

3 梯度分析(排序)的基础(Basics of gradient analysis)

群落物种组成数据的分析方法通常有两种：梯度分析（排序）和分类方法（聚类、TWINSPAN 等）。这里的梯度分析是广义的梯度分析，泛指任何以揭示物种组成数据与实测或潜在的环境因子之间关系的方法。

植物群落学上的分类方法，用来划分群落的间断分布（有时也称 Clement 超有机体）。而梯度分析主要是用来描述群落的连续分布（也称 Gleason 个体论）。这种认识是传统上的认识，也是关于排序与分类的历史由来。然而，现在这种认识已经不够准确。这些方法可以相互补充，到底用什么方法，完全取决于研究的目的。打个比方，制作植被图的时候，植被分类是基础。如果植被类型之间并没有明显的边界，我们不得不切开这种连续分布，以产生可以分辨的植被类型单元用于植被制图。通过群落内物种组成的排序，可以让我们找到植被内部的相对间隔的地方，或是找出过渡类型。这些方法现在是普遍被植物社会学接受。当然，今天，排序也不仅仅用于植物群落分析，大部分考察物种组成分布与环境因子关系的群落学研究，都可以用到排序。实际上，梯度分析扩展应用已经延伸到以植被科学为基础的学科，比如，用于水生生物学的研究就是很好的例子（see the bibliographies by Birks et al. 1996,1998）。

3.1 梯度分析（排序）的技术(Techniques of gradient analysis)

表 3-1（见原书）展示了研究问题类型与排序类型的对应关系。排序方法的选择主要取决于两方面，一是用于分析的变量数量，二则是期望的结果（也就是说我们需要回答什么问题）。

回归的分析目标是发现一个一元响应变量（a univariate response）的变化（通常是物种的多度，一个群落的综合特征，比如多样性或生物量）对于解释变量（环境因子）的依赖程度。当环境因子为因子时，方差分析（ANOVA）实际上也是回归分析。如果环境因子与物种组成建立起回归关系，我们不仅可以通过环境因子预测物种的分布，同样可以通过物种组成数据来估计环境因子的值。当然，在预测之前，我们首先要建立起回归关系，这也需要预先知道物种数据与被估计的环境因子的之间的关系。（有关数据与方法的选择，见帖子：<http://www.planta.cn/forum/viewtopic.php?t=15624>）

简单排序（间接梯度分析）的目标就是发现这样的坐标轴，让群落中的样方或是物种的最大变化量能够在坐标轴上体现出来。换句话说，让尽可能多的变化量能够在尽可能少的轴上展示出来，并且让样方或物种在排序图能够可视化展示出来。当然，我们会经常期望这些轴能够代表一些潜在的环境变量。而约束排序（constrained ordination）的目的就是发现物种在环境梯度上的变化情况。简单排序和约束排序都有偏分析版本（partial versions）。在偏分析中，我们可以预先剔除由协变量产生的那部分物种的变化量，然后再通过排序展示出剩下的变化量。当然，还有混合排序分析（hybrid ordination），可以让前面若干排序轴（一般是前 1、2、3 轴）是约束排序，后剩下的轴是非约束排序（间接梯度分析）。

3.2 物种响应环境梯度模型（Models of species response to environmental gradients）

物种对连续的环境梯度的响应方式很多种，也可以用很多图像来表示。现代回归方

法可以通过很多分布模型来展示物种与环境因子的关系（见第八章）。所有排序方法都是基于一定的模型之上，这种模型反映植物种和环境之间的关系以及在某一环境梯度上的种间关系。最常用的关系模型有两种：一种是线形模型（linear model），另一种是单峰模型（unimodal model）。线形模型的含义表示某个植物种随着某一环境因子的变化而呈线性变化或叫线性响应（linear response）。单峰模型的含义是某个植物种的个体数随某个环境因子值的增加而增加。当环境因子增加到某一值时，植物种的个体数达到最大值，此时的环境因子值称为该种的最适值（optimum）；随后当环境因子值继续增加时，种的个体数逐渐下降。为了简化单峰模型，我们经常假设单峰曲线以峰值的为中心，两边是对称的（见Hutchinson 1957,资源梯度与物种最适值）。在排序之前，我们首先决定用那种模型。大部分情况下，两种模型都仅仅是实际数据的近似拟合，所以我们要决定的那种模型对我们所要分析的拟合更好。其实，单峰模型也是理想化的模型，在现实中，物种对环境才响应很少以最适值为中心两边对称的，更复杂的响应曲线也经常出现（比如双峰模型）。

一般来说，在短的梯度，线性模型拟合更好；在长的梯度，线性模型效果就欠佳了（图 3-1，图表翻译成中文的意义不大，故直接从原文 copy 过来），单峰模型更适合长的梯度。

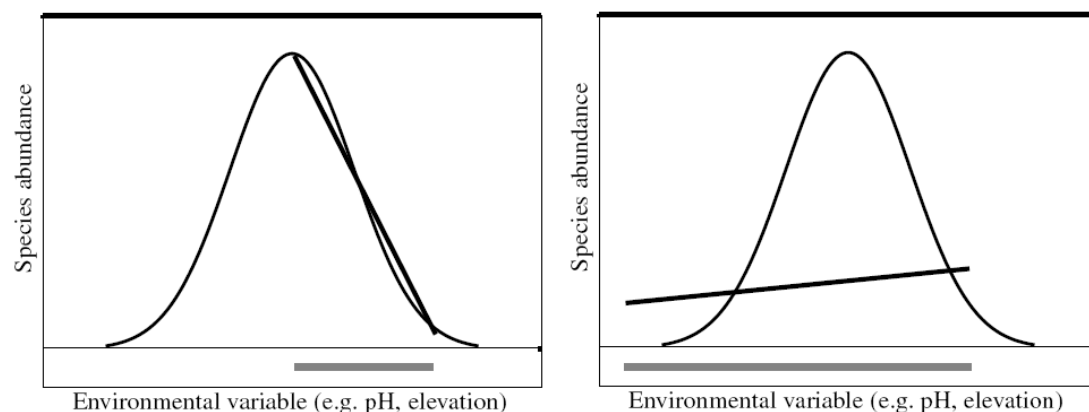


Figure 3-1. Comparison of the suitability of linear approximation of a unimodal response curve over a short part of the gradient (left diagram) and over a larger part of the gradient (right diagram). The horizontal bar shows the part of the gradient we are interested in.

3.3 通过加权平均估计物种最适值 (Estimating species optima by the weighted averaging method)

线性响应模型经常可以通过传统的方法（最小二乘法）回归拟合。但对于单峰响应模型，估计物种在环境梯度上最适值最简单的方法就是通过基于所有包含该物种的 n 个样方中环境因子值的加权平均得到。具体算法如下：

$$WA(Sp) = \frac{\sum_{i=1}^n Env_i \times Abund_i}{\sum_{i=1}^n Abund_i}$$

上式中， Env_i 是第 i 个样方的环境因子值， $Abund_i$ 是物种在 i 样方的多度，如果需要，物种的适应度（钟型曲线的宽度）也可以下面这个公式计算这个宽度的指标类型标准差（在第 10.2 节里的，算轴梯度的长度，也是用这个公式算的）

$$SD = \sqrt{\frac{\sum_{i=1}^n (Env_i - WA(Sp))^2 \times Abund_i}{\sum_{i=1}^n Abund_i}}$$

当物种占据着尽量多的样方时候（每个梯度上都有分布），加权平均的办法效果更好。图3-2和表3-2也展示了当物种分布于整个环境梯度的时候，最适值的估计是准确的。

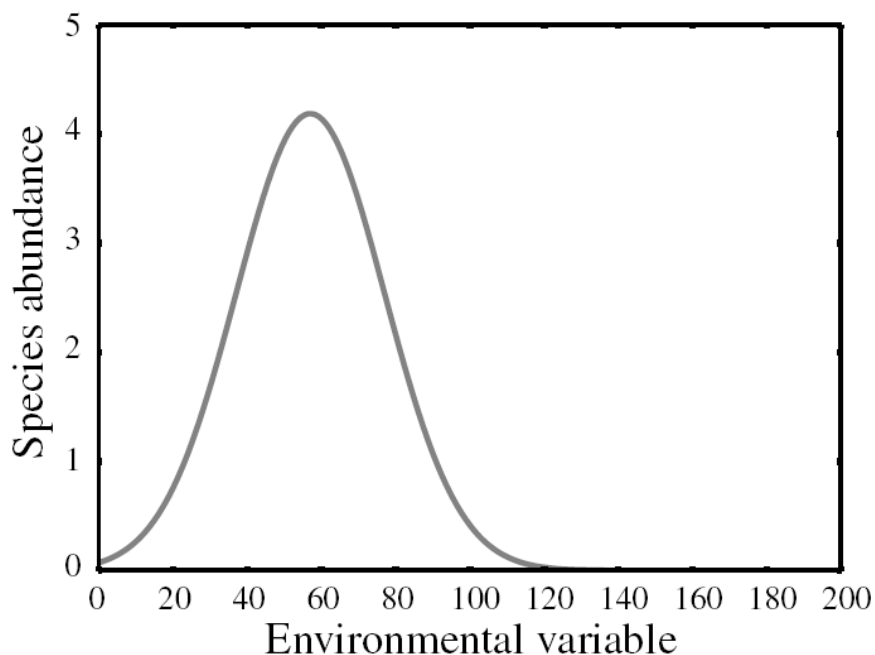


Figure 3-2. Example of a unimodal species response with a range completely covering the response curve.

Table 3-2. Estimation of species optimum from the response curve displayed in Figure 3-2, using a weighted averaging algorithm when a complete range of species distribution is covered

Environmental value	Species abundance	Product
0	0.1	0
20	0.5	10
40	2.0	80
60	4.2	252
80	2.0	160
100	0.5	50
120	0.1	12
Total	9.4	564

$$WA(Sp) = \frac{\sum_{i=1}^n Env_i \times Abund_i}{\sum_{i=1}^n Abund_i} = 564/9.4 = 60$$

另一方面，如果物种仅仅分布于部分的梯度之上，最适值的估计是有偏差的。比如跟表 3-2 同一个例子，如果仅仅拿右半部分的数据来加权平均，最适值会向尾部转移，造成估计偏差（表 3-3）。

Table 3-3. Estimation of species optimum from the response curve displayed in Figure 3-2, using a weighted averaging algorithm when only a part of the range of species distribution is covered

Environmental value	Species abundance	Product
60	4.2	252
80	2.0	160
100	0.5	50
120	0.1	12
Total	6.8	474

$$WA(Sp) = \frac{\sum_{i=1}^n Env_i \times Abund_i}{\sum_{i=1}^n Abund_i} = 474/6.8 = 69.7$$

当被覆盖的梯度很短的时候，大部分物种的分布幅度不能被体现出来，这也会造成最适值的估计偏差。所以如果轴的越长，代表的梯度越长，估计出来的物种最适值也比较准确。如果数据比较均一，短的梯度也是合适的。所以，线形模型更适合比较同质的数据，而单峰模型更适合比较异质性的数据。

模型选择的决定经常取决于 DCA 的梯度长度的分析，这个梯度分析展示了物种组成数据的异质性（见 4.3 节）

3.4 标定 (calibration)

标定是环境因素标定的简称，标定的目标是用植物种的测量值去推断环境因子值的数量方法。由于植物和环境相互作用的生态关系，植物种对环境的指示作用是常见的，对某一环境因子有明显指示作用的植物种在生态学上称做该因子的指示种 (indicator species)。

最常用的标定方法是也是加权平均。如果要通过加权平均来标定，首先应该获得该物种的在所标定的环境因子的最适值。这个估计最适值有时被称为指示值。例如，在中欧，大部分物种的对于光照、养分含量、土壤湿度等其他环境因子的指示值现在都已经可获得了 (e.g. Ellenberg 1991)。一个样方的环境因子的值可以通过该样方内所有物种的指示值跟它们多度的加权平均获得。

$$WA(Samp) = \frac{\sum_{i=1}^s IV_i \times Abund_i}{\sum_{i=1}^s Abund_i}$$

上式中 IV_i 是第*i*的物种的指示值(大概等于它的最适值), $Abund_i$ 第*i*物种在此样方的多度。这个算法是基于物种存在最适值(具有单峰响应)。当然, 基于线形模型的标定也是可以的, 只不过很少用加权平均来算。表3-4是给的是标定计算过程的例子。

Table 3-4. Estimation of nitrogen availability for two samples, Sample 1 and Sample 2

	Nitrogen IV	Sample 1	IV × abund.	Sample 2	IV × abund.
<i>Drosera rotundifolia</i>	1	2	2	0	0
<i>Andromeda polypifila</i>	1	3	3	0	0
<i>Vaccinium oxycoccus</i>	1	5	5	0	0
<i>Vaccinium uliginosum</i>	3	2	6	1	3
<i>Urtica dioica</i>	8	0	0	5	40
<i>Phalaris arundinacea</i>	7	0	0	5	35
Total		12	16	11	78
Nitrogen (WA)		<u>1.333</u>		<u>7.090</u>	
		(= 16/12)		(= 78/11)	

The nitrogen indicator values (IV) are listed first (according to Ellenberg, 1991). The a priori known indicator values are shown in bold style (in the Nitrogen IV column), the collected data and the calculations are shown in standard typeface.

选用指示值还是选用标定的算法来估计环境因子, 在生态学家之中意见常常难以统一。有人主张直接测量环境因子比标定得到的结果更可靠。然而, 对于历史纪录的样本(比如过去做的样方, 还有孢粉数据等等, 标定的结果虽说不一定准确, 但它的确是唯一的办法。标定也是一种复杂的分析, 也是排序很重要的一个步骤。

3.5 排序 (Ordination)

非约束排序可以下面两种方式构建:

1. 在排序空间给样方布局, 让样方间距离的远近度量代表它们之间物种组成的相似度。这类排序方法称作多维标定排序 (Multidimensional scaling) (Kruskal 1964; Legendre & Legendre 1998)。如果排序依赖于相异系数的数量值, 就叫有度量多维标定法 (metric multi—dimensional scaling); 如果排序仅仅决定于相异系数的大小顺序, 则称为无度量多维标定法 (Non—Metric Multi—Dimensional Scaling; NM—MDS)。

这种排序方式的目的是建立一个低维的空间, 让原来多维的空间影射到这个低维的空间来, 但让物种或样方空间关系失真最小。值得注意的是, 排序结果好坏很大程度上取决于我们如何所定义的最小失真。在无度量多维标定法, 我们不用影射规则, 而是相异系数的大小顺序, 通过数量算法求得样方间的距离(见 6.5 节)。

2. 发现潜在的变量(排序轴), 这个潜在的变量能够最好预测所有物种的值。这种途径要求预先知道物种与这些潜在变量的响应关系: 一般线形响应模型用于线形排序, 而单峰相应模型则用于加权平均排序。在线性排序, 样方的得分(score)是物种得分的线性组合(加权和)。在加权平均方法中, 样方的得分是物种得分的加权平均。(注: 加权平均的方式实际上已经包含了物种和样方的潜在在标准化。相对而言, 我们可以选择标准化和非标准化的线性排序。)

这两种排序构建方法可能殊途同归。比如, 主成分分析可以用欧几里德空间距离影射得到, 也可以通过线性的组合寻找排序轴的方法得到。其结果相差很小。

在 CANOCO 软件里，采用的是第二种排序的构建方式。排序的基本要领是能够被排序过程算法所解释。我们经常用的加权平均就是很好的例子。我们尽力去构建“潜在”的变量（第一排序轴），以致所有物种都能够在这个轴上得到最好的预测。排序的结果也就是让样方和物种的得分（坐标）在轴上体现出来。紧接着，我们要求物种的最适值能够通过样方得分的加权平均得到正确估计。这个过程如下：

第 1 步：给出任意的样方排序初始值 {xi}

第 2 步：通过 {xi} 的加权平均计算新的物种坐标 {yi}

第 3 步：通过标 {yi} 加权平均重新计算新的样方坐标 {xi}

第 4 步：对 {yi} 和 {xi} 进行修正，这是为了阻止排序坐标值在迭代过程中逐步变小。也是排除对去除任意初始值的影响。

第 5 步：回到第 2 步，重复迭代，直到两次迭代的结果基本一致为止。最终我们就得到样方和物种在第一排序轴上的坐标值。

Table 3-5. Calculation of the first ordination axis by the weighted averaging (WA) method. Further explanation is in the text

	Samp1	Samp2	Samp3	SPWA1	SPWA2	SPWA3	SPWA4
Cirsium	0	0	3	10.000	10.000	10.000	10.000
Glechoma	5	2	1	2.250	1.355	1.312	1.310
Rubus	6	2	0	1.000	0.105	0.062	0.060
Urtica	8	1	0	0.444	0.047	0.028	0.027
<i>Initial value</i>	<i>0</i>	<i>4</i>	<i>10</i>				
SAWA1	1.095	1.389	8.063				
SAWA1resc.	0.000	0.422	10.000				
SAWA2	0.410	0.594	7.839				
SAWA2resc.	0.000	0.248	10.000				
SAWA3	0.376	0.555	7.828				
SAWA3resc.	0.000	0.240	10.000				
SAWA4	0.375	0.553	7.827				
SAWA4resc.	0.000	0.239	10.000				

我们可以通过表 3-5 的例子来演示一下刚才迭代运算过程。这是一个有 3 个物种在 4 个样方的数据（用黑体表示）。首先给出样方的初始排序值（比如 0、4、10 斜体表示）。接着通过对样方初始值的加权平均，我们可以算得物种的第一次排序值 SPWA1。通过对 SPWA1 的加权平均，我们可以算得新的样方的坐标 SAWA1。此时你会发现轴的梯度越来越短（范围从初始的 0-10 变为现在的 1.095-8.063），这是为了阻止排序坐标值在迭代过程中逐步变小和剔除初始值对排序结果的影响，必须用下面公式对排序值进行标定。

$$x_{\text{rescaled}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \times \text{length}$$

这里的 Xmax 和 Xmin 分别是未标定前的排序值，而“length”是轴的长度。在上面例子中轴的长度是任意取的，是 10。但在一些排序方法中，轴的长度是真的，反映的是原始数据的异质性程度（详情请看 10.2 节 DCA 的 Hill 标度）。现在，我们可以比较一下原始值新的值 SAWA2，发现差别很大。我们需要继续迭代收敛为止（让 SAWNn 与 SAWn+1 的值基本相同为止，比如本列中，迭代到第四次，我们认为 0.240 与 0.239 就基本相等了）。

表 3-5 是才能从 orin.xle 这个文件拷贝过来的。如果改变一下初始值，你会发现最终 3 的结果跟初始值没有关系。所以，通过加权平均的迭代，我们可以分别获得物种和样方的第一轴的坐标。更多轴的推导也是跟第一轴算法一样，唯一不同的是，后面的轴要根据前面的轴

来确定初始排序值（也就是前面轴正交化）。

注解：线性模型算法也是按照上面的那样的步骤进行，只是将加权平均改为线性回归拟合最小二乘法

上面这些算法也可以用矩阵和特征根的代数语言来描述。为了实践的需要，我们经常要注意，轴在迭代过程收缩得越少（比如上例中 SAWA 和 SAWAresc 之间的差别越小），物种与排序轴的拟合就越好（更多的物种变量在这个排序轴被体现）。结果，这个值：

$$\lambda = \frac{x_{\max} - x_{\min}}{\text{length}}$$

被作为轴解释能力的度量，如果按照矩阵算法看，这个值被称为特征根。使用上面描述的方法，每个轴的建造，在前一轴构建之后，以便使它能够解释尽可能多的变量。结果，特征根是随着轴的次序逐渐降低的。

3.6 约束排序（Constrained ordination）

约束排序与非约束排序，通俗来讲，区别在于约束排序是确认最好的解释变量（在已有的环境因子基础上寻找），而非约束排序是寻找最能展示物种组成变化潜在的解释变量（这个变量就是排序轴）。约束排序的排序轴是环境因子的加权和（weighted sums）。约束排序的运算过程跟上面的非约束排序算法基本一致，只不过需要步骤 3 后加一步，

步骤 3b, 用多元回归法计算样方坐标与环境因子之间的回归系数，用回归系数重新计算样方的坐标（得分）。

由于回归系数是自变量的线性组合，所以，样方的新得分也是环境因子的线性组合。环境因子越少，约束就越严格。如果这个环境因子比样方的数量减去 2 的还多的话，排序就变成非约束排序（如果样方数量比环境因子数量少，那么回归系数就求不出来）。

非约束的排序轴是展示数据的最大的变化方向。而约束排序是展示数据在环境梯度轴上最大的变化量（能被环境因子解释的最大的变量）。因此约束轴的数量不会比环境因子的数量多。

3.7 基本的排序技术（Basic ordination techniques）

有 4 种基于物种与环境因子关系的排序的技术是最基本的（见表 3-6, 引自 Ter Braak & Prentice, 1988）。非约束排序也叫间接梯度分析，而约束排序也叫直接梯度排序分析。

Table 3-6. *Basic types of ordination techniques*

	Linear methods	Weighted averaging
Unconstrained	Principal components analysis (PCA)	Correspondence analysis (CA)
Constrained	Redundancy analysis (RDA)	Canonical correspondence analysis (CCA)

在加权平均的排序样方中，有去趋势的版本（如 DCA 和 DCCA，见 4.5 节），目的在于去除“马蹄形效应”。所有的排序方法，都有偏分析版本（*partial analyses*）。在偏分析过程，协变量的影响会预先被剔除，然后再分析剩下的变量的影响。约束排序的排序轴不会超过环境因子的变量。当我们只有一个环境因子的时候，仅只有一个排序轴是约束的，其它排序轴不会被约束。

混合排序（*hybrid analyses*）代表着排序过程，约束排序与非约束排序同时并存。在标准的约束排序，有几个环境因子，就有几个约束排序轴，剩下的排序轴都是非约束的。但在混合排序里面，你可以根据需要自己设定约束排序轴的数量（小于环境因子的数量），然后让约束轴后面的排序轴等成为非约束排序。

当约束排序轴被运算确定后，程序还可以继续分析，经常能够发现潜在的变量，这些变量甚至可以比前面用来约束的环境变量，解释能力还强。结果，在很多情况下，第一个非约束轴的解释能力比前面几个约束排序轴要强，也就是相应的特征根要比前面几轴特征根大。

3.8 排序图（*Ordination diagrams*）

排序的结果可以在排序图上展示出来。在所有排序图里面，样方经常用点来表示。在基于线性模型的排序图中，样方用箭头来表示（箭头的方向代表物种多度增加的方向），而在加权平均的排序方法中，物种用点来表示（该点表示物种的最适值的位置）。数量型的环境因子用箭头来表示（箭头的方向代表因子值增加的方向）。定性的环境变量用样方的质心（*cantroid*）来表示。关于排序图更详细的解释，可以参考第 10 章。

图 3-3（见原书）展示的四种基本排序技术所产生排序图。

3.9 两个途径（*Two approaches*）

当我们同时有环境数据和物种组成数据的时候，我们可以首先做非约束排序分析，看看物种数据整体的情况，然后再作约束排序（将环境因子加上到排序图上），当然可以直接进行约束排序。这两种途径是互补的，但两种方法必须都用。首先做非约束排序分析，我们才不会丢失物种组成变化量的大部分信息，如果直接做约束排序，很多不是由所测的环境因子的引起变化量就丢掉了。通过计算约束排序，我们不会丢掉与所测环境因子相关的大部分变量信息，但会丢掉与未测的环境因子相关的物种变量信息。

当你要发表你的排序结果的时候，你总是要特别小心确认排序的方法的使用。从排序图上，分辨约束排序与非约束排序有时难度很大。因为很多文章作者不按常规，在线性排序里面并不用箭头来表示物种，而且在说明里面，到底是线性还是单峰排序方法，模棱两可。

3.10 检验环境因子相关显著性（*Testing significance of the relation with environmental variables*）

在一般的统计检验中，数据的统计检验值来自实际分布与零和假设下期望分布的比较。基于这样的比较，我们可以得的分析所得的结果（事件）在零和假设条件下的发生的概率。实际结果的统计检验值的分布来源于原始数据的分布假设（这也是为什么在最小二乘回归中

我们要假设因变量的残差符合正态分布)。在 CANOCO 里面，独立条件下零和假设的统计检验值分布往往是未知的。这个分布取决于环境因子的数量及它们之间的相关结构，还有物种的多度分布等等。但这个分布可以被蒙特卡罗置换检验模拟 (Monte Carlo permutation test)。

在蒙特卡罗置换检验中，基于零和假设下统计检验分布估计可以按照下列步骤进行。蒙特卡罗置换检验零和假设是“响应变量（物种组成数据）独立于环境因子”。如果假设成立，表明所观测到的环境因子数据与物种数据之间毫不相干。既然环境因子数据与物种数据之间毫不相干，那么环境因子的值可以随机赋于任何一个样方也不会影响物种的分布。CANOCO 可以置换（移动）数据，这产生统计检验数据。利用这种方式，相应变量的分布和环境因子的相关结构能够保持在相同的真实数据与零和模拟数据之间。检验显著性水平（犯第一类错误的概率）可以用下面的公式计算。

$$P = \frac{n_x + 1}{N + 1}$$

上式 n_x 表示产生的置换数不低于随机置换分析的数量。N 地表总的置换数。这个检验是完全自由分布。这意味着检验完全不依赖于物种分布的假设检验。关于置换检验更详细的解释可以参考 Legendre & Legendre (1998) 第 20-26 页码。

这个置换的方案可以通过实验的设计的方案来规定。上面关于置换检验的描述是最基础的，其实在 CANOCO 里面，有多种置换方案。具体可以参考本书第五章和 CANOCO 的使用手册 (Ter Braak & Smilauer 2002)。

3.11 回归显著性的蒙特卡罗置换检验 (Monte Carlo permutation tests for the significance of regression)

置换检验与传统的分布检验的区别在于检验的分布自己产生分布，而传统检验是利用现成的分布来作为零和假设的分布（译者加）。

实际上置换检验能够检验很多类型的相关关系。为了展示这个检验算法，我们举一个最简单的回归模型显著性检验，这个例子是用置换检验来检验植物高度与土壤氮的浓度的关系。

我们测得 5 个植物的高度与氮浓度之间的对应关系（见表 3-7）。我们可以在植物高度与氮浓度之间建立回归关系，这个关系的显著性可以通过回归模型的方差分析的 F 值来衡量。在一些假设分布前提条件下（数据的正态分布），我们知道 F 值在独立零和假设条件下的分布。（自由度为 1 和 4 的 F 分布）。假设我们不能使用这个分布（比如原始数据不符合正态分布），我们可以通过随机给植物高度赋予氮浓度的，这样可以模拟一个分布。我们通过多次置换，每次可以算一个物种高度与氮的回归值（相应的 F 值）。当氮的值随机赋予植物的高度，那么 F 值的分布就相当于零和独立分布。回归的显著性可以按以下方式估计

$$\frac{1 + \text{no. of permutations where } (F \geq 10.058)}{1 + \text{total number of permutations}}$$

在 CANOCO 里面，F-ratio 的值跟回归模型里面方差分析 F 值是一个意思。另外，在 CANOCO 里面，蒙特卡罗检验带有对话框。

Table 3-7. *Example of permutation test for a simple linear regression*

Plant height	Nitrogen (as measured)	1st permutation	2nd permutation	3rd permutation	4th permutation	5th permutation etc.
5	3	3	8	5	5	...
7	5	8	5	5	8	...
6	5	4	4	3	4	...
10	8	5	3	8	5	...
3	4	5	5	4	3	...
<i>F</i> -value	10.058	0.214	1.428	4.494	0.826	0.###