第12卷 第3期
1994年 8月

武汉植物学研究
JOURNAL OF WUHAN BOTANICAL RESEARCH

Vol.12 No.3
Aug. 1994

植被生态信息的典范相关分析技术理论*

张全发 阎耀川 金义兴 郑重

（中国科学院武汉植物研究所 武汉 430074）

提 要 植被生态信息分析中对两（多）组取样集间的相关分析的需求导致了典范相关分析技术的应用和发展。本文对典范相关分析技术的原理、典范参数的生态学意义进行了揭示，比较了典范相关分析与PCA之间的差异，并对在典范相关分析技术的应用过程中，原始数据的预处理过程提出了建议。

关键词 典范相关分析，典范参数，PCA

植被生态信息的研究，通常是针对两组变量（取样）集：(1) 组成植被的植物种变量；(2) 环境数据变量。对这两组变量之间相关分析技术的探索，一直为植被生态学家所瞩目，然而仅有较少的多元统计分析技术得到发展[1,2]。

应用于植被生态信息处理中的多元分析技术，即是由直接梯度分析、分类和排序三者所构成的一整套分析数据的方法。三者的区别在于使用的目标和策略不同。直接梯度分析是用来展示有机体沿主要环境梯度的分布[2,3]，但实质上环境梯度是多个环境因子的复合体，因而它所揭示的是物种在多个因子相互影响的环境梯度上的分布。植被的分类，是人对植被带有明显的主观性的认识。随着计算机技术的发展，人们在人与植被之间加入计算机，形成人——计算机——植被的工作系统，但其对植被与环境之间的关系注意较少，只是用一多维坐标中的点来表示植被的样地，且将这一多维空间用超曲面来划分出多个对立于植被单位的离散的单元。一般的排序技术，是在代表植物种的二维或三维空间的坐标系内布置样方，力图通过自然群落取样来产生揭示群落与环境相互关系的一种数学生态技术[2,3]，现代大多数的研究工作是根据这种空间的散点图来推断植被潜在的环境梯度，并且多用于植被分类的解释[2]。

对两组变量之间的相关分析技术有多种[4-6]，但应用于生态学中的主要是多元回归分析[7,8]，但其并没有得到十分广泛的应用。原因是：①每个物种需要单独的分析，工作量大；②植被取样中物种有许多为零的数据，给分析造成了困难；③种与环境之间的非线性；④环境变量之间的非正交性，从而难以区别各环境变量的独立作用[9]；⑤同一组变量内因子间的相关信息的丢失。一般的线性模型解决了问题的②、③，但①、④、⑤

本文于1992年7月15日收到，1993年11月13日收到修改稿
* 中国科学院植被数量生态学开放实验室资助项目。

仍然存在。

典范相关分析是对两组变量间同时直接进行相关分析的技术，这种分析的目标是寻找几个排序轴，它们能最大限度地揭示两个矩阵联合或共同的结构[10]，这个目标非常接近植被生态学家对植被及环境之间相关关系研究的需要。

本文力图分析典范参数在典范相关分析中的产生过程，揭示在典范相关分析应用于生态学的实践中，典范参数所能代表的生态意义。

1 典范相关分析的原理及典范参数

1.1 典范相关分析的原理

假定有两组相互独立的变量X_1、X_2，典范相关分析就是对这两组变量相关分析的简单相关与多元回归分析的逻辑延伸，其目标就是找出这两组变量各自的一个线性函数，并且这两个线性函数之间的相关系数最大。这个技术也可认为是在由两组变量所构成的轴的多维空间中，研究总的分布规律。其详细的数学计算及推导过程见文献〔1〕。

1.2 典范参数

在典范相关分析的计算过程中，我们可以直接得到三个典范参数，即典范相关系数、典范权重、典范变量。

1.2.1 典范相关系数

典范相关系数是在对原始数据进行线性化的过程中，使成对的线性函数最大地相关的相关值。它具有如下特征：

(1) $|r_k| \leqslant 1$，且$|r_1| \geqslant |r_2| \geqslant \cdots\cdots \geqslant |r_p|$；其绝对值的大小表达了成对典范变量间的相关性程度，正负则表达了成对典范变量间的正反向关系；

(2) 在对原始数据进行非单一线性转换时，其值不变；

(3) 典范相关系数与一般的相关系数相似，表示为已解释的方差对所有方差的比率[11]。

1.2.2 典范权重a，b

典范权重是在对原始数据进行转换的过程中，每一取样特征对典范变量贡献大小的取值。它具有以下几项特征：

(1) 典范权重的绝对值大小表示各变量对典范变量的影响程度，其正负则为影响的正向或反向；

(2) 典范权重的值决定于原始变量的选择及其值的大小，变量的增加或减少对典范权重有极大的影响；

(3) 典范权重具有高度的不稳定性，即使是对同一对象进行重复取样。

1.2.3 典范变量U、V

典范变量是两组原始数据各自线性的组合，并使得两组之间（即成对典范变量之间）的相关系数最大。从其计算过程中[1]，我们知道它受到原始数据及典范权重大小的影响。

除上述三个典范参数外，我们下面引入一些新的典范变量。

1.2.4 结构相关

结构相关是指典范变量与原始变量的相关，这种相关的程度可通过诸如 Pearson's

相关系数的计算得出，其相关的数值大小，可以揭示出各原始变量对典范变量的影响程度及正负作用方向。其意义非常相似于典范权重，但它比典范权重更稳定；无论对原始数据进行标准化与否，原始数据的增减，或是重复取样，结构相关数值变化都较小。由于典范变量是对两组取样集的同时分析，因此结构相关可分为组内相关和组间相关。

1.2.4.1 组内相关

组内相关是指一组取样集与其相应的典范变量间的相关性。同样，组内相关亦有两组，在植被生态信息的分析中，即为植物和环境各分量间的相关性。

1.2.4.2 组间相关

组间相关是指一组原始数据取样集与来源于另一组原始数据取样集的典范变量的相关。同样，组间相关也有两组。在生态学上，即为植被原始数据与环境数据的典范变量的相关，环境原始数据与来源于植被数据的典范变量的相关。

1.2.5 含于典范变量中的方差（方差揭示度）

含于典范变量中的方差是指原始数据的所有方差中，典范变量所能揭示的那一部分。其计算过程为一组变量组内相关的平方和的平均值。

1.2.6 总残余量

残余量是指来源于一组变量的典范变量所解释的另一组变量的方差程度。其计算过程为一组变量组间相关的平方和的平均值。

总残余量则为一组变量的残余量的和。

1.2.7 方差分配

方差分配是指典范相关分析后，原始数据中所有方差的分配方式。其计算过程为一组变量的结构相关的平方和。它同样有两个。

1.2.7.1 组内分配

组内分配是指一组变量组内相关的平方和。理论上，全部的典范变量（P 对）的组内分配为 1，它保留了原始数据的全部信息，若取得K对（$K<P$）典范变量，则组内分配可表达出典范分析对一组变量信息的揭示程度。

1.2.7.2 组间分配

组间分配是指一组变量组间相关的平方和。它表示来自于另一组原始变量的典范变量的揭示程度。

2 典范参数生态学意义的评价

典范相关分析可产生许多参数[10]，下面就上文所列的 7 个参数进行植被生态信息的应用意义的评价。

2.1 典范相关系数与结构相关

典范相关系数是成对典范变量间相关性程度的度量[10]。但由于典范变量是原始变量的线性组合，且对于两组变量间是否存在着相关性，研究者都有一个先验的判断。因此，它除了得到一个定量的相关性系数外，对于植被生态信息处理中的单个环境因子（或多个复合环境因子）对某一物种（或植被类型）的影响方面，典范相关系数并没有多少分析价值。此外，典范相关系数还随着变量的增加或减少出现一定的波动，即使它对数据的转换是稳定的。它的较强的生态学意义在于计算过程中可以了解典范相关分析应具有

的维性（即取几对典范变量）。

结构相关是原始数据的变量与典范变量相关性的度量，其两种形式组内相关和组间相关都具有较强的生态学意义。典范变量是占有原始数据最大信息量的各变量之间的线性组合。而组内相关则是揭示这些信息量主要来源于哪一个或几个变量（植被生态信息中则为单个或几个物种或环境因子），这就可以得到各变量对典范变量的贡献的重要性程度，从而更稳定地表达了典范权重所应表达的生态学意义。组间相关则揭示两组变量中几个重要变量间的相互影响（植被生态信息中即为一个或几个环境因子对单个或几个物种的影响），这就揭示出了植被生态学家所需研究的物种与环境之间的相关关系的内容。此外，结构相关对于原始数据的增减、转换也较为稳定。因此，它在典范相关分析中是一个具有重要生态意义的参数。

2.2 典范权重

典范权重的特点相似于多元回归分析中的回归权重。我们知道典范权重的大小数值是指原始变量对具有典范相关 r 的成对典范变量的贡献大小，然而它用于生态意义的解释是很困难的，其原因是：①典范权重是在要求正交的条件下，使两组变量间组间协方差最大、而组内方差最小的一种组合；②典范权重与多元回归中的回归系数相似，依赖于变量的选择及其数值的大小。而在生态学的信息分析中，正交条件极难满足，并且其变动的幅度随变量的增减也产生较大的变化。因此，典范权重没有多大的生态学意义。

2.3 典范变量

典范变量相似于PCA中的主分量，二者都是原始变量数据的线性组合。典范变量的解释原来主要考虑的是原始变量对典范变量的贡献大小，并且这种贡献还得通过典范权重来表达。然而，由于生态学中数据的非独立性，这种解释通常是无效的[12]。因此，典范变量的生态意义在于极大地提取原始数据信息量的前提下，压缩原始数据信息的多维空间为一个低维（三维）空间，并作出图示，从而它更适用于分类的解释。

2.4 方差揭示度、总残余量

方差揭示度在生态意义上可用于分析典范相关分析过程中对原始数据信息的提取（占有）率。

总残余量为一组原始变量被另一组原始变量来源的典范变量所解释的方差。但它又不同于典范相关系数，相关系数表示各组变量的线性组合之间的相关，总残余量则是直接表示测量的原始数据之间的相关。因此，总残余量对于揭示物种与环境之间的关系具有重要的意义。

3 典范相关分析与PCA的比较

典范相关分析与 PCA 的分析对策存在着许多相似之处，二者都是要求在信息量损失最小的情形下压缩原始数据的变量空间，在轴的旋转过程中，使某些特征得到强调（其它的排序技术也与之相似），但PCA是对一组变量的方差分析，而典范相关分析是对两组变量的协方差分析。现将二者比较如下：

3.1 轴的导出

轴对于典范相关分析及 PCA 分别为求得典范变量和主分量。对于主分量分析，轴

的导出经过几个过程：①原始数据空间原点的形心化；②作一刚性旋转，使得第一个轴穿过点群最长的部分，即在第一轴上，样方点排序得分的方差被最大化了；③然后寻找下一个轴，令其垂直上一个轴，并占剩余方差的最大部分。典范相关分析的轴的导出，也存在着一个原始轴的旋转过程，不过其要求是成对的典范变量轴之间的相关最大。这样，二者在轴的导出上有相同之处：①各样方取样的位置没有改变；②新导出的轴都是一个线性组合；③前一个轴比后一个轴占有较多的信息量；④都有特征根的计算过程。但由于PCA分析存在着原点的形心化，因此相关系数会发生变化。

3.2 典范变量与主分量的解释

典范相关分析通过典范变量对结果的解释，与PCA中的主分量相似，亦可用图示表示。若在二维空间上，典范相关分析的两个轴可分别为：

①U_k与U_m，V_k与V_m（$k \neq m$）

②U_k与V_k

③U_k与V_m（$k \neq m$）

U_k与V_k代表成对的典范变量。

如在图示上，U_k与V_k为典范相关系数的图示，U_k与V_m（$k \neq m$）则是两组变量不相关的揭示，而较有意义的U_k与U_m，V_k与V_m（$k \neq m$）则与PCA分析后的图示相似，可用作对另一组变量梯度的揭示及分类。

3.3 群落模型与基础模型

群落模型通常为高斯模型（Gaussian model），即物种沿着一维环境梯度呈高斯曲线分布。但典范相关分析及PCA的基础模型均为线性模型，即认为物种沿环境梯度是呈线性变化的。这样，二者的分析都存在着非线性问题[13,14]。此外，典范相关分析还要求潜在梯度或因子的垂直性（不相关）[3]，这与群落模型中各梯度为不同因子相互作用的复杂现象不相适应，同时它成为早期典范分析在生态学中应用常得到模糊结论的原因之一[15-17]。

4 讨论

典范相关分析作为一种对两组变量间同时进行相关性分析的多元分析方法，看起来在理论上非常相似于植被生态学的研究目标。然而，在应用上成功与不成功实例的冲突，以及理论上群落模型与其基础模型之间的差距，具有生态意义的参数解释上的困难等方面的因素，仍给典范相关分析技术在生态学中广泛应用蒙上了一层阴影。

4.1 原始数据

植被生态信息分析的两组变量中，组成植被的植物种类较多（若种类较少时可直接运用多元回归分析），并且有许多为零的数据；且不同类型的环境数据在量纲化方面也存在着较大的差别，这给分析过程及其结果的解释造成了一定的困难。此外随着变量数目的增加，还会导致单个不重要变量间的偶然相关，从而出现明显较高的相关系数和没有意义的典范变量[10,18]。因此，对原始数据有必要进行一些分析前的处理。

（1）原始数据的非量纲化。数据的非量纲化有多种方法[1]，这种数据的转换过程，其目的是为了减少由于变量在量纲上的不同而出现变量值在分析过程中所产生的差异。

（2）减少原始数据。由于典范变量是一个线性的组合，较多的变量则对结果产生不

利的影响。虽然减少数据会损失一些信息，并且会对典范权重、典范变量的值产生变化，但对于结构相关、方差揭示度的影响较小；另一方面，减少的数据主要是对物种，且出现于少数样方中的物种（尤其是伴生种，受种源等因子的影响）具有一定的偶然性，与环境因子之间的必然相关较小。数据的减少过程也可通过PCA与典范相关分析联合进行[17]。这样，通过PCA分析减少原始数据，会最大限度地保留其信息。

(3) 原始数据的线性转换。典范相关分析与PCA相似，是从群落的非线性数据中寻找线性特征[14,19]，因此，有必要对分析前的数据进行线性转换。这种转换可通过PCA探索，根据数据特点进行中心化或非中心化[1]。

4.2 结果分析

典范相关技术与其它一般的排序技术的最大区别在于它是对两组数据同时分析，因而它能更为直观地反映出两组数据间的相关关系，但它必须通过一些典范参数来进行生态学意义上的揭示。本文对几种典范参数的分析认为，结构相关、典范变量、方差揭示度、总残余量是一些较为有意义的参数。这些参数用以揭示植物生态信息中物种与环境的关系，但典范变量最好用于作图示，它能直观地表达分类的意图。

参 考 文 献

1 阳含熙，卢泽愚. 植物生态学的数量分类方法. 北京：科学出版社，1981
2 Whittaker P H 著，王伯荪译. 植物群落排序. 北京：科学出版社，1986
3 Gauch H G Jr 著，杨持等译. 群落生态学中的多元分析：北京. 科学出版社，1989
4 Noy-Meir I, Anderson D J. Multiple pattern analysisor multiscale orclination, towards a vegetation hologram? In Statistical Ecology. G. P. Patil, E. C. Pielou and W. E. Water eds. Pennsylvania State University Press, University Park, Pennsylvania. 1973, 3: 207—225
5 Macnaughton-Smith P. Some statistical and other numberical techniques for classifying individuals. HMSO, London, 1965
6 Orloci L. Multivariate Analysis in vegetation research. Second Edition Junk. The Hague. 1978.
7 Austin M P. Role of regression analysis in plant ecology. Proceeding of the Ecological society of Australia, 1971, 6: 63—75
8 Yarranton G A. A quantitative study of the bryophyte and macrolichen vegetation of the Dartmoor granite. Lichenologist, 1967, 3: 392—408
9 Gauch H G Jr, Chase G B, Whittaker R H. Ordination of vegetation samples by Gaussian species distribution. Ecology, 1974, 55: 1382—1390
10 Gittin R. Ecological application of canonical correlation analysis in Multivariate Method in Ecological Work. eds. L. Orloci, C. R. Rao & W. M. Stitler. Burtonsville, Md, International Cooperative. 1979: 309—535
11 Bartlett M S. Multivariate analysis. Journal of the Royal Statistical Society, Series B, 1947, 9: 176—197
12 Meredith W. Canonical correlation with fallible data. Psychometrika, 1964, 29: 55—65
13 Goodall D W. Objective methods for the classification of vegetation. III An essay in the use of factor analysis. Aust J Bot, 1954, 2: 304—324
14 Swan J M A. An examination of some ordination problems by use of simulated vegetation data. Ecology, 1970, 51: 89—102
15 Austin M P. Models and analysis of deseriptivs vegetation data. In Mathsmatical Models in Ecology. ed. by Jeffers, J. N. R. Blackwell, Oxford. 1972. 61—86
16 Dale M. On objectives of methods of Ordination. Vegetation, 1975, 30: 15—32

17 Gauch H G Jr, Wentworth T R. Canonical correlation analysis as an ordination technique. Vegetation, 1976, 33 : 17—22
18 Miller J K. The sampling distribution and a test for the significance of the bimultivariate redundancy statistic, a monte carlo study. Multivariate Behavioral Research, 1975, 10 : 233—244
19 Gauch H G Jr, Whittaker R H, Wentworth T R. A comparative study of reciprocal averaging and other ordination techniques. J Ecol, 1977, 65 : 157—174
20 Pielou E C 著，石绍业，陈华豪译. 生态学数据的解释. 哈尔滨：东北林业出版社，1978

THE THEORY OF CANONICAL CORRELATION ANALYSIS IN ECOLOGY

Zhang Quanfa Yin Yaochuan Jin Yixing Zheng Zhong

(Wuhan Institite of Botany, The Cihinese Academy of Sciences Wuhan 430074)

Abstract The correlation analysis between two sets of variates results in the development of Canonical Correlation Analysis(CCA). In this paper, the theory of CCA and the explanation of canonical parameters are reviewed, a few new parameters for the explanation of CCA in ecology also are developed.

Key words Canonical correlation analysis, Canonical parameters, PCA