74,76,100,112$时有$F(i)>F_{0.05}(1;104)=3.94$，且残差$e_i$、$r_i$均显过大，又对照 Cook 距离$D_i$，怀疑这7块样地数据异常，经与河南省马尾松协作组同志协商后，决定从数据中剔除之。值得特别提出的是这七个样本点除$i=26$，63外并不是原先计算中复相关系数r增大最快的十个点。而原先 LSE 计算中出现的十个点中只有两个通过F检验可以定为异常点。另外从残差分析中我们还看到由于异常点的存在，强权点（观察h_{ii}）只有为数很少的一个（$i=80$）满足$h_{ii}>2m/n=42/126=0.333$；这从实际说明了异常点对线性模型的拟合精度影响是很大的。

剔除异常点后，仍补充7个独立方程使方程组有唯一解，并利用主成分估计监测复共线性关系，取$k=20$个主成分，结果精度有了显著提高：$\hat{b}_0=9.810, \hat{b}_{11}=1.188$，$\hat{b}_{12}=1.465$，$\hat{b}_{13}=0.601$，$\hat{b}_{14}=-1.325$，$\hat{b}_{61}=-1.593$，$\hat{b}_{62}=-1.073$，$\hat{b}_{51}=0.748$，$\hat{b}_{52}=1.155$，$\hat{b}_{53}=0.477$，$\hat{b}_{71}=-0.414$，$\hat{b}_{24}=0.073$，$\hat{b}_{31}=0.230$，$\hat{b}_{32}=1.353$，$\hat{b}_{33}=-0.134$，$\hat{b}_{21}=1.535$，$\hat{b}_{22}=-0.242$，$\hat{b}_{23}=-0.610$，$\hat{b}_{72}=0.705$，$\hat{b}_{41}=0.992$，$\hat{b}_{42}=0.227$，$\hat{b}_{43}=0.079$，$r=0.7063, \sigma=1.69$；误差中间值超过20%的仅20个样点，仅占15.8%。

三、项目的筛选和数量化表的编制

按照数量化表的一般理论，应以各项目的得分范围：$\text{range}(j)=\max\hat{b}_{jk}-\min\hat{b}_{jk}$，$1\leqslant k\leqslant r_j$，$(j=1,\cdots,m)$的大小作为评价该项目贡献的主要标准。于是将得分范围最小的项目首先剔除并再次上机计算，依次类推做下去，最后可得七个予测数模，这七个数模由于各个方程中项目数不同，各项目的类目得分也不同。为清楚地表达各项目中所有类目得分，把予

测数模的各参数估计列为表4，即为数量化立地指数表。从表4可以立即看出，影响河南省马尾松人工林生长的主要立地因子为A层土厚（贡献最大），土壤质地、坡度、石砾含量次之，其它为次要因子。

四、结　论

1. 该表通过复相关系数检验$r=0.7063>0.514=r_{0.01}(20,100)$差异显者，$\hat{\sigma}=1.69$说明建立的予测数模精度较高，可用于本省及环境条件类似的其它地区的林业生产。例如信阳县查山乡马庄大队林场马尾松人工林5号样地：16年生，优势木平均高7.6米、坡位上、坡度6°、坡向S·WS、土厚24cm、土质重砂、石砾含量较多、A层土厚大于25cm。

用7项因子：$9.81+1.188+1.593+0.743-0.414+1.353-0.610+0.079=10.561$

用6项因子：$9.656+1.294-1.625+0.679+1.336-0.699+0.0759=10.717$

用5项因子：$9.349+1.216+0.340+1.305-0.572-0.026=11.612$

用4项因子：$9.496+1.364+1.178-0.541+0.0747=11.581$

均属11米指数级范围；故该林分予估11米。16年时优势高(平均)7.6米，查立地指数表属11米指数级，予估值与实测吻合。

2. 该表编制方法比原表有了改进，且所用样地较多，剩余标准差较小；故可对林分生长作进一步精确的予测，并能适用于更广的地区。

3. 样本地继续剔除，虽可使r增大，但减小了模型的适用范围；并更难于适合估计的一些先决条件（独立性、方差齐性、线性性、正态）。

（本文所用数据来自信阳地区林科所、马尾松协作组）

参 考 资 料

[1] 潘德惠《数学模型的统计方法》，辽宁科技出版社，1986年5月；153页，156页

[2] 王松桂《线性回归诊断》Ⅰ、Ⅱ“数理统计与管理”1985年第6期，1986年第1期

[3] 张尧庭、方开泰《多元统计分析引论》科学出版社，1982年，322页

The Application of Main Composition Estimation and Linear Regression Diagnosis to Numeration Counting

Tao Jing xuan　Li Xiu lan

(Xin Yang Normal institute)

Abstract

This paper, based on the former Henan Province Pony-tail Pines Numeration Table, eliminates collinearity relationship by way of princi-

ple component estinate and linear regression diagnosis, and makes a comprehensive diagnosis of the outliers of the model, and puts forward a more accurate project as well as a new table. At the same also in this paper, some noticeable questions are put forward for the later numeration counting.

全国工科院校应用概率统计委员会将举办微机统计分析软件包推广应用班

为适应科学研究、经济管理和概率统计课程教学的需要，全国工科院校应用概率统计委员会交流培训部和合肥工业大学数力系，将于1988年3月中下旬在合肥工大计算中心联合举办微机用统计分析与数据处理软件包推广应用班。

该培训班将用一周的时间，讲授目前受欢迎的美国、日本和国内研制的统计分析与数据处理的通用微机软件。这些软件，几乎包括了所有常用的统计分析模型：一元描述量统计、相关分析、回归分析、因子分析、判别分析、聚类分析、方差分析、各种统计检验、时间序列分析等。

该班除讲解上述软件的功能和使用方法外，将安排参加者上机操作实习，学会使用，并且提供有关资料、手册和软件。

为了促进国内高等院校概率统计教学与计算机的结合，培训班将介绍有关统计教学的小型软件。

全国工科院校应用概率统计委员会，欢迎广大数学教师、应用概率统计工作者、工程技术人员、经济管理教师和经济管理工作者参加这个推广应用班，并欢迎有关单位和个人携带自己研制的较为成熟的软件赴会介绍。

参加者請同合肥工业大学数力系经济数学教研室联系

（应用概率统计委员会交流培训部）

研究变量之间关联性的信息论方法及其应用

潘国成 夏立显

（长春地质学院）

摘 要

多元统计分析中的一个重要课题是研究变量之间的关联性。特别是在数量化理论和多维标度法中，常常需要根据原始数据算出研究对象之间的相似性或非相似性计量，由此算出研究对象在低维空间中的标度。因此这种计量的定义方法一直受到人们的关注[1][7]，它直接关系到各种方法的应用效果。

我们这里以信息论中熵的概念为基础，对随机变量之间的关联性给出一种计量，论证它的一些优良性质，指出它的统计算法，将它与前人已有的一些关联性指标进行比较，最后给出一个在地质上的应用实例。

一、信息量和关联信息

为了叙述方便，这里着重讨论两个离散型随机变量之间的关联性，尽管有关的结论也都适用于连续型变量的情形。设有二维离散型随机变量（X，Y），它们的概率分布为

$$P\{X=x_i,\ Y=y_j\}=P_{ij},\ i=1,\cdots,r,\ j=1,\cdots,s$$

则边缘分布为

$$P\{X=x_i\}=\sum_{j=1}^{s}P_{ij}=P_{i.},\ i=1,\ 2,\cdots,r$$

$$P\{Y=y_j\}=\sum_{i=1}^{r}P_{ij}=P_{.j},\ j=1,\ 2,\cdots,s$$

$$\sum_{i=1}^{r}\sum_{j=1}^{s}P_{ij}=\sum_{i=1}^{r}P_{i.}=\sum_{j=1}^{s}P_{.j}=1$$

作为这些随机变量取值的不肯定程度的度量，按 Shannon 的定义[2]，它们的熵为

$$H(X)=-\sum_{i=1}^{r}P_{i.}\log P_{i.}$$

收稿日期：1987年1月13日。

$$H(Y)=-\sum_{j=1}^{s}P_{\cdot j}\log P_{\cdot j}$$

$$H(X,Y)=-\sum_{i=1}^{r}\sum_{j=1}^{s}P_{ij}\log P_{ij}$$

$$H_X(Y)=-\sum_{i=1}^{r}\sum_{j=1}^{s}P_{ij}\log\frac{P_{ij}}{P_{i\cdot}}\text{（条件熵）}$$

$$H_Y(X)=-\sum_{i=1}^{r}\sum_{j=1}^{s}P_{ij}\log\frac{P_{ij}}{P_{\cdot j}}\text{（条件熵）}$$

这些熵具有如下的众所周知的性质：

(1) $H_X(Y)\geqslant 0$, $H_Y(X)\geqslant 0$

(2) $H_X(Y)\leqslant H(Y)$, $H_Y(X)\leqslant H(X)$

(3) $H(X,Y)=H(X)+H_X(Y)=H(Y)+H_Y(X)$

(4) 若X与Y独立，则

$$H_X(Y)=H(Y),\quad H_Y(X)=H(X)$$

$$H(X,Y)=H(X)+H(Y)$$

X和Y之间的信息量为

$$I(X,Y)=H(Y)-H_X(Y)=H(X)-H_Y(X)\geqslant 0$$

它表示某一变量确定以后，另一变量不肯定程度的减少量，因此可以度量两个变量之间的关联性。于是有

定义1. 令

$$RI(X\to Y)=\frac{H(Y)-H_X(Y)}{H(Y)}=1-H_X(Y)/H(Y)$$

称它为X对Y的有向关联信息，类似地，Y对X的有向关联信息定义为

$$RI(Y\to X)=\frac{H(X)-H_Y(X)}{H(X)}=1-H_Y(X)/H(X)$$

这种关联信息实质上是相对的信息量，下面的几条性质表明它可以合理地度量变量间的关联性：

(1) $0\leqslant RI(X\to Y)\leqslant 1$, $0\leqslant R(Y\to X)\leqslant 1$

(2) X与Y独立$\Leftrightarrow RI(X\to Y)=RI(Y\to X)=0$

(3) $RI(X\to Y)=1$的充要条件是 $s\leqslant r$ 且有函数关系φ，使

$$Y=\varphi(X),\ a.s.$$

$RI(Y\to X)=1$的充要条件是 $r\leqslant s$且有函数关系 ψ，使

$$X=\psi(Y),\ a.s.$$

$RI(X\to Y)=R(Y\to X)=1$的充要条件是 $r=s$ 且有函数关系φ，使

$$Y=\varphi(X),\ X=\varphi^{-1}(Y),\ a.s.$$

前两条性质由熵的性质可明显地看出来，这里仅证明性质(3)。

由定义1可知，$RI(X\to Y)=1$的充要条件是$H_X(Y)=0$。由条件熵定义可知，这相

当于：对任意的 $i(i=1, 2, \cdots, r)$，有唯一的 j^*，使

$$P\{Y=y_j \mid x_i\}=\begin{cases}1, & j=j^* \\ 0, & j\neq j^*, \quad j=1, 2, \cdots, s\end{cases}$$

此即成立函数关系 φ，依概率 1 成立

$$y_j^*=\varphi(x_i)$$

由 i 的任意性可知

$$Y=\varphi(X) \quad a.s., \quad s\leqslant r$$

这就证明了性质(3)的第一条结论，同理可证第二条结论。由已证的结果，当 $RI(X\to Y)=RI(Y\to X)=1$ 时，X 和 Y 之间依概率 1 成立函数关系

$$Y=\varphi(X), \quad X=\psi(Y), \quad r=s$$

进一步可证 φ 和 ψ 互为反函数。设有

$$y_j^*=\varphi(x_i^*), \quad \text{即 } P\{Y=y_j^* \mid X=x_i^*\}=1$$

则有

$$\begin{aligned}P\{Y=y_j^*\}P\{X=x_i^* \mid Y=y_j^*\}&=P\{X=x_i^*, \ Y=y_j^*\}\\&=P\{X=x_i^*\}P\{Y=y_j^* \mid X=x_i^*\}=P\{X=x_i^*\}\\&=P_{i\cdot}>0\end{aligned}$$

于是必有 $P\{X=x_j^* \mid Y=y_j^*\}>0$，由 $H_Y(X)=0$ 可推出

$$P\{X=x_i^* \mid Y=y_j^*\}=1$$

此即

$$x_i^*=\psi(y_j^*)=\varphi^{-1}(y_j^*)$$

由 i^*，j^* 的任意性可知 φ 与 ψ 互为反函数。

在一般情况下，变量之间的两个有向关联信息是不相等的，它们反映了变量间的非对称的关联性。在许多实际问题中确实需要考虑这种关联性。例如考虑父子两代人职业上的联系以及其它的遗传因素，各种控矿地质因素间的关系，地层序列状态转移规律等都具有非对称性。由于在数学上处理非对称的关联性有许多困难，也由于在实际问题中经常只须考虑对称的关联性，我们可用两个有向关联信息的几何平均值作为变量间关联性的度量。

定义2. 令

$$\begin{aligned}RI(X, Y)&=\sqrt{RI(X\to Y)RI(Y\to X)}\\&=I(X, Y)/\sqrt{H(X)H(Y)}\\&=[H(Y)-H_X(Y)]/\sqrt{H(X)H(Y)}\\&=[H(X)-H_Y(X)]/\sqrt{H(X)H(Y)}\end{aligned}$$

称它为变量 X，Y 之间的关联信息，简记为 RI。

据前面证明过的有向关联信息的几条性质，也可以看出这里定义的关联信息具有类似的性质，这些性质表明用它来度量变量之间的关联性是很合适的。

(1) $RI(X, Y)=RI(Y, X)$

(2) $0\leqslant RI(X, Y)\leqslant 1$

(3) X 与 Y 独立 $\Leftrightarrow RI(X, Y)=0$

(4) $RI(X, Y)=\sqrt{H(Y)/H(X)}$ 的充要条件是有单值函数关系 φ，使

$$Y=\varphi(X), \quad a.s. \qquad s\leqslant r$$

$RI(X,Y)=\sqrt{H(X)/H(Y)}$ 的充要条件是有单值函数关系 ψ，使

$X=\psi(Y)$，$a.s.$ $\quad r\leqslant s$

$RI(X,Y)=1$ 的充要条件是有函数关系 φ，使

$Y=\varphi(X)$，$X=\varphi^{-1}(Y)$，$a.s.$ $\quad r=s$

上面仅就离散情形给出了关联信息的定义，其基本原理也适用于连续型随机变量的情形。我们既可以将连续随机变量离散化，然后再用上述定义；也可以根据连续型变量的熵，仿照如上方式定义关联信息。关联信息是按随机变量之间所提供的信息量的多少来度量变量之间的关联性的，它反映了变量间的各种可能的依赖关系。关联信息的重要特点是：$RI(X,Y)$数值的大小取决于Y对X和X对Y两方面的依赖关系，当且仅当X和Y互为函数关系时才有$RI(X,Y)=1$，否则恒有$RI(X,Y)<1$。

在实际问题中，可对(X,Y)进行n次独立观测，按X的r种观测结果和Y的s种观测结果对n次观测进行分类，得到一个$r\times s$列联表如表1所示。

表1 X—Y列联表

X \ Y	1	2	…	j	…	s	Σ
1	n_{11}	n_{12}	…	n_{1j}	…	n_{1s}	$n_{1.}$
2	n_{21}	n_{22}	…	n_{2j}	…	n_{2s}	$n_{2.}$
⋮	⋮	⋮		⋮		⋮	⋮
r	n_{r1}	n_{r2}	…	n_{rj}	…	n_{rs}	$n_{r.}$
Σ	$n_{.1}$	$n_{.2}$	…	$n_{.j}$	…	$n_{.s}$	n

根据表中所列的观测频数，我们可以得到(X,Y)的概率分布的最大似然估计

$$\hat{P}_{ij}=n_{ij}/n,\ \hat{P}_{i.}=n_{i.}/n,\ \hat{P}_{.j}=n_{.j}/n$$

$$i=1,2,\cdots,r;\ j=1,2,\cdots,s$$

由此可以得到关联信息的近似计算公式

$$\widehat{RI}(X\to Y)=1-\frac{\sum_{i=1}^{r}\sum_{j=1}^{s}\hat{P}_{ij}\log(\hat{P}_{ij}/\hat{P}_{i.})}{\sum_{j=1}^{s}\hat{P}_{.j}\log\hat{P}_{.j}}$$

$$=1-\frac{\sum_{i}\sum_{j}n_{ij}\log(n_{ij}/n_{i.})}{\sum_{j}n_{.j}\log(n_{.j}/n)}$$

$$\widehat{RI}(Y\to X)=1-\frac{\sum_{i=1}^{r}\sum_{j=1}^{s}\hat{P}_{ij}\log(\hat{P}_{ij}/\hat{P}_{.j})}{\sum_{i=1}^{r}\hat{P}_{i.}\log\hat{P}_{i.}}$$

$$=1-\frac{\sum_{i}\sum_{j}n_{ij}\log(n_{ij}/n_{.j})}{\sum_{i}n_{i.}\log n_{i.}}$$

$$\widehat{RI}(X,Y)=\frac{-\sum_{j=1}^{s}\widehat{P}_{\cdot j}\log\widehat{P}_{\cdot j}+\sum_{i=1}^{r}\sum_{j=1}^{s}\widehat{P}_{ij}\log\widehat{P}_{ij}/\widehat{P}_{i\cdot}}{\left(\sum_{i=1}^{r}\widehat{P}_{i\cdot}\log\widehat{P}_{i\cdot}\sum_{j=1}^{s}\widehat{P}_{\cdot j}\log\widehat{P}_{\cdot j}\right)^{1/2}}$$

$$=\frac{\sum_{i}\sum_{j}n_{ij}\log(n_{ij}/n_{i\cdot})-\sum_{j}n_{\cdot j}\log(n_{\cdot j}/n)}{\left(\sum_{i=1}^{r}\widehat{P}_{i\cdot}\log\widehat{P}_{i\cdot}\sum_{j=1}^{s}\widehat{P}_{\cdot j}\log\widehat{P}_{\cdot j}\right)^{1/2}}$$

二、关联信息的显著性检验

1. 相关系数检验

当随机试验次数增多，即样本容量增大时，服从多项分布的属性将渐近正态分布。而二元正态分布的相关系数为0等价于变量的独立性，因此可通过建立关联信息和相关系数的关系来实现对变量间的关联性的显著性检验。

设有两个标准化了的随机变量X，Y，其密度函数分别为

$$X\sim P(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2},\quad Y\sim P(y)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}$$

在$X=x$条件下Y的密度为

$$P(y|x)=\frac{1}{\sqrt{2\pi(1-r^2)}}e^{-\frac{1}{2(1-r^2)}(y-rx)^2}$$

其中r是X与Y的相关系数。由熵的定义有

$$H(Y)=-\int_{-\infty}^{+\infty}P(y)\log P(y)\mathrm{d}y=\log\sqrt{2\pi e}$$

$$H_x(Y)=-\iint_{-\infty}^{+\infty}P(x)P(y|x)\log P(y|x)\mathrm{d}x\mathrm{d}y$$

$$=\log\sqrt{2\pi e(1-r^2)}$$

则信息量为

$$I(X,Y)=-\log\sqrt{1-r^2}$$

这个关系式是仅就标准正态分布的情形得到的，但它完全适用于一般情况。当我们算出信息量$I(X,Y)$以后，可以按这公式算出相关系数r，再由对相关系数的显著性检验即可实现对变量关联性的显著性检验。

2. χ^2检验

对X与Y的独立性检验相当于检验假设

$$H_0: I(X,Y)=0$$

对于算得的信息量$I(X,Y)$，令

$$\chi^2 = 2nI(X, Y)$$

则 χ^2 渐近于自由度为 $(r-1)(s-1)$ 的 χ^2 分布[3]。于是可用这个统计量对变量间的独立性进行检验。

3. Kimball 分解检验

前面两种检验方法都是对两个变量关联性的总体检验，但一个变量的哪些可能的取值对另一变量的影响是主要的？这个问题可用 Kimball 法[4]进行研究。为简便起见，这里只引入 $2\times s$ 列联表分解的 Kimball 检验公式

$$\chi_k^2 = \frac{n^2(n_{t,k+1}S_k^{(l)} - n_{l,k+1}S_k^{(t)})^2}{n_{l\cdot}n_{t\cdot}n_{\cdot k+1}S_kS_{k+1}}$$

其中 l 和 t 是表中的任意两行，$k=1, 2, \cdots, s-1$。

$$S_k^{(l)} = \sum_{j=1}^{k} n_{lj},\ S_k^{(t)} = \sum_{j=1}^{k} n_{tj}, S_k = \sum_{j=1}^{k} n_{\cdot j},\quad l \neq t$$

这些统计量都服从自由度为 1 的 χ^2 分布。

Kimball 分解使得列联表的结构表现得更加清楚了，其基本思想是逐渐合并 2 行的各列，分别分解成 2×2 的表进行检验。比如有 2×4 的表，Kimball 分解检验的步骤是：先进行头两列的 χ^2 检验，然后再把这两列合并并同第 3 列又构成一个 2×2 表进行检验，再将这 3 列合并同第 4 列又构成一个 2×2 表进行检验……经过这种分解检验，就可以剔除一些极为次要的可能结果。如若头 3 列合并后同第 4 列构成的 2×2 表经检验后不显著，则认为前 3 列完全可以反映原变量的变化特点，第 4 列可以被剔除。

三、计算实例和关联信息权

为了说明关联信息的实际意义，我们重新研究[1]中引用的一组数据，如表 2 所示，这里给出了 8 个变量在 8 个样品上的取值，这些变量有的是二态变量，有的是三态变量。

表 2 原始数据表

样品 \ 变量	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1	1	0	−1	−1	1	−1	−1	1
2	1	0	−1	−1	1	−1	−1	1
3	1	1	0	0	−1	1	1	1
4	1	1	−1	−1	−1	1	1	1
5	1	1	1	1	−1	1	1	1
6	1	1	1	1	−1	1	0	0
7	1	1	1	1	−1	1	1	−1
8	−1	1	1	1	−1	1	0	−1

将这些变量两两组合，根据它们在 8 个样品上的取值构成一个如表1形式的列联表。例如变量对 (X_4, X_5) 相应的列联表如表 3 所示。根据这个列联表，据第一段所介绍的计算

公式，即可算出

$$RI(X_4 \to X_5)=0.576$$
$$RI(X_5 \to X_4)=0.332$$
$$RI(X_4,\ X_5)=0.437$$

表3 X_4，X_5的列联表

X_4 \ X_5	1	−1	Σ
1	0	4	4
0	0	1	1
−1	2	1	3
Σ	2	6	8

按这种方式，我们可以算出每两个变量之间的有向关联信息，列于表4中，其中第i行第j列元素表示$RI(X_i \to X_j)$ (i，$j=1$，2，…，8)，一般的关联信息在表5中给出。

表4 8个变量的有向关联信息

变量	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	1	0.069	0.098	0.098	0.069	0.069	0.196	0.226
X_2	0.103	1	0.332	0.332	1.000	1.000	0.541	0.157
X_3	0.254	0.576	1	1.000	0.576	0.576	0.437	0.423
X_4	0.254	0.576	1.000	1	0.576	0.576	0.437	0.423
X_5	0.103	1.000	0.332	0.332	1	1.000	0.541	0.157
X_6	0.103	1.000	0.332	0.332	1.000	1	0.541	0.157
X_7	0.540	1.000	0.466	0.466	1.000	1.000	1	0.495
X_8	0.540	0.252	0.390	0.390	0.252	0.252	0.429	1

表5 8个变量间的关联信息

变量	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
X_1	1	0.084	0.158	0.158	0.084	0.084	0.325	0.349
X_2		1	0.437	0.437	1.000	1.000	0.735	0.196
X_3			1	1.000	0.437	0.437	0.452	0.406
X_4				1	0.437	0.437	0.452	0.406
X_5					1	1.000	0.735	0.199
X_6						1	0.735	0.199
X_7							1	0.461
X_8								1

由表2中的原始数据可以看出，变量X_3，X_4在各样品上取相同数值，它们是同一变量，因此这对变量与其余变量有完全相同的关联信息。变量X_2，X_5，X_6的取值有一一对应关系，表4和表5的计算结果表明，它们和其余变量有完全相同的关联信息，其作用相当于一个变量。变量X_2，X_5，X_6可以看作是X_7的函数，因此X_7给它们提供的关联信息为1，即有

$$RI(X_7 \to X_i)=1,\ i=2,\ 5,\ 6$$

反之，X_7不是它们的函数，于是有

$$RI(X_i \to X_7)=0.541,\ i=2,\ 5,\ 6$$

它们仅能刻划X_7的“一半”的变化量，而X_7却能完全刻划它们的变化。它们的一般的关系信息为