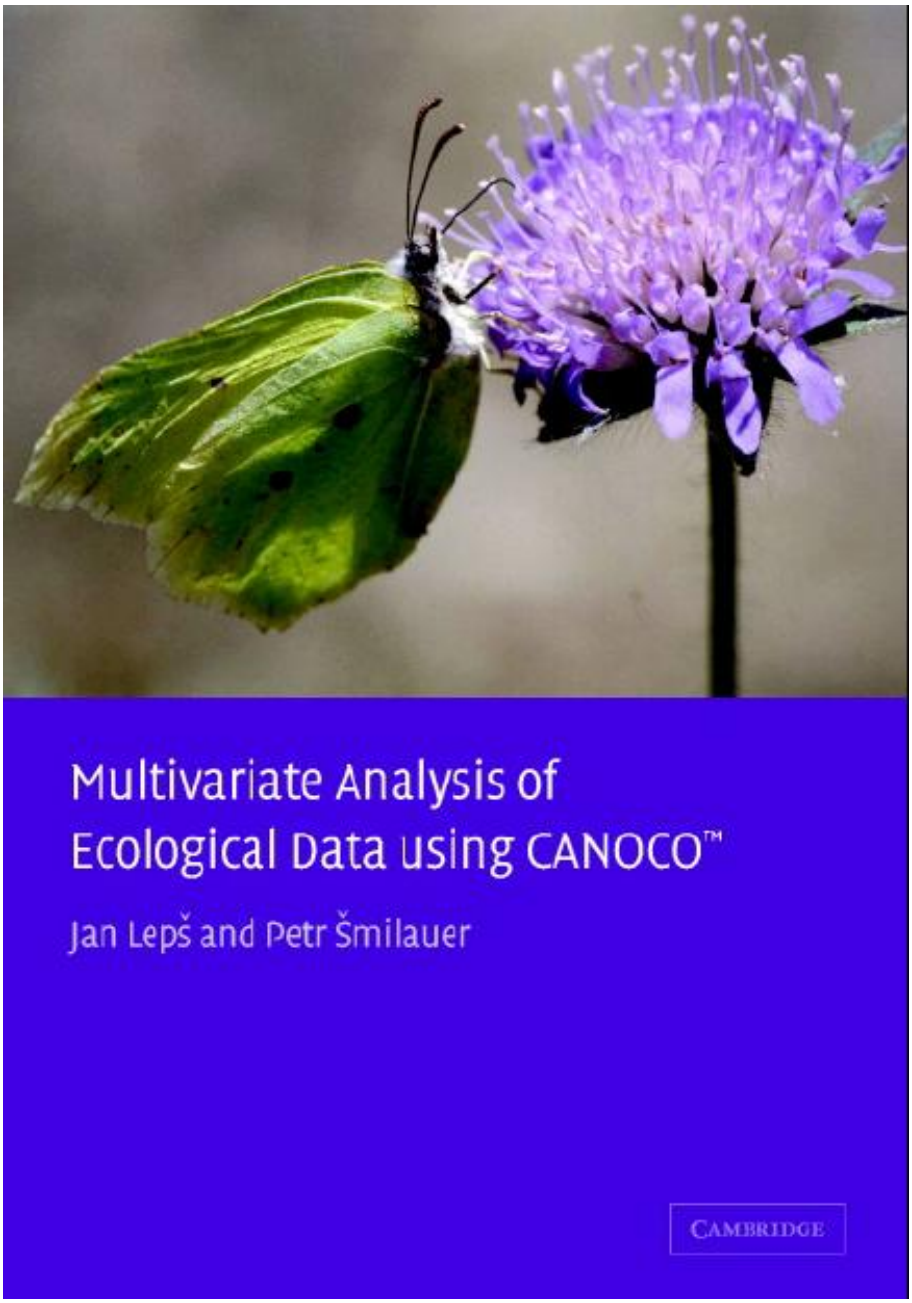

基于 CANOCO 的生态学数据的多元统计分析

著者：Jan Leps　捷克南波希米亚大学植物学系和捷克科学院昆虫研究所生态学教授
Petr Smilauer　捷克南波希米亚大学多元统计分析讲师
译者：赖江山　中国科学院植物研究所生物多样性与生物安全研究组助理研究员

这本书目的主要在于帮助生态学者分析野外观测数据和实验获得的数据。本书对于学生或研究人员处理复杂的生态学问题非常有用，比如生物群落随环境条件的如何变化，或是生物群落在控制实验中的变化。在简单介绍排序原理之后，本书的着重介绍约束排序方法（RDA 和 CCA）和置换统计检验在多元数据中的应用。同时介绍了如何利用分类的方法及现代回归技术（GLM，GAM，loess）来正确解读排序图。最后，用 CANOCO 软件分析了 7 个难度不同的研究案例。这些案例对于大家选择排序方法及分析排序结果很有帮助。案例的数据均可以从网络本书的主页（http://regent.bf.jcu.cz/maed/）上获得。

原书前言

群落的组成的多维数据，比如种群的属性，或是环境因子的属性，是生态学家研究生涯的面包与黄油。这些数据被分析时候需要考虑它们的多维性。用多元统计的方法来分析群落数据是比较适合的。

在这本书，我们尽量使用一套一致的方法来回答生态学家在研究中常遇到的问题。然而，我们也经常用自己观点来表述一些内容，同时，我们也关注一些非参数的方法，比如非度量多维尺度分析（NMDS）的算法等等。我们并不要是强调不同的方法对于分析多元数据的差异，而是想说明要解决一个问题，可以用很多方法。

在本书主要内容讲排序的方法，但并不意味着分类的方法没有用（译者注：排序与分类密不可分，分类分析群落的间断分布，排序分析群落的连续分布 ）。同时，我们也对回归方做了一些总结，包括最新发展的内容比如广义可加模型（generalized additive models）。

在这本书的所描述的方法可以广泛被研究植物、动物和土壤的研究人员利用，当然也可以是水生生物方面的人员。由于本书的两位作者的背景，本书的内容偏向植物生态学。

这本手册的材料原先是作为“生态数据多元分析”的课件。我们也希望这本书能用于其它相关类似的课程，也期望每个学生能够从这本书提高他们的分析数据的能力。

我们希望这本书可以作为 Canoco 4.5 使用手册的简明的补充材料。

Jan Lep 和 Petr Smilauer

译者前言

四年前，我开始接触 CANOCO 软件时候，也是菜鸟一个，自己并不是学统计出身，学习的过程也非常缓慢。当时也不会想到四年后今天，我居然还能为大家翻译这样一本有关 CANOCO 的书。这个过程，不得不承认普兰塔（www.planta.cn）的作用，正是为了回答塔友有关排序和 CANOCO 相关的问题，我不断翻文献、看软件说明书和自己摸索，积累了一点关于多元统计方法和 CANOCO 软件的一些知识。我相信很多的塔友的排序知识比我丰富，CANOCO 软件用得比我熟练，但至今不愿“出山”写本中文的参考书，哪怕是翻译一本也行。没办法，只能由我这个半桶水的家伙来承担此任务。

《士兵突击》的许三多有口头禅：“俺就是想做有意义的事情”。我一直觉得，翻译这本书就是非常有意义的事情，尽管现在的评价考核体系根本不会考虑我翻译了这本书，但我还是很乐意做这个事情。如果有很多 CANOCO 的初学者将从中受益，我也深感欣慰。当然，我发现每天早上上班和晚上下班前翻译一两段这本书还是一件很惬意的事情，而在翻译的过程，我也学到很多的东西。

由于本人非统计科班出身，而且时间比较仓促，翻译过程中可能有不少错误。有些统计学术语内容也可能斟酌不够，把握不准。因此，希望各个兄弟姐妹发现错误后给直接在博客里面回复，我会尽快修改。

赖江山

2009/6/25

1．导论和数据处理（Introduction and data manipulation）

1.1 为什么排序？（Why ordination?）

当考察植物或动物群落沿着一系列环境条件下的变化情况，我们经常发现在不同条件的群落不仅物种组成变化很大，而且这些变化往往具有连续性（consistency）和可预测性(predictability)。例如，我们为了要考察景观尺度上的草地群落的变化，可以通过观测样方内物种组成变化来描述，我们可以在一、两个或是三个虚拟坐标轴上将这些样方一个一个进行排列。当我们的目光在虚拟的排序轴上从一个样方移动到下一个相邻的样方，我们就会发现群落内物种组成变化通常很小。

群落中物种组成渐变跟每个物种对环境条件有着需求不同但又有重叠的生物学性质息息相关，这些环境因子如土壤平均湿度及随季节的波动变化、物种间竞争养分和光照能力等等。如果我们原来排列样方的虚拟轴走向恰好能反映某种环境因子的变化规律（比如土壤湿度或是土壤养分丰富度），这些排序轴便可以被称为土壤湿度梯度、养分梯度等等。这些环境梯度偶尔恰好又能跟实际的景观联系起来，例如土壤湿度梯度与河岸坡面，经常是沿着坡面从下到上，土壤湿度逐渐降低。但大部分情况下，我们并不能发现这些轴具体反映什么环境因子，或是反映什么空间变化，因此我们只能称这些轴为群落组成变化梯度。

生物群落的变化可以用很多统计方法来描述，但我们如果着重考察群落变化的连续性，所谓的“排序方法”是很好的选择。自从上个世纪五十年代开始，生态学家就开始用排序的方法分析生态学数据，经过半个世纪的发展，现在已经创制出种类繁多排序技术。我们利用刚才那个草地群落的例子来说明一下最简单的排序使用。当我们通过样方调查法来描述群落物种变化规律的时候，把样方数据总结在一个表格里面形成一个物种-样方矩阵，矩阵的列代表物种，行代表样方。如果用排序的方法分析数据矩阵并在排序图上表示出来（图 1-1），我们可以获得对于这个草地群落相当直观的认识。

排序图的解读规则将在随后的第 10 章进行讨论。但即使现在不知道这些规则，只要脑子里有群落连续性分布的思想和相似相近的原则（Proximity implies similarity），我们也能从这个图解读出一些信息。在图 1-1 中灰色的圆圈代表样方，我们可以相信如果两个样方在排序图上挨得越近，它们的物种组成和种间数量比例应该越相似。

在图 1-1 中用三角型代表物种。或许这些物种的生态学特征能够帮我们解读排序轴所表示的生态学梯度。有几个偏好丰富养分的土壤的物种（如 *Urtica dioica, Aegopodium podagraria, or Filipendula ulmaria*）排在图的右边，另外一些偏好养分匮乏的土壤的物种排在左边（如*(Viola palustris, Carex echinata, or Nardus stricta)*。因此，排序图中的水平轴（第一轴）可以解读为表示土壤养分的梯度，从左到右，养分越丰富。同样的道理，排序图下面几个物种（*如 Galium palustre, Scirpus sylvaticus, or Ranunculus repens)*比排在上面的一些物种（*如 Achillea millefolium, Trisetum flavescens, or Veronica chamaedrys*）更喜欢湿生环境。因此，纵轴（第二轴）可以解读为表示土壤的湿度梯度。

同样，或许大家都可以猜到，在排序图上，如果某一物种的越靠近某一个样方，表示该物种在此样方内个体数量越多。同样，两个物种离一个样方的距离也可以代表它们在该样方所占比例的差异，离得越近该物种相对数量越多。

上面的例子已经展示了排序方法对于群落分析的最基本的用途。通过排序分析，我们可以认识群落格局，也可以将排序轴跟我们已知的环境条件联系起来，看是否代表某一环境梯度。当然，也许我们必须用统计手段来检验排序轴到底是否真能代表环境因子

的梯度，比如，上面这个例子，我们可以这样问个问题：群落物种组成分布真的是随土壤湿度的变化，还是仅仅是一个巧合呢？通过约束排序法（constrained ordination methods）可以帮我们回答这样的问题。这些内容通通将要在本书的后半部分介绍。

然而，这本书并没有止步于仅仅用排序的方法来探讨上述这些简单的确定性的分析。这本书还介绍了各种类型的回归和方差分析，包括了固定样点重复观察的数据分析，空间结构数据分析和各种等级的方差分析。这些方法能够让生态学家探讨更复杂、更现实的科学问题。另外，这本书不仅是告诉如何分析问题，还手把手教大家怎么做。

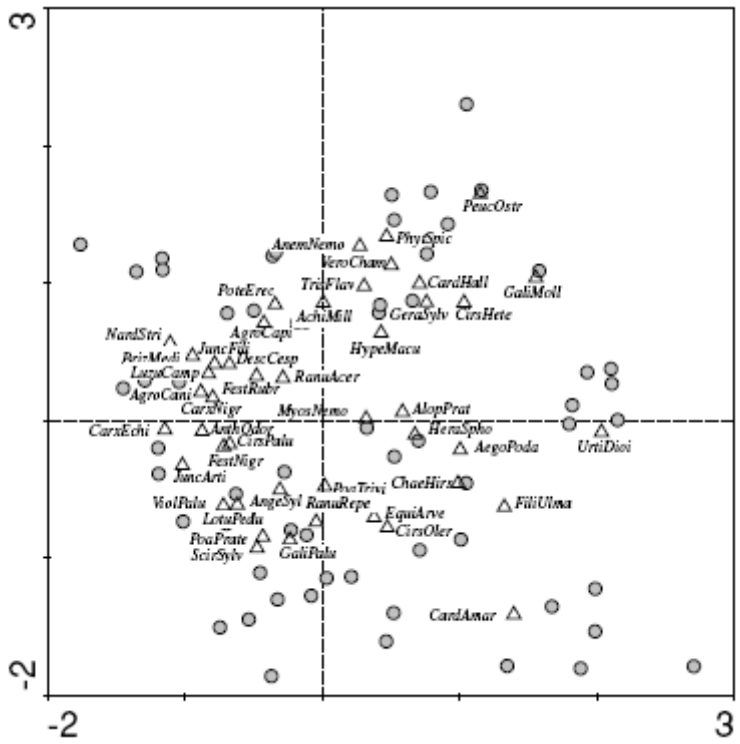

图 1.1 草地植被物种组成的 CA 排序图

1.2 专业术语（Terminology）

多元统计分析的专业术语非常复杂。本书内至少有两套不同的术语，一套是很多学科领域共同使用的、纯粹统计学术语，本书中我们将这部分术语斜体并放在括号里面；另外一套是群落生态学惯用的多元统计学术语。本书也大部分统计术语基于群落生态学，偶尔用纯粹统计学术语来表达一些常用的统计学理论。本书中的术语跟 CANOCO 软件中的术语是一致的。

在群落学分析中，大部分情况是下是基于样方单元（sampling units）观测的原始数据。每个样方内包含很多物种的数量信息，或是其他属性的信息。原始数据常用矩阵来表示，一般是一行代表一个样方，一列代表一个属性特征（如物种，水分或土壤的物理化学特征等等）

生态学原始数据一般由两个部分构成，一组是响应变量 (response variable)，另外一组是解释变量（explanatory variables）。在群落学分析里面，响应变量经常是物种的组成数据，而解释变量通常是环境因子，比如土壤或水的特征属性等。在一个模型里，我们要利用解释变量来预测响应变量（群落的组成）。在排序分析中，解释变量又经常可以分为两组：一类主环境变量，我们主要关心它们与群落内物种分布的关系；另外一组叫

协环境变量 (在一般统计方法里面也叫协同变量 covariates)。协环境变量与主环境变量同时对响应应变起作用，因此我们要分析主环境变量效应之前，应先将协环境变量的效应剔除出来，以便更准确考察主环境变量与物种分布的关系。

举个例子，我们要分析一个特定区域土壤的属性特征和管理模式（刈割或放牧）对草地群落物种组成的影响。当我们感兴趣的土壤的属性的影响，不关心管理模型的影响时，物种组成数据作为响应变量，土壤的属性数据作为解释变量，得出的结论可以看出每个物种分布与土壤梯度的关系。同样，我们考察管理模式对物种分布的影响，不关心土壤的属性的影响时，只用管理模型数据作为解释变量即可。然而，假设管理模式可能改变土壤的属性，这样就能通过影响土壤属性改变来间接影响物种分布。现在我们只要分析管理模式单独对群落内物种组成的影响，需要排除掉通过影响土壤间接影响物种分布的这部分效应。此时，就应该把管理模式作为主环境变量，而土壤属性数据作为协环境变量。

在 CANOCO 程序里面，理解“物种数据”内涵是很关键的。其实只要是需要我们去预测的数据，都可以成为物种数据。例如我们预测集水区不同景观内多种水金属离子的数量的时候，此时一种金属离子在 CANOCO 软件里面即代表一种物种。在群落学里，CANOCO 里面的物种数据经常是用物种的组成来表示，描述物种组成的属性通常用不同的多度指标来表示，例如个体数，频率估计和生物量估计。当然也可以用表示物种在样方内存在与否的二元数据描述物种组成属性。同样，数量型变量和 0-1 型（presence-absence）都可以成为环境变量。对于这部分内容将在下面详细讨论。

1.3 分析类型（Types of analyses）

如果我们要使用数量统计方法描述一个或多个响应变量，如何选择合适的统计模型要依赖于所研究的响应变量是一个还是多个，以及是否有解释变量。

表 1-1 总结了不同变量条件与统计方法的对应关系。如果只有一个响应变量数据，而没预测器（解释变量），我们仅仅需要、也只能归纳这个变量的分布特征（如通过直方图、中值，标准差、四分位极差等）。如果有多个响应变量，依然没有解释变量，我们可以用排序（间接梯度分析）来分析数据，例如可以用主成分分析（PCA）、对应分析（CA）、去趋势对应分析（DCA）和非度量多维尺度分析（NMDS），当然也可以用等级分类，如聚类的方法将样方分为有区别的几类（详见第 7 章的聚类分析）。

如果我们有一个或多个的解释变量，要分析一个响应变量，可以用广义的回归模型，包括传统的回归模型和方差分析、协方差分析。这类分析统称为一般线性模型（general linear model），最近在一般线性模型基础上，发展出了广义线性模型（generalized linear models, GLM）和广义可加模型(generalized additive models, GAM)。有关这回归模型更多的信息，我们将在第 8 章讨论。

如果有多个响应变量需要分析，解释变量一个或多个，我们可以通过直接梯度排序来分析解释变量与多个响应变量（群落学里通常是物种）之间的关系。常用的有冗余分析（RDA）和典范对应分析（CCA）等排序技术。

Table 1-1. *The types of the statistical models*

Response variable(s)...	Predictor(s)	
	Absent	Present
...is one	• distribution summary	• regression models *sensu lato*
...are many	• indirect gradient analysis (PCA, DCA, NMDS)	• direct gradient analysis
	• cluster analysis	• discriminant analysis (CVA)

CVA, canonical variate analysis; DCA, detrended correspondence analysis; NMDS, non-metric multidimensional scaling; PCA, principal components analysis.

1.4 响应变量（Response variables）

响应变量是多元分析固有的成分。如果解释变量（见 1.5 节）并没有被测量，统计的方法可以设定虚拟的解释变量，但响应变量是不可虚拟的。

响应变量（在群落学分析中习惯称物种数据）通常能够用比较准确的量化方法观测或测量获得，比如植物地上部分生物量干重、掉到捕获器里昆虫种类和个体数量、景观尺度内各种植被群落的盖度百分比等等。这样的观测得到的量化数值不仅可以用大于、小于、相等这样的关系来比较，还可以用比例来比较（比如一个值是另外一个值的两倍）。

还有一些情况下，响应变量是半变量化的，比如样方内物种多度分级估计（Braun-Blanquet 等级估计法等）就是典型的例子。最简单响应变量数据是二元数据，比如一个物种是否存在于一个样方中，存在记为 1，不存在记为 0，就构成了二元生态数据。种存在与否的二元数据在数量分析中用的也非常广泛，有些分析方法只适合分析二元数据。

如果响应变量是一组描述不同属性的数据，比如描述水的温度、离子浓度、酸度等等，尽管每种属性都有量化的数据，但是不同属性的量纲是不一致的。这种类型的响应变量在有些排序方法是不适用的，另外一些排序方法中需要经过标准化才能适用（见 4.4 节）。

1.5 解释变量（explanatory variables）

解释变量（统计学也叫预测器 predictors 或自变量 independent variables）就是从样方测量得到用来预测响应变量（不同物种的在样方内多度）的环境因子数据。例如，我们可以用土壤属性和土地管理方式预测群落内物种组成，此时土壤属性和土地管理方式即可称为解释变量。需要注意的是，解释变量不是用来预测它本身，而是使用“预测规则”（通常通过排序图获得）来获得更多关于被研究的有机体或系统信息。

解释变量可以是数量数据（如土壤中硝酸盐离子的浓度），半量化数据（如人类影响程度，比如可分为 0-3 级），或是因子（属性或等级数据）。最简单的解释变量是二元变量（0-1 数据），1 表示某特征或某事件存在或发生，0 表示不存在或不发生。

因子（factors）经常是样方或是事件的分类数据，例如草地管理模式分类、研究污染对河流影响时河流的类型等，或是 0-1 数据。在 CANOCO 程序里面，我们经常叫因子为哑变量（dummy variables），有时也叫指示变量（indicator variables）（或二元变量）。对于每个因子水平，应该有一个分离的哑变量代表。一个因子水平，在一个样方内用哑

变量 1 来表示，对于在这个样方内的同一因子另外的水平，只能哑变量 0 来表示。比如，如果我们要标记一个样方草地植被类型为牧场、草甸或是废弃草地则三种类型中任何一种，此时需要二个哑变量(D1,D2)来代表这个因子三个水平，

D1	D2	
0	0	牧场
0	1	草甸
1	0	废弃草地

另外，可以将因子更精细分解到哑变量，允许我们去产生所谓的模糊编码（fuzzy coding）。比如用我们刚举的例子，如果有一个样点（样方），前几年是作为割草的草地，这几年是作为放牧的草地。如果我们期望考察两种类型的管理模式对于物种组成影响的时候，这时候我们就可以用一个大于 0 而小于 1 的数值来代表所处的因子水平。这里一个很重要的限制是，所有因子水平加起来的数值应该等于 1（类似于正常的因子的虚拟变量）。除非我们能够量化两种管理模式对于该样点的相对重要性，否则我们最好设立三个虚拟变量为 0.5，0.5 和 0。

当我们在解释变量（环境因子数据）与响应变量（物种数据）之间建立预测模型的时候，经常会遇到这样的情况，往往我们仅仅考察解释变量中某几个环境因子的对物种数据的影响，但剩下的环境因子也会对物种产生影响，这些剩余环境因子我们经常称为协变量(Covariables) 。在 CANOCO 中，协变量的影响可以用偏分析（partial analyze）剔除出来。

实际上，任何一个环境因子变量均可以成为协变量。例如，我们要研究管理模式对蝴蝶群落中组成的影响，我们可以在不同的海拔地点取样，海拔也许对群落物种组成影响很大，但此时我们感兴趣的是管理模式的影响，而非海拔梯度的影响。这个时候，如果能剔除出海拔的影响，我们能管理模型与蝴蝶种群之间更清晰的关系。

1.6 如何处理丢失的数据（Handing missing values in data）

无论准备工作多么充分，我们也经常不能够完全收集到我们所需要的数据。比如土壤在从野外到实验室运送过程丢失了一部分，或是我们在记录数据时忘了某一个数据。

大部分情况，我们很难将漏掉数据补回来。所有经常把漏掉数据让它空着，但这样做往往不是明智的。例如，在记录一个只含有零星个体的群落数据时（比如整个群落物种数为 300，但每个样方的平均物种数很少），如果我们会将忘了记录的数据当作该物种在此样方内不存在，比如记录为 0，但实际上忘了记录跟物种在此样方内不存在是两码事，差别很大。有些统计软件会提供处理缺失值的表示方法（ 经常用 NA 表示），但也仅仅是表述上的方便。真正统计学上的方法必须严格处理缺失值，下面几个处理缺失值的方法是值得我们参考的：

1. 如果仅仅在少数几个样方里存在缺失值，可以将这些带有缺失值的样方去掉。当然，切记仅仅是少数的样方带缺失值才可以做。例如，一个带有 30 个变量的 500 个样方矩阵数据，如果有 20 个数据丢失，且这 20 个缺失的数据仅仅发生在 3 个样方里

面。在数据分析之前，将这 3 个样方去掉可能是明智之举。这个策略在很多常用统计软件里面称为“个案删除”（case-wise deletion）。

2. 同样，如果缺失值集中在少数几个非关键性的变量里，我们也可以将这几个变量去掉。当我们做化学分析的时候，这样的情况比较常发生。例如，我们如果知道空气沉淀物镉的浓度，我们经常能够通过镉的浓度合理推断出水银的浓度。这两种物质的浓度具有高度的相关性，因此如果我们只知道一种的浓度，就可以很好预测另一种的浓度。如果我们缺失了镉的浓度的值，只要有水银的浓度，我们完全可以把镉的数据去掉。
3. 上面两种处理缺失值的方法可能是比较粗糙的，因为我们经常花了不少财力和物力来收集数据，最终因为缺失值而把整个样方或变量去掉，显然是很可惜的。实际上有很多内插（imputation）可以来处理缺失值。最简单办法就是用未缺失的平均值来填充缺失值。或许，更复杂一点，可以利用现有的数据先建立一个回归模型，通过其它变量来预测缺失值。用这种方式可以将我们缺失的数据一一填补，而不必将样方或是变量删除。当然，我们不能欺骗自己，我们仅仅能近似补充这些数据，但是数据中的自由度是没法恢复回来的。

值得注意的是，如果我们用估计的数据来替代缺失的数据，在统计检验的时候经常会造成跟自由度水平判断失误，从而让显著性水平估计并不是很准确（经常是过分乐观估计）。我们可以通过降低这些带有补充数据的样方权重以减轻这种影响。这个算法比较简单，例如，有个 20 个变量的数据矩阵，如果一个样方缺失了 5 个变量数据，那么这个样方的权重为 0.75（等于 1.0-5/20），即将样方内所有值乘以 0.75，以减少该方法在统计分析中的权重。当然，这种处理法并不是很完美。如果缺失的变量在统计分析中是很关键变量，我们这么将样方权重整体降低，将会影响关键变量的作用。

有关如何处理缺失数据的方法讨论，可以参考 Little & Rubin (1987).

1.7 从表格内输入数据—WCanoimp 程序（Importing data from spreadsheets – WCanoImp program）

数据输入对于初学者常常是很大障碍。在 CANOCO 的老版本中，对数据格式具有严格的要求，必须是按照它自带的格式才能识别。在 CANOCO4.0 以后的版本，增加一个从普通的数据表格转为 CANOCO 可识别数据的 WCanoimp 程序[在开始菜单里有]。下面将演示如果将 Excel 表格中的数据转化成 CANOCO 可识别的数据。

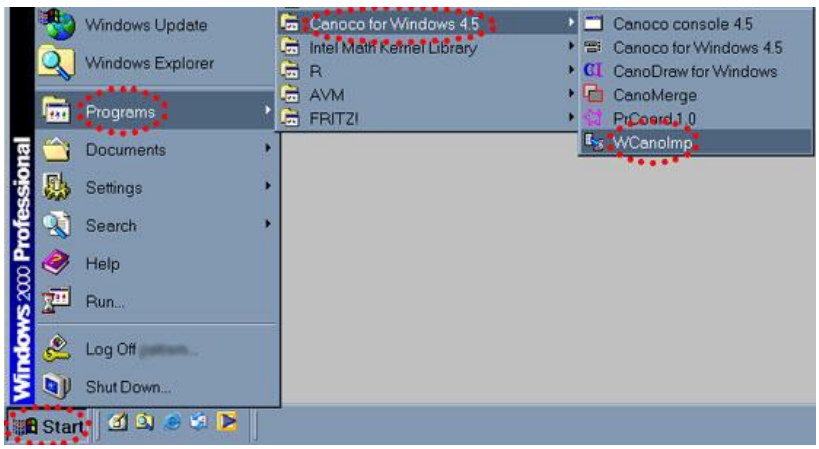

图 1-2a WCanoimp 程序打开途径

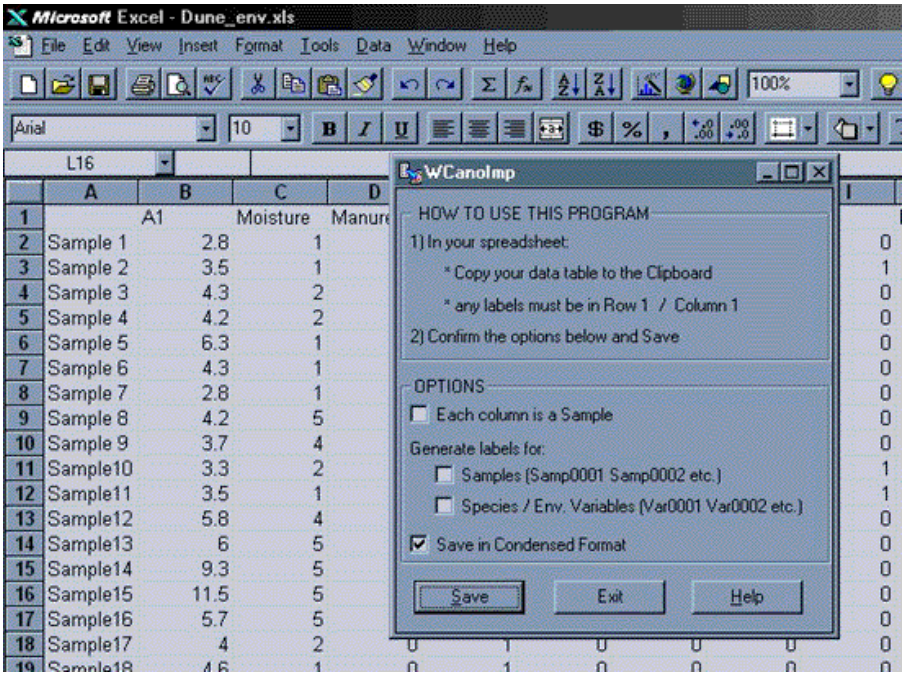

图 1-2 b 数据的参考模式及 WCanoimp 对话框

在 Excel 表格里面，你必须将数据做成矩形形式。默认的方式（也是常用的方式）是一行代表一个样方，一列代表一个变量。表格左顶格最好是空着。最好第一列和第一行分别有样方编号和变量的名称。必须注意的是名称不能超过 8 个字符，如果超过 8 个字符，CANOCO 会自动截取前 8 个字符作为名称。变量名称最好是英文字母、数字 、圆点或是连字符，空格也可以。

除了第一行和第一列，表格内剩下的填充内容必须是数字或是空着，绝对不能使用字符型数据。定性变量（因子）必须转换为哑变量（0-1 数据）方可进入 CANOCO 分析。

当数据在 Excel 表格里按要求整理好后，将包含数据的矩形方阵选定，然后选择“复制”按钮，此时数据便复制到剪贴板中。WCanoImp 便可以从剪贴板中读取数据。如图 1-2a 所示，WCanoImp 可以从“开始”菜单中 Canoco for windows 下来菜单中打开。此时会弹出 WCanoImp 对话框，上半部分包含如何使用该程序的简短信息，下半部分是一些可选框。如果在 Excel 表格数据是按照默认方式组织你的数据，第一选项不必选，相反，如果是数据结构正好相反，以列代表样方，以行代表变量，必须选中这个“Each column is a Sample”选项。除非你的数据是样方很少而变量很多(Excel 表格里面列数不能超过 256 列)，否则不推荐用这种方式组织数据。如果你没有样方或是变量没有编号或是名称，可以选择下面两个选框，程序会帮你给各行各列附上默认名称（Sample1,）。最后一个选项是问你是否存为压缩型数据类型，除非你觉得硬盘空间不够大，否则不必选这个选项，是否选这个选项中对于分析结果毫不影响。

当你确定所以的选择是正确的，你就可以按下 save 按钮，系统弹出新的对话框让你选择保存新文件地方和取个文件名，之后会让你给这个文件加个标注，这个标注内容将显示在新文件的数据内容第一行，以便日后数据内容的识别。选定确认后，程序会告诉你保存成功

1.8 物种数据的转化（Transformation of species data）

排序的过程在于寻找最佳预测响应变量的坐标轴，此时坐标轴代表回归预测器（解释变量）（这些内容将第 3 章详细解释）。因此，在排序中对响应变量进行转化，就好比在多重回归中要将很多物种数据转化一个单因变量的形式一样。有点限制的是，在排序中，所有的响应变量应该是做一致的转化，因为响应变量经常是同一属性的数据，具有一致的量纲。在基于单峰模型（加权平均）的排序（见 3.2 节），所有响应变量的数据不能是负值，这就要求某些带负值的数据必须转化，而且对于转化的结 果有更严格的要求（不能为负值）。

这个限定（非负值）对于对数转化更应该值得注意。因为 1 的对数为 0，而处于 0-1 之间的值取对数是负值。因此，在 CANOCO 里面提供了变通的对数转化公式：

$$y' = \log(A \cdot y + C)$$

在对 y 转化之前，你可设定上面公式中 A 和 C 的值，让输出的 y`值保证不小于 0。在系统中，A 和 C 默认值均为 1，这样可以保证本来是 0 的值，转化后仍为 0，而其他的值依然是正的。然而，如果你的原始值很小（比如说处于 0-0.1 之间），可以将 A 的值适当增大，比如说设为 10。但对于百分比数据和普通的点数据，默认的转化（log(y+1)）是比较合适的。

什么情况下需要对响应变量进行对数转化是个很难回答的问题，统计学家的答案也是五花八门。我们建议你不必太在意关于数据的分布特征，比如原始数据不一定符合理想的正态分布，对于排序来说，也不一定非要通过对数转化为正态分布的类型。是否需要对数转化，关键还是比较原始数据和转化数据分析处理的最终结果哪个更好解释你所要探讨的问题。

正如上面所描述那样，排序可以被看作多重回归的扩展，所以整个排序方法可以用简单回归的语言来描述。你可以通过一个或多个预测器（环境因子或排序轴）来预测一个响应变量（比如物种的多度）。比如，在一元线形回归方程中（$y=B_0+BX+E$），你可以问当 x 变化一个单位时，y 的平均值是如何变化的？如果自变量和因变量都没有对数转化，你可以回答这个问题：当 x 增加一个单位时候，y 的增量是 B。但在很多情况下，你可能更倾向听到这样的解释，如果变量 X 增加一个单位，Y 的量将增加 10%，或是，y 增加 1.1 倍这样的话。显然，这已经并不是线形回归模型所能体现出来的，因此，这种前情况下，你需要对响应变量进行对数转化。

同样，如果预测器（环境因子）变化是成倍增长，此时的环境变量也应该被对数转化。

植物群落组成数据有时是半量化估计尺度数据，比如最典型的例子是多度的 Braun-Blanquet 等级估计（7 个等级水平，分别为 r,+,1,2,3,4,5 这 7 个标号表示）。这个等级估计经常在数据表格里用 1-7 的数据来代替原来的标号进行分析。其实，这个量化 1-7 的数字已经相当原始多度数据的对数转化，因为不同等级的多度变化往往是成倍增加的，不是简单的单位量的变化。

在 CANOCO 里面另外一种有用的数据转化模式是平方根转化。平方根转化更适合观测计数数据（count data），比如在土壤收集器中收集到标本个体的数量，或是通过某一条标志线蚂蚁的数量等等这样的观测数据。但对数转化对这样的数据进行转化也是可以的。

当然，如果你觉得需要某种除了对数转化和平方根之外的数据转化，你可以在数据

输入 CANOCO 之前通过别的数据软件进行。

1.1 解释变量的转化（Transformation of explanatory variables）

因为解释变量（环境因子变量，包括协变量）经常是样方的多属性数据，量纲往往不一样的，所以你经常需要选择合适的转化方法分别对环境变量进行单独转化。CANOCO 里面并没有提供这样的转化，因为很多环境因子在被输入 CANOCO 之前，就应该被转化好。

但你应该知道，CANOCO 读了了环境变量或协变量后，它们会自动被中心化和标准化，让它们的均值为 0 和方差为 1（这个转化通常被称为“单位方差标准化”）。