返回总目录

目　　录

第八部分
鉴 别 分 析

第 38 章　SAS 系统内的三种执行鉴别分析程序概述

38.1　三种执行鉴别分析程序

鉴别分析是一套用来分类的统计方法。它包括好几种鉴别法，如：分类鉴别分析，典型鉴别分析及逐步鉴别分析等。一般而言，鉴别法的用途可分三种：

1. 找出一个鉴别函数，这个函数的值可用来猜测到底某一个观察体 (如：牵牛花) 是属于甲类 (如：草本类) 还是属于乙类的 (如：木本类)。
2. 找出一组数值变量的线性组合，此线性组合可用来强调各类别之间的不同。
3. 从一组数值变量中，挑选出一部分的变量，其线性组合可以最有效地显出各类别之间的不同。

SAS 里有三种统计程序可用来执行鉴别分析它们的简介如下：

■PROC DISCRIM 程序

执行分类鉴别分析。将观察体分到某一类别里 (即上述第一种功能)。

在新版下执行这个程序，读者可选择参数鉴别法或无参数鉴别法来分析数据。若你选用参数鉴别法，则 DISCRIM 程序假设每一类别里的观察体来自多元常态分配。若你选用无参数的鉴别法，则 DISCRIM 的分析方法会与旧版的程序 NEIGHBOR 完全相同。有关这种无参数的分析法，请参阅第 39 章或第 39 章第 39.4 节例四的说明。

■PROC CANDISC 程序

执行典型鉴别分析，亦即上述第二种功能。它的原理与主成份分析及典型相关分析有关。

■PROC STEPDISC 程序

执行逐步鉴别分析，以便达成上述第三种功能。它的原理是逐步回归分析。此程序分析的结果可再输入 DISCRIM 或 CANDISC 程序，以便得到更详尽的分析。

38.2　鉴别法与集群法的异同

与鉴别法相似的另一类统计分析法称为集群法 (见第九部分)。鉴别法与集群法的目的都是为了分类。但鉴别法必须从已知的，事先确定的类别 (如：草本类和木本类两个类别) 选出具有代表性的样本 (如：牵牛花、剑兰代表草本类植物，而玫瑰代表木本类植物)，然后由这些植物的属性中找出一套最有效的鉴别函数，这个 (些) 函数可用来执行分类的工作。而集群法假设我们事先并不知道到底有那些类别，而以各观察体的属性为依据，建立集群并分派个体到合适的集群里。所以它们分类的原理是不一样的。

第 39 章　分类鉴别分析：统计程序 PROC DISCRIM

39.1　PROC DISCRIM 程序概述

PROC DISCRIM 程序的主要用途在于将观察体分类。分类的根据是观察体在一个或一个以上连续变量上的值；分类的结果则是两个或两个以上的类别。

若读者假设各类别内的观察体来自一个多元常态分配，则 PROC DISCRIM 利用参数统计法导出一个线性的或二项式判别函数作分类之用。

若多元常态分配的假设不成立，则 PROC DISCRIM 利用无参数统计法导出分类用的判别函数。

下面简单地介绍本章内提及的几个专有名词：

■ 判别指标（Discriminant Criterion）

判别指标是 PROC DISCRIM 根据一组连续变量，外加一个分类变量的值所导出的分类标准。这个分类标准可应用在第二组数据上，以便测试分类的准确性。

■ 参考数据集（Calibration 或 Training Data Set）

用来导出上述判别指标的第一组数据称为参考数据集。

■ 判别函数（Discimrinant Function）或分类指标（Classification Criterion）

若各类别数据符合多元常态的分配假设，则 PROC DIRCRIM 利用参数统计法导出一个判别函数。此函数的值就是分类的依据。

这个判别函数的决定可根据三种不同的数据结构：

甲、各类别里的共变异数矩阵，由此而导出一个二项式判别函数。

乙、各类别里共变异数矩阵的平均 (Pooled Covariance Matrix)。由此导出一个线性的判别函数。

丙、分类前各类别可能出现的概率 (Prior Probability)。

由此可知，判别函数是由已知的类别导出。分类之后，其结果与已知的类别作比较，可得知错误分类的比率。这个错分的比率是定义判别函数据确度的根据。有关判别函数的理论部分，请参考饶氏的著作 (Rao，1973)。

■ 颗粒法（Kernel Method）

若各类别内的数据无法符合多元常态的分配假设，则 PROC DISCRIM 利用无参数统计法 (如颗粒法) 来估计观察体隶属于各类别的事后概率密度。

颗粒法使用均等 (Uniform)，常态 (Normal)，伊氏 (Epanechnikov)、双加权 (Biweight)，参加权 (Triweight) 的颗粒选择法来计算事后概率密度。数据来源是各类别内

的共变异数矩阵或矩阵的平均。

■ K 个最近邻法 (K-Nearest-Neighbor)

亦即旧版所介绍的近邻鉴别分析法 (PROC NEIGHBOR)。这是一种无参数的统计方法；分类的根据是观察体与各类别间的玛氏距离 (Mahalanobis Distance)；数据的来源则是各类别内共变异数矩阵的平均。

***** ***** *****

除了上述介绍的几种鉴别分析法之外，PROC DISCRIM 也可用来执行典型鉴别分析，并且在分析完毕之后导出判别指标。若读者只想得到单纯的典型鉴别之结果，则应执行第 40 章介绍的 PROC CANDISC 程序。

PROC DISCRIM 所产生的输出资料文件内含多种统计值，视不同的分析方法而定。在下面 39.3 节“如何撰写 PROC DISCRIM 程序?”指令 #1 的选项串部分，我们会对这些输出资料文件的内容作进一步的说明。

鉴别分类的准确度以错分观察体的比率来表示。错分的比率或以实际错分的频率或事后概率值的错估值来表示。

39.2 计 算 公 式

首先介绍几个统一的数学代号：

x　一个观察体在 p 个连续变量上的值 (也是一个含 p 个元素的向量)。

s　共变异数矩阵的平均。

t　识别各组的下标。

n_t　第 t 组内参考数据集的个数。

m_t　第 t 组在 p 个连续变量上的平均数 (也是一个含 p 个元素的向量)

s_t　第 t 组之观察体在 p 个连续变量上的共变异数矩阵。

$|s_t|$　S_t 矩阵的行列式值。

q_t　观察体隶属于第 t 组的事前概率。

$p^{(t|x)}$ 观察体 x 隶属于第 t 组的事后概率。

■贝氏定理 (Baye's Theorem)

DISCRIM 利用已知的事前概率 (亦即各组在 x 点上的概率密度)，以及贝氏定理来计算事后概率 p(tx) 如下：

$$p(t|x)=q_t f_t(x)/f(x)$$

在此，$f_t(x)$=在 x 点上，从 t 组所估计出来的概率密度。

$f(x)=\sum_t q_t f_t(x)$=在 x 点上的无条件的概率密度。

DISCRIM 分析的原理是：首先从 p 向量空间中分割出一个区域，以 Rt 表示。 Rt 区内所含的向量，其对应的观察体隶属于第 t 组的事后概率应该是最大的。换言之，若一个观察体所对应的 p 元素向量座落在 Rt 区域内，则该观察体被认为属于第 t 组。

■参数统计法

这种方法的统计假设，如前所述，是多元常态分配。所导出的判别函数 (或分类指标) 以通用的平方距离表示。因此，若观察体与某 t 组的通用平方距离是最小的，则此观察体就隶属于该组。

平方距离的计算公式为：

$$d_t^2(x)=(x-m_t)'V_t^{-1}(x-m_t)$$

在此，V_t 可以等于 S_t (若读者决定使用组内的变异数矩阵)，或

V_t 等于 S (若读者决定使用组内变异数矩阵的平均)。

接下来，DISCRIM 程序计算判别函数在观察体 x 点上的概率密度 [以 $f_t(x)$ 表示]：

$$f_t(x)=(2\pi)^{-p/2}|V_t|^{-1/2}\exp(-0.5d_t^2(x))$$

利用前述的贝叶斯定理，观察体隶属于第 t 组的事后概率为：

$$p(t|x)=\frac{q_t f_t(x)}{\sum_u q_u f_u(x)}$$

所以，观察体与第 t 组之间的平方距离等于

$$D_t^2(x)=d_t^2(x)+g_1(t)+g_2(t)$$

在此，$\begin{cases} g_1(t)=\log_e|St| & \text{，若读者决定使用组内的共变异数矩阵；或} \\ g_1(t)=0 & \text{，若读者决定使用组内共变异数矩阵的平均。} \end{cases}$

$\begin{cases} g_2(t)=-2\log_e(q_t) & \text{，若各组的事前概率不等；或} \\ g_2(t)=0 & \text{，若各组的事前概率均等。} \end{cases}$

利用前述平方距离的定义，事后概率也可以表示成

$$p(t|x)=\frac{\exp(-0.5D_t^2(x))}{\sum_u \exp(-0.5D_u^2(x)))}$$

因此，某个观察体属于第 t 组的先决条件是分派至该组的事后概率是最大的，或该观察体距 t 组的平方距离是最小的。若某一观察体的事后概率未达预设的最低标准，则该观察体被纳入其它 (OTHER) 的组别。

■无参数统计法

无参数的统计法首先在估计各组在观察体 x 点上的概率密度。有两种估计法：颗粒法 (KernelMethod) 以及 K 个最近邻法 (K-Nearest-Neighbor Method)。

颗粒法使用均等 (Uniform)、常态 (Normal)、伊氏 (Epanechnikov)、双加权 (Biweight)、参加权 (Triweight) 等颗粒进行概率密度的估计。

观察体与各组间的距离可以界定成欧氏距离 (Euclidean Distance) 或玛氏距离 (Mahalanobis Distance)。欧氏距离 $d^2(x_1,x_2)=(x_1-x_2)'(x_1-x_2)$，玛氏距离 $d^2(x_1,x_2)=(x_1-x_2)'T^{-1}(x_1-x_2)$，T 是整个资料文件的共变异数矩阵。

DISCRIM 程序在颗粒法下计算玛氏距离时，允许读者挑选 (1) 组内共变异数矩阵，或 (2) 共变异数矩阵的平均作为距离的单位。

然而，DISCRIM 程序在 K 个最近邻法下计算玛氏距离时，只允许读者选择其变异

数矩阵的平均作为距离的单位。

无参数的鉴别法一般而言遵循下列的几个步骤：

第一 计算观察体 x，y 在第 t 组内的距离：

$$d_t^2(x, y)=(x-y)'V_t^{-1}(x-y)$$

此外，V_t 的定义可以有下列几种：

$V_t=S$，共变异数矩阵的平均。

$V_t=\text{diag}(S)$，共变异数平均的矩阵对角元素

$V_t=S_t$，第 t 组内的共变异数矩阵

$V_t=\text{diag}(S_t)$，第 t 组内共变异数矩阵的对角元素

$V_t=I$，X 的单位矩阵

第二 根据参考数据集的资料，估计概率密度。

第三 根据上述估计的概率密度，计算观察体属于各组的事后概率。

第四 针对各观察体，比较各组的事后概率。最大的概率值就决定该观察体属于该组。若最大的事后概率值仍未达到最低标准或最大概率值有两个时，该观察体会被分派到“其它(OTHER)”的组内。

颗粒法利用一个定值的半径 (r) 以及读者选定的的颗粒函数 (K_t) 来估计在观察体 x 点上的概率密度。因此，对任何一组 t 而言，以 r 为半径的椭圆体的体积是

$V_r(t)=r^p|V_t|^{1/2}V_0$，p= 向量空间的维度，

$$V_0=\frac{\pi^{p/2}}{\Gamma(p/2+1)},$$ ，单位球体的体积 (Γ 是伽玛函数)

V_t=定义如上述第一步骤所示

颗粒函数的设定则可有下列数据 (z 是一个含 p 元素的向量)：

均等颗粒 (Uniform Kernel)

$$K_t(z)=\frac{1}{V_r(t)}, \quad \text{若 } z'V_t^{-1}z\leqq r^2;$$

$$=0 \quad \text{，其它条件。}$$

常态颗粒 (Normal Kernel)

$$K_t(z)=\frac{1}{C_0(t)}\exp(-0.5z'V_t^{-1}z/r^2)$$

在此，$C_0(t)=(2\pi)^{p/2}r^p|V_t|^{1/2}$，此常态颗粒的平均数为零，变异数为 r^2V_t。

伊氏颗粒 (Epanechnikov Kernel)

$$K_t(z)=C_1(t)(1-z'V_{t-1}z/r^2) \quad \text{，若 } z'V_t^{-1}z\leqq r^2$$

$$=0 \quad \text{，其它条件}$$

在此，$C_1(t)=(1+p/2)/V_r(t)$。

双加权颗粒 (Biweight Kernel)

$$K_t(z)=C_2(t)(1-z'V_t^{-1}z/r^2)^2 \quad \text{，若 } z'V_t^{-1}z\leqq r^2$$

$$=0 \quad \text{，其它条件}$$

在此，$C_2(t)=(1+p/4)C_1(t)$。

参加权颗粒（Triweight Kernel）

$K_t(z)=C_3(t)(1-z'V_t^{-1}z/r^2)^3$　　，若 $z'V_t^{-1}z \leqq r^2$

$=0$　　，其它条件

在此，$C_3(t)=(1+p/6)C_2(t)$。

*****　　*****　　*****

有了上述定义的颗粒函数，DISCRIM 可进一步估计在 x 点上的概率密度为：

$$f_t(x)=\frac{1}{n_t}\sum_y K_t(x-y)$$

此处，K_t 是前述的颗粒函数的一种，y 代表所有属于 t 组的观察体，n_t 是 t 组内观察体的个数。

观察体 x 被分到 t 组的事后概率为

$$p(t|x)=\frac{q_t f_t(x)}{f(x)}$$

此处，f(x) 是边际概率密度。

若某个观察体对应出来的 f(x) 函数值等于零，则该观察体会被列入其它 (OTHER) 的组别。

■错误分类的比率是如何定义的？

由于鉴别分析的方法形形色色，DISCRIM 程序决定采用两个标准来鉴定这些方法的优劣、亦即错误分类的比率。

第一种标准是计算错误分类的次数，第二种标准是计算事后概率的误差。

第一种标准的执行可藉两组独立的数据，或同一组数据。若有两组独立的数据 (或样本)，则第一组数据是参考数据集，由此导出分类的指标 (Classification Criterion) 。然后将这个分类指标应用在第二组数据上，以便试验出到底有多少个观察体被分到错的类别里。

这种方法唯一的缺点是：当第二组数据太小时，分错的可能性会增加。

当数据只有一组，并且含原始资料时，DISCRIM 程序可同时导出分类指标，并且评鉴它的优劣。不过这种方法的缺点是评鉴的结果不够客观，容易造成较优的印象。补救的方法有二：(甲)将数据等分为二，一半当作参考数据集，另一半当作评系数据。不过，这种补救法会造成小样本的误差 (如上所述)。第二种补救法是 (乙) 执行交叉确认 (Crossvalidation，见 Lachenbruch 与 Mickey，1968 年参考文献)。交叉确认的原理是将数据中前（n-1）个观察体当作参考数据集。由这（n-1）个观察体所导出的分类指标再应用在第 n 个观察体上。此法执行 n 次，每次以一个不同的观察体当作评系数据，如此，就可比较各种鉴别法的优劣。

第二种标准是在实际执行过程里计算事后概率的误差。当这种标准与参数统计法 (线性或非线性的判别函数) 联用时，读者必须确定多元常态分配的假设是成立的。

若交叉确认法配合著事后概率的计算同时进行，则评鉴的结果会更准确 (参考 Hora

与 Wilcox1982 年论文)。

39.3　如何撰写 PROC DISCRIM 程序

PROC DISCRIM 程序含十二道指令，它们的格式如下：

```
PROC DISCRIM      选项串;
    CLASS         变量名称;
    VAR           变量名称串;
    PRIORS        概率串;
    FREQ          变量名称;
    WEIGHT        变量名称;
    ID            变量名称;
    TESTCLASS     变量名称;
    TESTFREQ      变量名称;
    TESTID        变量名称;
    BY            变量名称串;
```

指令 #1　PROC DISCRIM 选项串;

此指令的选项可分下列七大类讨论：第一类的选项与输入资料文件的界定有关，第二类选项与输出资料文件的界定有关，第三类选项用来控制报表的印出，第四类选项界定有关判别函数的各项事宜，第五类选项与参数的鉴别分析有关，第六类选项与典型鉴别分析有关，第七项选项与分析结果的评鉴有关。

第一类选项　下列两个选项可用来界定输入资料文件：

(1) DATA= 第一个输入资料文件名称

为第一个输入资料文件命名。DISCRIM 程序接受 TYPE=CORR，COV，CSSCP，SSCP，LINEAR，QUAD，MIXED，以及 DISCAL 等资料文件。当选用 METHOD=NPAR 时，输入资料必须是原始数据。若省略此选项，则 SAS 会自动找出在此程序之前最后形成的 SAS 资料文件，对它执行分类鉴别分析。

(2) TESTDATA= 第二个输入资料文件名称

为第二个输入资料文件命名。这个资料文件内的变量名称必须与第一个输入资料文件的变量名称完全相同。这个资料文件的数据是用来交叉确认判别函数的有效度。TESTDATA= 可与 TESTLIST 或 TESTLISTERR 选项 (见下述) 或 TESTCLASS，TESTID，TESTFREQ 等指令合用。

第二类选项　下列六个选项可用来界定各式的输出资料文件：

(1) OUTSTAT=资料文件文件名

此资料文件必含变量的平均数、标准差，以及相关系数矩阵。若读者同时选用 CANONICAL 的选项，则此输出资料文件同时含典型相关、典型结构、典型系数，以及典型变量在各组内的平均数值。若鉴别分析的方法界定为

METHOD=NORMAL，则此资料文件同时含判别函数的系数；而且输出资料的 TYPE 会依另一个选项 POOL= 而定。当 POOL=YES 时，TYPE=LINEAR；当 POOL=NO 时，TYPE=QUAD；当 POOL=TEST 时，TYPE=MIXED。

如果分析的方法界定为 METHOD=NPAR，则资料文件的 TYPE 必定是相关系数 (即 TYPE=CORR)。

最后，选项 OUTSTAT= 可与原始数据、相关系数矩阵、变异数 / 共变异数矩阵、平方和与内乘积矩阵等类形的输入资料文件联用。

(2) OUT= 资料文件文件名

此输出资料文件只能与原始数据联用。它含所有输入资料文件的原始数据，分类后概率以及观察体经替代法分类后所属的组别。若读者同时选用 CANONICAL 选项，则此输出资料文件同时含各观察体的典型变量值。

(3) OUTCROSS=资料文件文件名

此输出资料文件只能与原始数据联用。它含所有输入资料文件的原始数据、分类后概率以及观察体经双重检验法分类后所属的组别。若读者同时选用 CANONICAL 选项，则此输出资料文件同时含各观察体的典型变量值。

(4) OUTD=资料文件文件名

此输出资料文件只能与原始数据联用。它含所有输入资料文件的原始数据，以及观察体属于各分组的概率密度。

[下面两个选项与前述 TESTDATA= 选项联用，旨在收集鉴别分析的结果以及分组后的概率密度。]

(5) TESTOUT=资料文件文件名

这个输出资料文件含所有 TESTDATA= 界定的原始数据、分类后概率以及各观察体经分组后所属的组别，若读者同时选用 CANONICAL 选项，则此输出资料文件同含各观察体的典型变量值。

(6) TESTOUTD=资料文件文件名

这个输出资料文件含有 TESTDATA= 的原始数据以及观察体属于各分组的概率密度。

第三类选项　下列二十个选项可用来控制报表的打印：

(1) ALL

打印出下述 (2)～(7)，(9)～(20) 个选项的结果。

(2) SIMPLE

印出每一变量在各组内或整个样本内的描述性统计值。

(3) STDMEAN

印出每一变量在各组内或整个样本内的标准化平均数。

(4) ANOVA

列出每一变量在各组内平均数检定的变异数分析之结果。

(5) MANOVA

列出所有变量在各组内平均数检定的多变量变异数分析之结果。

(6) DISTANCE

打印出各组间的平方距离；这个平方距离的大小依选项 POOL 及 METRIC 而定。

(7) SHORT

要求将精简的分析结果印出即可。若分析的方法界定为 METHOD=NORMAL，则这个选项会导致 DISCRIM 程序省略平方距离，行列式值，以及判别函数的系数。若分析方法为 CANONICAL (典型鉴别分析)，则 DISCRIM 程序略去典型结构，典型系数以及典型变量各组内的平均数，唯有典型相关系数印在报表上。

(8) NOPRINT

不印出任何分析的结果。

(9) BCORR

要求印出各组间的相关系数矩阵。

(10) WCORR

要求印出各组内的相关系数矩阵。

(11) PCORR

要求印出各组内相关系数矩阵之平均。

(12) TCORR

要求印出以整个样本为单位的相关系数矩阵。

(13) BCOV

要求印出各组间的共变异数矩阵。

(14) WCOV

要求印出各组内的共变异数矩阵。

(15) PCOV

要求印出各组内共变异数矩阵的平均。

(16) TCOV

要求印出以整个样本为单位的共变异数矩阵。

(17) BSSCP

要求印出各组间的向量内乘积矩阵。

(18) WSSCP

要求印出各组内的向量内乘积矩阵。

(19) PSSCP

要求印出各组内向量内乘积矩阵之平均。

(20) TSSCP

要求印出以整个样本为单位的向量内乘积矩阵。

第四类选项　下列三个选项可用来界定有关判别函数的各项事宜：

(1) METHOD=NORMAL 或

METHOD=NPAR

这个选项是用来决定一个导出判别指标 (Classification Criterion) 的方法。当读者界定 METHOD=NORMAL 时，导出 (一个线性或二项式) 判别函数的方法会是

根据参数统计程序。这个统计程序假设各组内观察体在母群的分配上是一个多元常态分配。

若读者界定 METHOD=NPAR，则判别函数的导出是根据无参数的统计程序，这种统计程序对观察体在母群内的分配不作任何假设，必须与 K= 或 R= 选项联用。此选项的内设值是 METHOD=NORMAL。

(2) POOL=YES　或
POOL=NO　或
POOL=TEST

内设值是 POOL=YES，若读者省略此选项或选用 POOL=YES，则分析用的判别函数来自各类别的共变异数矩阵的平均。若读者选用 POOL=NO，则判别函数来自各个类别里的共变异数矩阵。若读者选用 POOL=TEST，则 PROC DISCRIM 利用可能比测试 (Likelihood Ratio Test) 来检验各类别里的共变异数矩阵是否相同。此检定的显著度在下述选项 SLPOOL=中界定。若检定结果达显著水准，则判别函数将来自各类别里的共变异数矩阵 (相当于 POOL=NO)。若检定结果不显著，则判别函数将来自于各类别内共变异数矩阵的平均 (相当于 POOL=YES)。

(3) SLPOOL=显著度

与上述 POOL=TEST 选项合用，旨在界定可能比测试的显著水准，内设值是 .10。

第五类选项　与无参数的鉴别分析有关，有下列四个选项：

(1) K= 正整数 (如 3)

界定 k 个最近邻法分析中的 k 值。不可与下一个选项 R= 联用。

(2) R=正实数

界定各颗粒法分析中椭圆体的半径值。不可与上一个选项 K=联用。

(3) METRIC=DIAGONAL　或
METRIC=FULL　或
METRIC=IDENITITY

界定平方距离的计算单位。若读者界定 METRIC=FULL，则距离的单位可以是组内共变异数矩阵的平均 (若 POOL=YES) 或各个组内的共变异数矩阵 (若 POOL=NO)。同理，METRIC=DIAGONAL 的单位也视选项 POOL= 的值而定：若 POOL=YES，则 (组内) 共变异数平均矩阵的对角线成为距离的计算单位；若 POOL=NO，则各个 (组内) 共变异数矩阵的对角线成为距离的单位。METRIC=IDENTITY 会导致欧氏距离的单位。当 METHOD=NORMAL 时，METRIC=FULL 是内设值。

(4) KERNEL=BIWEIGHT(BIW)　或
KERNEL=EPANECHNIKOV(EPA)　或
KERNEL=NORMAL(NOR)　或
KERNEL=TRIWEIGHT(TRI)　或
KERNEL=UNIFORM (UNI)　(内设值)

这个选项界定颗粒函数的形式。这五种选项的值及其定义在第 39.2 节已有介绍，故不另赘述。值得注意的是，此选项必须与 R= 的选项一起界定。

第六类选项　与典型鉴别分析有关，有下列三个选项：

(1) CANONICAL (或 CAN)

要求 DISCRIM 程序执行典型鉴别分析。

(2) NCAN=正整数 (如 3)

界定典型鉴别分析中典型变量的个数。因此这个选项的值 (如 3) 必须小于或等于变量的总个数。这个选项的内设值等于 (v,c-1) 之中较小的值,在此 v=VAR 指令界定变量总数，c=CLASS 指令所界定的分类变量的总数。因此，若读者在 DISCRIM 程序中要求一个或一个以上的输出资料文件 (由 OUT=，OUTCROSS=，TESTOUT= 等选项界定)，则典型变量的最后 [v-(c-1)] 个无法导出，其数值会等于遗漏数据。若 NCAN=0，则 DISCRIM 程序只打印典型相关系数的值，而无典型系数、结构或其它相关的统计值。

值得读者注意的是：当你选用 NCAN= 或 CANPREFIX= 之后，DISCRIM 自动执行典型鉴别分析。然而，若你希望省略判别指标的计算，则应执行第 39 章所介绍的 CANDISC 程序。

(3) CANPREFIX= 典型变量的名称 (如 XYZ)

若典型变量的名字界定成 XYZ，则第一变量自动命名为 XYZ1，第二变量则是 XYZ2，以此类推。这个选项的内设值是 CAN，与数字 1，2 合起来，其长度不可超过八个字母。

第七类选项　与分析结果的评鉴 (如交叉确认的结果，事后概率的误差等) 有关，含十一个选项，介绍如下：

(1) THRESHOLD=概率值

这个选项界定各观察体隶属于某组别的事后概率之最低标准。因此，若某个观察体的事后概率低于此标准，则它会被分派到其它 (OTHER)。这个选项的内设值等于零。

(2) LIST

印出各观察体被分类后的结果；此选项只可与原始数据联用。

(3) LISTERR

印出被分到错误类别的观察体；此选项只可与原始数据联用。

(4) TESTLIST

与选项 TESTDATA= 合用，列出第二个输入资料文件内所有的观察体分类后的结果。

(5) TESTLISTERR

与选项 TESTDATA= 及指令 TESTCLASS 合用，列出第二个输入资料文件内被分到错误类别里的观察体。

(6) CROSSLIST

印出交叉确认后各观察体分类的结果。

(7) CROSSLISTERR

印出交叉确认后被分派到错误类别里的观察体。

(8) CROSSVALIDATE

要求 DISCRIM 程序对 DATA= 资料文件内的观察体进行交叉确认。交叉确认的过程是这样的：若读者界定一种参数分析的方法，则 DISCRIM 程序自 (n-1)个观察体中导出一个判别函数。依据这个函数再对第 n 个观察体进行分类。如此重复 n 次，看交叉确认的结果如何。若读者界定一种无参数的分析方法，则概率函数的计算 (与选项 R= 联用)或 k 个最近邻的定义 (与选项 K=联用)，将不含第 n 个观察体。如此重复 n 次，以便审核分析结果的优劣。

若读者已经在程序中界定 CROSSLIST，CROSSLISTERR，或 OUTCROSS= 等选项，则 CROSSVALIDATE 也会自动被界定。

(9) POSTERR

要求打印出分类指标所造成的错误分类之事后概率。

(10) SINGULAR=　小的正实数

此选项定义分析中矩阵之奇异性。内设值等于 10 的 -8 次方。

(11) NOCLASSIFY

抑制分类的结果，不使其印出，必须与含原始数据的输入资料文件联用。

指令 #2　CLASS 变量名称;

读者一定要在 PROC DISCRIM 中列出此指令，不可省略。此指令指定一个分类变量，这个分类变量可以是数值的 (如：性别，男=1，女=2) ，也可以是文字的 (如男=M，女=F)。

指令 #3　VAR 变量名称串;

列举所有鉴别分析可能用到的变量名称。若省略此指令，则在本程序中其它指令未曾提到的所有数值变量将被纳入分类鉴别分析之内。

指令 #4　PRIORS 概率串;

这些概率串表示在分类鉴别分析前，各类别可能出现的概率。若你觉得各类别可能出现的概率相等，则你可以省略此指令。若你觉得各类别出现的概率与其人数多少成正比，则你可以订 PRIORS PROPORTIONAL (或 PRIORS PROP;)。如果你想订其它的概率值，则可在此指令中列出各类别的概率。比方说：一班学生中，只有 10% 的人得 A，30% 得 B，50% 的人得 C，10% 的人得 D，则下式可表示这样的成绩分配：

```
PRIORS A=.1 B=.3 C=.5 D=.1;
```

这样的表示法也可以用在数值变量上 (假设 男=1，女=2)，如：

```
PRIORS 1=.6  2=.4;
```

须注意的是，所有的概率值总和必须是 1 (即 100%)。

指令 #5　FREQ 变量名称;

此变量的值代表资料文件中各观察体重复出现的次数。当这个值小于 1 或是一个遗

漏数据时，其对应的观察体便被排除在分析之外。若 FREQ 变量的值不是一个整数 (如 5.8)，则 SAS 自动取其整数的部分 (如 5)。

指令 #6　WEIGHT 变量名称;

这个指令的作用是将原始数据中各观察体作不等的加权调整。调整的幅度视 WEIGHT 变量的值而定。若 WEIGHT 变量的值等于负值或遗漏数据，则 SAS 自动将其改为零。

这个指令与上述 FREQ 指令有异曲同工之妙。唯一不同的是 WEIGHT 指令不改变分析中自由度的计算，然而 FREQ 指令会。

指令 #7　ID 变量名称;

此指令必须与 PROC DISCRIM 中的 LIST 或 LISTERR 选项合用。若选用此指令，则此变量的值会出现在报表中的各观察体之前。若省略此指令，则报表以观察体的编号 (以 _N_ 表示) 来区分。

指令 #8　TESTCLASS 变量名称;

这个指令从 TESTDATA= 资料文件中挑选一变量，这个变量是用来检定资料文件中的观察体是否有被分错的现象。这个变量须与上述 CLASS 指令的变量具有相同的形式 (如两者都是数值变量) 及相同的长度 (如两者都是含十个字母的文字变量) 。当此变量的值与资料文件中观察体分类后的类别不符合时，PROC DISCRIM 认为这一个观察体的分类是错误的。

指令 #9　TESTFREQ 变量名称;

这个指令界定 TESTDATA= 资料文件加权变量的名称 。其值代表各观察体在 TESTDATA 中重复出现的次数。若 TESTFREQ 变量的值小于 1 或是遗漏数据，则此观察体被排除于分析之外。若变量的值含小数，则 SAS 自动取其整数的部分。

指令 #10　TESTID 变量名称;

须与 TESTLIST 或 TESTLISTERR 选项合用，而且这个变量必须来自 TESTDATA= 资料文件。在报表上，分类后的结果是以这个变量的值 (而非观察体的编号) 来表示的。

指令 #11　BY 变量名称串;

SAS 依据此指令所列举的变量将资料文件分成几个小的资料文件，然后对每一个小的资料文件分别执行分类鉴别分析。当读者选用此指令时，资料文件内的数据必须先按照 BY 变量串的值做由小到大的重新排列，这个步骤可藉 PROC SORT 达成。

39.4　范　　例

例一：费氏紫罗兰的分类鉴别分析

本资料文件 (IRIS) 的数据来自费契尔氏 (Fisher，1936)。它常被用来当做范例以解

释分类的方法。这一组资料是从三种不同属性的紫罗兰 (SETOSA=1，VERSICOLOR=2，及 VIRGINICA=3) 搜集而来的。每种紫罗兰各取五十个样本，然后测量它们花萼与花瓣的长与宽 (测量单位是厘米)。每一个观察体包括五个数据，依序是：花萼长，宽；花瓣长，宽；及属性号码。下列的程序分两部分来处理资料：第一部分是 PROC CHART，第二部分是 PROC DISCRIM。PROC CHART 的目的在于找出这三种紫罗兰在变量花瓣宽 (即 PETALWID) 上的分布图 (见报表 39.1a)。PROC DISCRIM 则导出一个二次式的判别函数以用来鉴别各属种的紫罗兰 (见报表 39.1b)。

PROC DISCRIM 以选项 SIMPLE 来要求将四个变量的简单统计值印出。同时用 WCOV，WCORR，PCOV 及 PCORR 等选项要求 SAS 印出三种紫罗兰类别里的共变量矩阵、相关系数矩阵、共变异数矩阵的平均以及净相关系数矩阵，并以选项 LISTERR 要求印出被分错的紫罗兰。最后用选项 POOL=TEST 来比较这三种类别里的共变异数矩阵是否相同。因省略 SLPOOL= 选项，故采内设的 .10 显著水准。检定的结果是显著的，故判别函数会以各个类别的共变异数矩阵以及一个二次方的判别函数为分类的基础。

程　序

```
PROC FORMAT;
     VALUE SPECNAME
     1='SETOSA'
     2='VERSICOLOR'
     3='VIRGINICA';
     VALUE SPECCHAR
     1='S'    2='O'    3='V';
DATA IRIS;
    TITLE 'FISHER (1936) IRIS DATA';
    INPUT SEPALLEN SEPALWID PETALLEN PETALWID SPECIES @@;
    FORMAT SPECIES SPECNAME.;
    LABEL SEPALLEN='SEPAL LENGTH IN MM.'
          SEPALWID='SEPAL WIDTH  IN MM.'
          PETALLEN='PETAL LENGTH IN MM.'
          PETALWID='PETAL WIDTH  IN MM.';CARDS;
  (数据及 PROC DISCRIM 的程序请见下页)
50  33  14  02  1  64  28  56  22  3  65  28  46  15  2
67  31  56  24  3  63  28  51  15  3  46  34  14  03  1
69  31  51  23  3  62  22  45  15  2  59  32  48  18  2
46  36  10  02  1  61  30  46  14  2  60  27  51  16  2
65  30  52  20  3  56  25  39  11  2  65  30  55  18  3
58  27  51  19  3  68  32  59  23  3  51  33  17  05  1
57  28  45  13  2  62  34  54  23  3  77  38  67  22  3
63  33  47  16  2  67  33  57  25  3  76  30  66  21  3
49  25  45  17  3  55  35  13  02  1  67  30  52  23  3
```

```
70  32  47  14  2  64  32  45  15  2  61  28  40  13  2
48  31  16  02  1  59  30  51  18  3  55  24  38  11  2
63  25  50  19  3  64  32  53  23  3  52  34  14  02  1
49  36  14  01  1  54  30  45  15  2  79  38  64  20  3
44  32  13  02  1  67  33  57  21  3  50  35  16  06  1
58  26  40  12  2  44  30  13  02  1  77  28  67  20  3
63  27  49  18  3  47  32  16  02  1  55  26  44  12  2
50  23  33  10  2  72  32  60  18  3  48  30  14  03  1
51  38  16  02  1  61  30  49  18  3  48  34  19  02  1
50  30  16  02  1  50  32  12  02  1  61  26  56  14  3
64  28  56  21  3  43  30  11  01  1  58  40  12  02  1
51  38  19  04  1  67  31  44  14  2  62  28  48  18  3
49  30  14  02  1  51  35  14  02  1  56  30  45  15  2
58  27  41  10  2  50  34  16  04  1  46  32  14  02  1
60  29  45  15  2  57  26  35  10  2  57  44  15  04  1
50  36  14  02  1  77  30  61  23  3  63  34  56  24  3
58  27  51  19  3  57  29  42  13  2  72  30  58  16  3
54  34  15  04  1  52  41  15  01  1  71  30  59  21  3
64  31  55  18  3  60  30  48  18  3  63  29  56  18  3
49  24  33  10  2  56  27  42  13  2  57  30  42  12  2
55  42  14  02  1  49  31  15  02  1  77  26  69  23  3
60  22  50  15  3  54  39  17  04  1  66  29  46  13  2
52  27  39  14  2  60  34  45  16  2  50  34  15  02  1
44  29  14  02  1  50  20  35  10  2  55  24  37  10  2
58  27  39  12  2  47  32  13  02  1  46  31  15  02  1
69  32  57  23  3  62  29  43  13  2  74  28  61  19  3
59  30  42  15  2  51  34  15  02  1  50  35  13  03  1
56  28  49  20  3  60  22  40  10  2  73  29  63  18  3
67  25  58  18  3  49  31  15  01  1  67  31  47  15  2
63  23  44  13  2  54  37  15  02  1  56  30  41  13  2
63  25  49  15  2  61  28  47  12  2  64  29  43  13  2
51  25  30  11  2  57  28  41  13  2  65  30  58  22  3
69  31  54  21  3  54  39  13  04  1  51  35  14  03  1
72  36  61  25  3  65  32  51  20  3  61  29  47  14  2
56  29  36  13  2  69  31  49  15  2  64  27  53  19  3
68  30  55  21  3  55  25  40  13  2  48  34  16  02  1
48  30  14  01  1  45  23  13  03  1  57  25  50  20  3
57  38  17  03  1  51  38  15  03  1  55  23  40  13  2
66  30  44  14  2  68  28  48  14  2  54  34  17  02  1
```

```
51  37  15  04  1  52  35  15  02  1  58  28  51  24  3
67  30  50  17  2  63  33  60  25  3  53  37  15  02  1
;
PROC CHART DATA=IRIS;
    VBAR PETALWID/SUBGROUP=SPECIES MIDPOINTS=0 TO 30;
    FORMAT SPECIES SPECCHAR.;
RUN;
PROC DISCRIM SIMPLE WCOV WCORR PCOV PCORR LISTERR POOL=TEST;
    CLASS SPECIES;
RUN;
```

结　果

报表 39.1a 的分布图显示：花瓣宽似乎足以区分紫罗兰的属种，因为 SETOSA 种的花瓣最窄 (介于 0 至 6 公厘间)，VERSICOLOR 种的次之 (介于 10 至 18 公厘间)，VIRGINICA 种最宽 (介于 16 至 25 公厘间)。

报表 39.1b 证实上述的想法；当我们采用各类别内的共变异数矩阵以及一个二次方的判别函数为分类基础时，五十株 SETOSA 种的紫罗兰没有一株被错分的，然而其它两品种则各有一株 (属 VIRGINICA 种) 或两株 (属 VERSICOLOR 种) 被彼此错分到对方一组内了。

报表 39.1a　三种紫罗兰在 PETALWID 上的分布图

FISHER'S (1936) IRIS DATA
FREQUENCY OF PETALWID

PETALWID MIDPOINT		FREQ	CUM FREQ	PERCENT	CUM PERCENT
0		0	0	0.00	0.00
1	SSSSS	5	5	3.33	3.33
2	SSSSSSSSSSSSSSSSSSSSSSSSSSSSSS	29	34	19.33	22.67
3	SSSSSSS	7	41	4.67	27.33
4	SSSSSSS	7	48	4.67	32.00
5	S	1	49	0.67	32.67
6	S	1	50	0.67	33.33
7		0	50	0.00	33.33
8		0	50	0.00	33.33
9		0	50	0.00	33.33
10	0000000	7	57	4.67	38.00
11	000	3	60	2.00	40.00
12	00000	5	65	3.33	43.33
13	0000000000000	13	78	8.67	52.00
14	0000000V	8	86	5.33	57.33
15	0000000000VV	12	98	8.00	65.33
16	000V	4	102	2.67	68.00
17	0V	2	104	1.33	69.33
18	0VVVVVVVVVV	12	116	8.00	77.33
19	VVVVV	5	121	3.33	80.67
20	VVVVVV	6	127	4.00	84.67

21	VVVVVV	6	133	4.00	88.67
22	VVV	3	136	2.00	90.67
23	VVVVVVVV	8	144	5.33	96.00
24	VVV	3	147	2.00	98.00
25	VVV	3	150	2.00	100.00
26		0	150	0.00	100.00
27		0	150	0.00	100.00
28		0	150	0.00	100.00
29		0	150	0.00	100.00
30		0	150	0.00	100.00

```
-----+----+----+----+----+----+----
     5    10   15   20   25
```

SYMBOL S = SPECIES S　　SYMBOL 0 = SPECIES 0　　SYMBOL V = SPECIES V

报表 39.1b　费氏紫罗兰的分类鉴别分析

FISHER (1936) IRIS DATA

DISCRIMINANT ANALYSIS

150 Observations　　149 DF Total

4 Variables　　147 DF Within Classes

3 Classes　　2 DF Between Classes

Class Level Information

SPECIES	Frequency	Weight	Proportion	Prior Probability
SETOSA	50	50.0000	0.333333	0.333333
VERSIC	50	50.0000	0.333333	0.333333
VIRGIN	50	50.0000	0.333333	0.333333

DISCRIMINANT ANALYSIS　　WITHIN-CLASS COVARIANCE MATRICES

SPECIES = SETOSA　　DF = 49

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID	
SEPALLEN	12.42489796	9.92163265	1.63551020	1.03306122	SEPAL LENGTH IN MM.

SEPALWID	9.92163265	14.36897959	1.16979592	0.92979592	SEPAL WIDTH IN MM.
PETALLEN	1.63551020	1.16979592	3.01591837	0.60693878	PETAL LENGTH IN MM.
PETALWID	1.03306122	0.92979592	0.60693878	1.11061224	PETAL WIDTH IN MM.

SPECIES = VERSIC DF = 49

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID	
SEPALLEN	26.64326531	8.51836735	18.28979592	5.57795918	SEPAL LENGTH IN MM.
SEPALWID	8.51836735	9.84693878	8.26530612	4.12040816	SEPAL WIDTH IN MM.
PETALLEN	18.28979592	8.26530612	22.08163265	7.31020408	PETAL LENGTH IN MM.
PETALWID	5.57795918	4.12040816	7.31020408	3.91061224	PETAL WIDTH IN MM.

SPECIES = VIRGIN DF = 49

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID	
SEPALLEN	40.43428571	9.37632653	30.32897959	4.90938776	SEPAL LENGTH IN MM.
SEPALWID	9.37632653	10.40040816	7.13795918	4.76285714	SEPAL WIDTH IN MM.
PETALLEN	30.32897959	7.13795918	30.45877551	4.88244898	PETAL LENGTH IN MM.
PETALWID	4.90938776	4.76285714	4.88244898	7.54326531	PETAL WIDTH IN MM.

Pooled Within-Class Covariance Matrix DF = 147

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID	
SEPALLEN	26.50081633	9.27210884	16.75142857	3.84013605	SEPAL LENGTH IN MM.
SEPALWID	9.27210884	11.53877551	5.52435374	3.27102041	SEPAL WIDTH IN MM.
PETALLEN	16.75142857	5.52435374	18.51877551	4.26653061	PETAL LENGTH IN MM.
PETALWID	3.84013605	3.27102041	4.26653061	4.18816327	PETAL WIDTH IN MM.

Within-Class Correlation Coefficients / Prob > |R|

SPECIES = SETOSA

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.00000	0.74255	0.26718	0.27810
SEPAL LENGTH IN MM.	0.0	0.0001	0.0607	0.0505
SEPALWID	0.74255	1.00000	0.17770	0.23275
SEPAL WIDTH IN MM.	0.0001	0.0	0.2170	0.1038

PETALLEN	0.26718	0.17770	1.00000	0.33163
PETAL LENGTH IN MM.	0.0607	0.2170	0.0	0.0186
PETALWID	0.27810	0.23275	0.33163	1.00000
PETAL WIDTH IN MM.	0.0505	0.1038	0.0186	0.0

SPECIES = VERSIC

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.00000	0.52591	0.75405	0.54646
SEPAL LENGTH IN MM.	0.0	0.0001	0.0001	0.0001
SEPALWID	0.52591	1.00000	0.56052	0.66400
SEPAL WIDTH IN MM.	0.0001	0.0	0.0001	0.0001
PETALLEN	0.75405	0.56052	1.00000	0.78667
PETAL LENGTH IN MM.	0.0001	0.0001	0.0	0.0001
PETALWID	0.54646	0.66400	0.78667	1.00000
PETAL WIDTH IN MM.	0.0001	0.0001	0.0001	0.0

SPECIES = VIRGIN

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.00000	0.45723	0.86422	0.28111
SEPAL LENGTH IN MM.	0.0	0.0008	0.0001	0.0480
SEPALWID	0.45723	1.00000	0.40104	0.53773
SEPAL WIDTH IN MM.	0.0008	0.0	0.0039	0.0001
PETALLEN	0.86422	0.40104	1.00000	0.32211
PETAL LENGTH IN MM.	0.0001	0.0039	0.0	0.0225
PETALWID	0.28111	0.53773	0.32211	1.00000
PETAL WIDTH IN MM.	0.0480	0.0001	0.0225	0.0

Pooled Within-Class Correlation Coefficients / Prob > |R|

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.00000	0.53024	0.75616	0.36451
SEPAL LENGTH IN MM.	0.0	0.0001	0.0001	0.0001
SEPALWID	0.53024	1.00000	0.37792	0.47053
SEPAL WIDTH IN MM.	0.0001	0.0	0.0001	0.0001
PETALLEN	0.75616	0.37792	1.00000	0.48446
PETAL LENGTH IN MM.	0.0001	0.0001	0.0	0.0001
PETALWID	0.36451	0.47053	0.48446	1.00000
PETAL WIDTH IN MM.	0.0001	0.0001	0.0001	0.0

Total-Sample

Variable	N	Sum	Mean	Variance	Std Dev	Label
SEPALLEN	150	8765	58.43333	68.56935	8.28066	SEPAL LENGTH IN MM.
SEPALWID	150	4586	30.57333	18.99794	4.35866	SEPAL WIDTH IN MM.
PETALLEN	150	5637	37.58000	311.62779	17.65298	PETAL LENGTH IN MM.
PETALWID	150	1799	11.99333	58.10063	7.62238	PETAL WIDTH IN MM.

SPECIES = SETOSA

Variable	N	Sum	Mean	Variance	Std Dev	Label
SEPALLEN	50	2503	50.06000	12.42490	3.52490	SEPAL LENGTH IN MM.
SEPALWID	50	1714	34.28000	14.36898	3.79064	SEPAL WIDTH IN MM.
PETALLEN	50	731.00000	14.62000	3.01592	1.73664	PETAL LENGTH IN MM.
PETALWID	50	123.00000	2.46000	1.11061	1.05386	PETAL WIDTH IN MM.

SPECIES = VERSIC

Variable	N	Sum	Mean	Variance	Std Dev	Label
SEPALLEN	50	2968	59.36000	26.64327	5.16171	SEPAL LENGTH IN MM.
SEPALWID	50	1385	27.70000	9.84694	3.13798	SEPAL WIDTH IN MM.
PETALLEN	50	2130	42.60000	22.08163	4.69911	PETAL LENGTH IN MM.
PETALWID	50	663.00000	13.26000	3.91061	1.97753	PETAL WIDTH IN MM.

SPECIES = VIRGIN

Variable	N	Sum	Mean	Variance	Std Dev	Label
SEPALLEN	50	3294	65.88000	40.43429	6.35880	SEPAL LENGTH IN MM.
SEPALWID	50	1487	29.74000	10.40041	3.22497	SEPAL WIDTH IN MM.
PETALLEN	50	2776	55.52000	30.45878	5.51895	PETAL LENGTH IN MM.
PETALWID	50	1013	20.26000	7.54327	2.74650	PETAL WIDTH IN MM.

DISCRIMINANT ANALYSIS　WITHIN COVARIANCE MATRIX INFORMATION

SPECIES	Covariance Matrix Rank	Natural Log of Determinant of the Covariance Matrix
SETOSA	4	5.35332
VERSIC	4	7.54636
VIRGIN	4	9.49362
Pooled	4	8.46214

DISCRIMINANT ANALYSIS　TEST OF HOMOGENEITY OF WITHIN COVARIANCE MATRICES

Notation: K = Number of Groups　　N = Total Number of Observations - Number of Groups

P = Number of Variables　　N(i) = Number of Observations in the i'th Group - 1

$$V = \frac{\prod |\text{Within SS Matrix(i)}|^{N(i)/2}}{|\text{Pooled SS Matrix}|^{N/2}}$$

$$RHO = 1.0 - \left(\sum \frac{1}{N(i)} - \frac{1}{N} \right) \frac{2P^2 + 3P - 1}{6(P+1)(K-1)}, \text{WHERE } DF = .5(K-1)P(P+1)$$

Under null hypothesis: $-2\ RHO \ln \left(\frac{N^{PN/2} \, V}{\prod N(i)^{PN(i)/2}} \right)$ is distributed approximately as chi-square(DF)

Test Chi-Square Value = 140.943050　with 20 DF　Prob > Chi-Sq = 0.0001

Since the chi-square value is significant at the 0.1000 level,
the within covariance matrices will be used in the discriminant function.

Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

Pairwise Generalized Squared Distances Between Groups

$$D^2(i|j) = (\bar{X}_i - \bar{X}_j)' \, COV_j^{-1} \, (\bar{X}_i - \bar{X}_j) + \ln |COV_j|$$

Generalized Squared Distance to SPECIES

From SPECIES	SETOSA	VERSIC	VIRGIN
SETOSA	5.35332	110.74017	178.26121
VERSIC	328.41535	7.54636	23.33238
VIRGIN	711.43826	25.41306	9.49362

Classification Results for Calibration Data: WORK.IRIS

Resubstitution Results using Quadratic Discriminant Function

Generalized Squared Distance Function:

$$D_j^2(X) = (X-\bar{X}_j)' \, COV_j^{-1} \, (X-\bar{X}_j) + \ln |COV_j|$$

Posterior Probability of Membership in each SPECIES:

$$Pr(j|X) = \exp(-.5 \, D_j^2(X)) \, / \, \sum_k \exp(-.5 \, D_k^2(X))$$

Posterior Probability of Membership in SPECIES:

Obs	From SPECIES	Classified into SPECIES	SETOSA	VERSIC	VIRGIN
5	VIRGIN	VERSIC *	0.0000	0.6050	0.3950
9	VERSIC	VIRGIN *	0.0000	0.3359	0.6641
12	VERSIC	VIRGIN *	0.0000	0.1543	0.8457

* Misclassified observation

Classification Summary for Calibration Data: WORK.IRIS

Resubstitution Summary using Quadratic Discriminant Function

Generalized Squared Distance Function:

$$D_j^2(X) = (X-\bar{X}_j)' \, COV_j^{-1} \, (X-\bar{X}_j) + \ln |COV_j|$$

Posterior Probability of Membership in each SPECIES:

$$Pr(j|X) = \exp(-.5\, D_j^2(X)) \,/\, SUM_k \exp(-.5\, D_k^2(X))$$

Number of Observations and Percent Classified into SPECIES:

From SPECIES	SETOSA	VERSIC	VIRGIN	Total
SETOSA	50	0	0	50
	100.00	0.00	0.00	100.00
VERSIC	0	48	2	50
	0.00	96.00	4.00	100.00
VIRGIN	0	1	49	50
	0.00	2.00	98.00	100.00
Total	50	49	51	150
Percent	33.33	32.67	34.00	100.00
Priors	0.3333	0.3333	0.3333	

Error Count Estimates for SPECIES:

	SETOSA	VERSIC	VIRGIN	Total
Rate	0.0000	0.0400	0.0200	0.0200
Priors	0.3333	0.3333	0.3333	

例二：费氏紫罗兰的无参数鉴别分析（上）

承袭例一的数据以及结论 (即花瓣宽=PETALWID 足以区分这三种不同的紫罗兰)，我们采用无参数的方法继续执行鉴别分析。本例中的无参数分析法是由 METHOD=NPAR，KERNEL=NORMAL (常态的颗粒函数) 以 R=.4 (椭圆形的半径) 定义的。此外 POOL=YES 会使参数的估计比较圆滑，分析的结果以 PROC PLOT 的图形来表示。

程　序

```
DATA PLOTDATA;
    DO PETALWID=-5 TO 30 BY .5;
      OUTPUT;
    END;
PROC DISCRIM DATA=IRIS TESTDATA=PLOTDATA TESTOUT=PLOTP TESTOUTD=PLOTD
          METHOD=NPAR KERNEL=NORMAL R=.4 POOL=YES
          SHORT NOCLASSIFY CROSSLISTERR;
    CLASS SPECIES;
    VAR PETALWID;
    TITLE2 'Using Kernel Density Estimates with Equal Bandwidth';
RUN;
DATA PLOTD;
    SET PLOTD;
    IF SETOSA<.002 THEN SETOSA=.;
    IF VERSICOL<.002 THEN VERSICOL=.;
    IF VIRGINIC<.002 THEN VERGINIC=.;
RUN;
PROC PLOT DATA=PLOTD;
    PLOT SETOSA*PETALWID='S'
        VERSICOL*PETALWID='O'
        VIRGINIC*PETALWID='V'
        /OVERLAY VPOS=27 VAXIS=0 TO .6 BY .1;
    TITLE3 'Plot of Estimated Densities';
RUN;
DATA PLOTP;
    SET PLOTP;
    IF SETOSA<.01 THEN SETOSA=.;
    IF VERSICOL<.01 THEN VERSICOL=.;
    IF VIRGINIC <.01 THEN VIRGINIC=.;
RUN;
PROC PLOT DATA=PLOTP;
    PLOT SETOSA*PETALWID='S'
        VERSICOL*PETALWID='O'
        VIRGINIC*PETALWID='V'
```

```
            /OVERLAY VPOS=18 VAXIS=0 TO 1 BY .2;
        TITLE3 'Plot of Posterior Probabilities';
    RUN;
```

结　果

根据错误分类的次数，花瓣宽的测量值是分辨这三种紫罗兰的良好指标 (对 SETOSA 属种而言，错误分类的比例是零，对 VERSICOLOR 及 VIRGINICA 两属种而言，其比例分别是 0.04 及 0.08)。PROC PLOT 的两个图形亦证实了这个结论。

报表 39.2　费氏紫罗兰的分类鉴别分析（上）

FISHER (1936) IRIS DATA

Using Kernel Density Estimates with Equal Bandwidth

Discriminant Analysis

150 Observations	149 DF Total
1 Variables	147 DF Within Classes
3 Classes	2 DF Between Classes

Class Level Information

SPECIES	Output SAS Name	Frequency	Weight	Proportion	Prior Probability
SETOSA	SETOSA	50	50.0000	0.333333	0.333333
VERSICOLOR	VERSICOL	50	50.0000	0.333333	0.333333
VIRGINICA	VIRGINIC	50	50.0000	0.333333	0.333333

Discriminant Analysis　Classification Results for Calibration Data: WORK.IRIS

Cross-validation Results using Normal Kernel Density

Squared Distance Function:

$$D^2(X,Y) = (X-Y)' \, COV^{-1} \, (X-Y)$$

Posterior Probability of Membership in each SPECIES:

$$F(X \mid j) = n_j^{-1} \sum_i \exp(-.5 \, D^2_{ji}(X,Y_i) / R^2)$$

$$Pr(j \mid X) = PRIOR_j \, F(Xj) \, / \, \sum_k PRIOR_k \, F(X \mid k)$$

Posterior Probability of Membership in SPECIES:

Obs	From SPECIES	Classified into SPECIES	SETOSA	VERSICOLOR	VIRGINICA
5	VIRGINICA	VERSICOLOR *	0.0000	0.8827	0.1173
9	VERSICOLOR	VIRGINICA *	0.0000	0.0438	0.9562
57	VIRGINICA	VERSICOLOR *	0.0000	0.9472	0.0528
78	VIRGINICA	VERSICOLOR *	0.0000	0.8061	0.1939
91	VIRGINICA	VERSICOLOR *	0.0000	0.8827	0.1173
148	VERSICOLOR	VIRGINICA *	0.0000	0.2586	0.7414

* Misclassified observation

Cross-validation Summary using Normal Kernel Density

Squared Distance Function:

$$D^2(X,Y) = (X-Y)' \; COV^{-1} \; (X-Y)$$

Posterior Probability of Membership in each SPECIES:

$$F(X|j) = n_j^{-1} \sum_i \exp(-.5 \; D^2_{ji}(X,Y_i) / R^2)$$

$$Pr(j|X) = PRIOR_j \; F(X|j) \; / \; \sum_k PRIOR_k \; F(X|k)$$

Number of Observations and Percent Classified into SPECIES:

From SPECIES	SETOSA	VERSICOLOR	VIRGINICA	Total
SETOSA	50	0	0	50
	100.00	0.00	0.00	100.00
VERSICOLOR	0	48	2	50
	0.00	96.00	4.00	100.00
VIRGINICA	0	4	46	50
	0.00	8.00	92.00	100.00
Total	50	52	48	150
Percent	33.33	34.67	32.00	100.00
Priors	0.3333	0.3333	0.3333	

Error Count Estimates for SPECIES:

	SETOSA	VERSICOLOR	VIRGINICA	Total
Rate	0.0000	0.0400	0.0800	0.0400
Priors	0.3333	0.3333	0.3333	

Classification Summary using Normal Kernel Density

Squared Distance Function:

$$D^2(X,Y) = (X-Y)' \, COV^{-1} \, (X-Y)$$

Posterior Probability of Membership in each SPECIES:

$$F(X|j) = n_j \sum_i \exp(-.5 \, D^2(X,Y_{ji}) / R^2)$$

$$Pr(j|X) = PRIOR_j \, F(X|j) / \sum_k PRIOR_k \, F(X|k)$$

Number of Observations and Percent Classified into SPECIES:

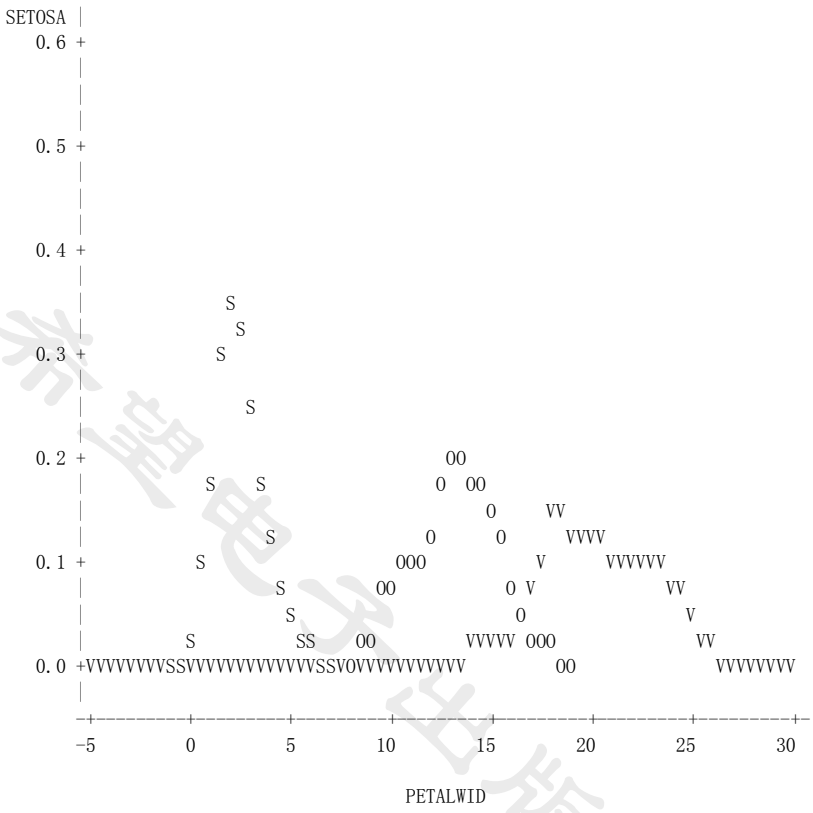
	SETOSA	VERSICOLOR	VIRGINICA	Total
Total	26	18	27	71
Percent	36.62	25.35	38.03	100.00
Priors	0.3333	0.3333	0.3333	

Plot of Estimated Densities

Plot of SETOSA*PETALWID.　Symbol used is 'S'.

Plot of VERSICOL*PETALWID.　Symbol used is 'O'.

Plot of VIRGINIC*PETALWID.　Symbol used is 'V'.



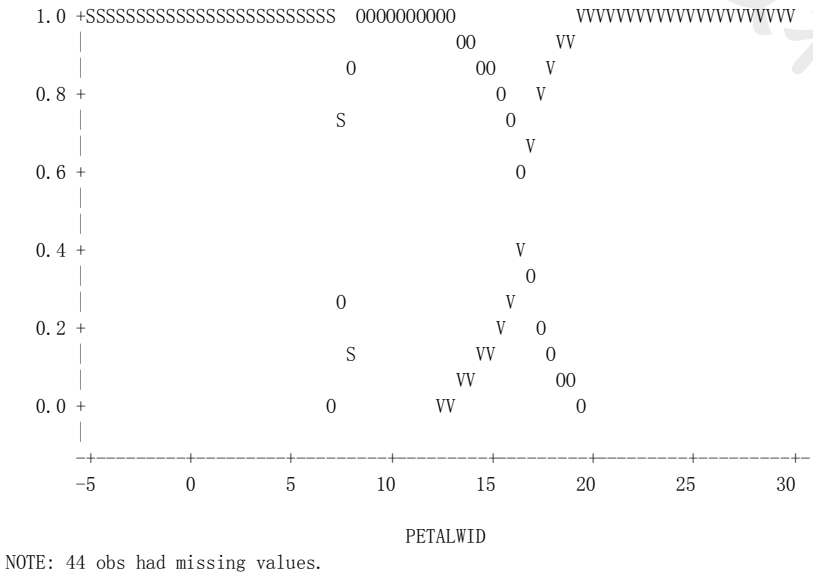
NOTE: 54 obs had missing values. 6 obs hidden.

Plot of Posterior Probabilities

Plot of SETOSA*PETALWID. Symbol used is 'S'.

Plot of VERSICOL*PETALWID. Symbol used is '0'.

Plot of VIRGINIC*PETALWID. Symbol used is 'V'.



NOTE: 44 obs had missing values.

例三：费氏紫罗兰的无参数鉴别分析（下）

本例的程序与例二完全一致：均采用了无参数的鉴别分析以及 .4 的椭圆半径。唯一不同的地方是：选项 POOL=NO 的界定会使参数的估计值分岐比较大；这个现象由 PROC PLOT 的第一个图形可看出来。

程　序

```
PROC DISCRIM DATA=IRIS TESTDATA=PLOTDATA TESTOUT=PLOTP TESTOUTD=PLOTD
          METHOD=NPAR KERNEL=NORMAL R=.4 POOL=NO
          SHORT NOCLASSIFY CROSSLISTERR;
    CLASS SPECIES;
    VAR PETALWID;
    TITLE2 'Using Kernel Density Estimates with Unequal Bandwidth';
RUN;
DATA PLOTD;
    SET PLOTD;
    IF SETOSA<.002 THEN SETOSA=.;
    IF VERSICOL<.002 THEN VERSICOL=.;
    IF VIRGINIC<.002 THEN VERGINIC=.;
RUN;
PROC PLOT DATA=PLOTD;
    PLOT SETOSA*PETALWID='S'
        VERSICOL*PETALWID='O'
        VIRGINIC*PETALWID='V'
        /OVERLAY VPOS=27 VAXIS=0 TO .6 BY .1;
    TITLE3 'Plot of Estimated Densities';
RUN;
DATA PLOTP;
     SET PLOTP;
     IF SETOSA<.01 THEN SETOSA=.;
     IF VERSICOL<.01 THEN VERSICOL=.;
     IF VIRGINIC <.01 THEN VIRGINIC=.;
RUN;
PROC PLOT DATA=PLOTP;
     PLOT SETOSA*PETALWID='S'
        VERSICOL*PETALWID='O'
        VIRGINIC*PETALWID='V'
        /OVERLAY VPOS=18 VAXIS=0 TO 1 BY .2;
    TITLE3 'Plot of Posterior Probabilities';          RUN;
```

结　果

根据错误分类的次数，本例分析的结果与例二完全一样。不过由于 POOL=NO 的设定，PROCPLOT 所产生的第一个图形看起来比例二相对的图形分歧更显著 (亦即三种紫罗兰的概率分布的差异很大)。

报表 39.3　费氏紫罗兰的分类鉴别分析 (下)

FISHER (1936) IRIS DATA

Using Kernel Density Estimates with Unequal Bandwidth

Discriminant Analysis

150 Observations　　149 DF Total

1 Variables　　147 DF Within Classes

3 Classes　　2 DF Between Classes

Class Level Information

SPECIES	Output SAS Name	Frequency	Weight	Proportion	Prior Probability
SETOSA	SETOSA	50	50.0000	0.333333	0.333333
VERSICOLOR	VERSICOL	50	50.0000	0.333333	0.333333
VIRGINICA	VIRGINIC	50	50.0000	0.333333	0.333333

Discriminant Analysis　Classification Results for Calibration Data: WORK.IRIS

Cross-validation Results using Normal Kernel Density

Squared Distance Function:

$$D^2_j(X,Y) = (X-Y)' \, COV_j^{-1} \, (X-Y)$$

Posterior Probability of Membership in each SPECIES:

$$F(Xj) = n_j \sum_i \exp(-.5 \, D^2_{ji}(X,Y_i) / R^2)$$

$$Pr(jX) = PRIOR_j \, F(Xj) / \sum_k PRIOR_k \, F(Xk)$$

Posterior Probability of Membership in SPECIES:

Obs	From SPECIES	Classified into SPECIES	SETOSA	VERSICOLOR	VIRGINICA
5	VIRGINICA	VERSICOLOR *	0.0000	0.8805	0.1195
9	VERSICOLOR	VIRGINICA *	0.0000	0.0466	0.9534
57	VIRGINICA	VERSICOLOR *	0.0000	0.9394	0.0606
78	VIRGINICA	VERSICOLOR *	0.0000	0.7193	0.2807
91	VIRGINICA	VERSICOLOR *	0.0000	0.8805	0.1195
148	VERSICOLOR	VIRGINICA *	0.0000	0.2275	0.7725

* Misclassified observation

Discriminant Analysis　Classification Summary for Calibration Data: WORK.IRIS

Cross-validation Summary using Normal Kernel Density

Squared Distance Function:

$$D^2(X,Y) = (X-Y)' \mathrm{COV}_j^{-1} (X-Y)$$

Posterior Probability of Membership in each SPECIES:

$$F(X|j) = n_j \sum_i \exp(-.5\, D^2_{ji}(X,Y_i) / R^2)$$

$$\Pr(j|X) = \mathrm{PRIOR}_j F(X|j) / \sum_k \mathrm{PRIOR}_k F(X|k)$$

Number of Observations and Percent Classified into SPECIES:

From SPECIES	SETOSA	VERSICOLOR	VIRGINICA	Total
SETOSA	50	0	0	50
	100.00	0.00	0.00	100.00
VERSICOLOR	0	48	2	50
	0.00	96.00	4.00	100.00
VIRGINICA	0	4	46	50
	0.00	8.00	92.00	100.00
Total	50	52	48	150
Percent	33.33	34.67	32.00	100.00
Priors	0.3333	0.3333	0.3333	

Error Count Estimates for SPECIES:

	SETOSA	VERSICOLOR	VIRGINICA	Total
Rate	0.0000	0.0400	0.0800	0.0400
Priors	0.3333	0.3333	0.3333	

Classification Summary using Normal Kernel Density

Squared Distance Function:

$$D^2(X,Y) = (X-Y)' \mathrm{COV}_j^{-1} (X-Y)$$

Posterior Probability of Membership in each SPECIES:

$$F(X|j) = n_j \sum_i \exp(-.5\, D^2_{ji}(X,Y_i) / R^2)$$

$$\Pr(j|X) = \mathrm{PRIOR}_j F(X|j) / \sum_k \mathrm{PRIOR}_k F(X|k)$$

Number of Observations and Percent Classified into SPECIES:

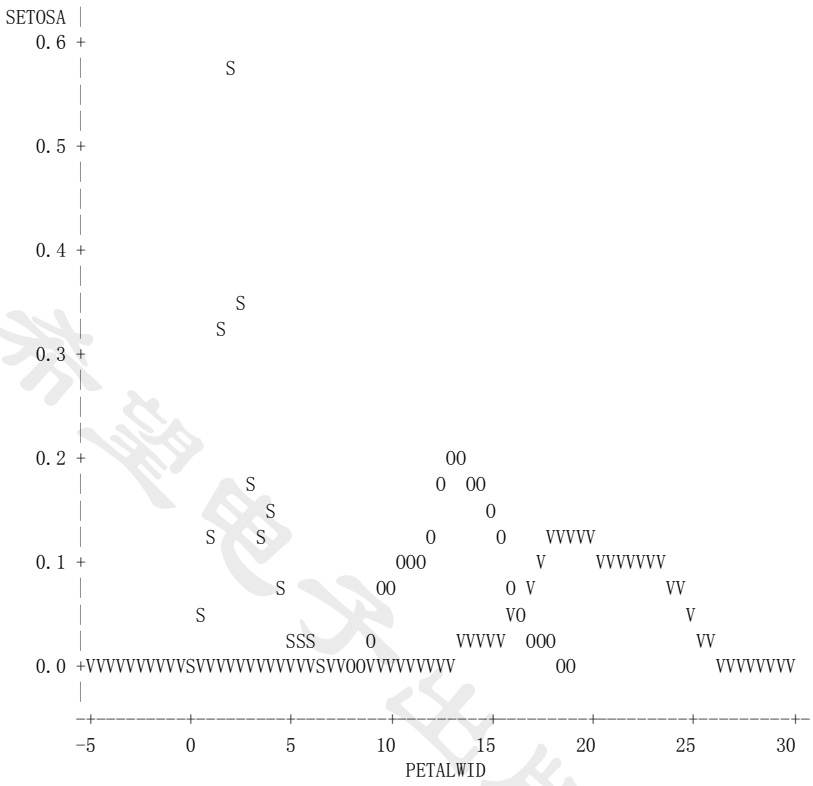
	SETOSA	VERSICOLOR	VIRGINICA	Total
Total	25	18	28	71
Percent	35.21	25.35	39.44	100.00
Priors	0.3333	0.3333	0.3333	

Plot of Estimated Densities

Plot of SETOSA*PETALWID.　Symbol used is 'S'.

Plot of VERSICOL*PETALWID.　Symbol used is 'O'.

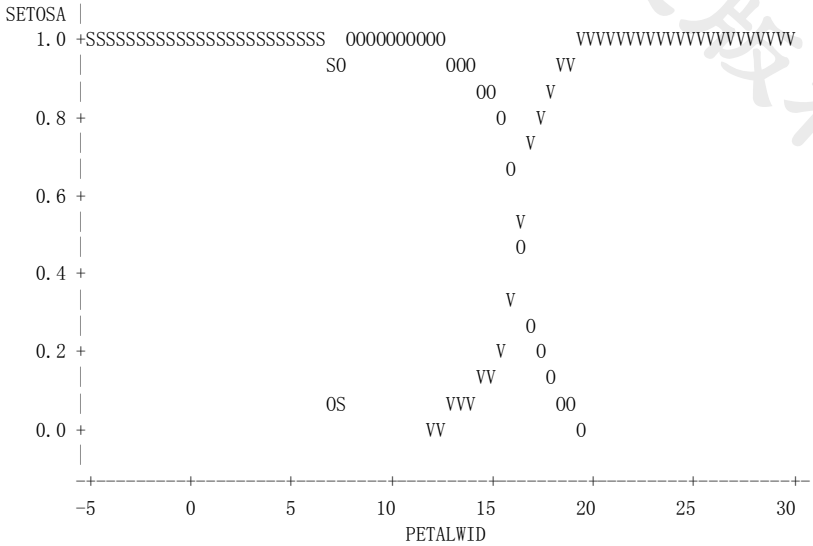
Plot of VIRGINIC*PETALWID.　Symbol used is 'V'.



NOTE: 57 obs had missing values. 5 obs hidden.

Plot of Posterior Probabilities

Plot of SETOSA*PETALWID. Symbol used is 'S'.
Plot of VERSICOL*PETALWID. Symbol used is '0'.
Plot of VIRGINIC*PETALWID. Symbol used is 'V'.



NOTE: 45 obs had missing values.

例四：杂粮的分类鉴别分析

这一个资料文件 (CROPS) 包含五种杂粮，即：玉米 (CORN)，棉花 (COTTON)，黄豆 (SOYBEANS)，甜菜 (SUGARBEETS) 及苜蓿 (CLOVER)。其中每一观察体又以四种属性 (简称 X1，X2，X3，X4) 来描述。首先用 PROC DISCRIM 对观察体执行分类鉴别分析，然后将分类的结果输入第二资料文件 (TESTDATA=TEST)，以决定判别函数的优劣。

程　序

```
DATA CROPS;
    TITLE 'REMOTE SENSING DATA ON FIVE CROPS';
    INPUT CROP $ 1-10 X1-X4 XVALUES $ 11-21; CARDS;
CORN      16 27 31 33
CORN      15 23 30 30
CORN      16 27 27 26
CORN      18 20 25 23
CORN      15 15 31 32
CORN      15 32 32 15
SOYBEANS  20 23 23 25
SOYBEANS  24 24 25 32
SOYBEANS  21 25 23 24
SOYBEANS  27 45 24 12
SOYBEANS  12 13 15 42
SOYBEANS  22 32 31 43
COTTON    31 32 33 34
COTTON    29 24 26 28
COTTON    34 32 28 45
COTTON    26 25 23 24
COTTON    53 48 75 26
COTTON    34 35 25 68
SUGARBEETS22 23 25 42
SUGARBEETS25 25 24 26
SUGARBEETS34 25 16 52
SUGARBEETS54 23 21 54
SUGARBEETS25 43 32 15
SUGARBEETS26 54  2 54
CLOVER    12 45 32 54
CLOVER    24 58 25 34
CLOVER    87 54 61 41
CLOVER    51 31 31 16
CLOVER    96 48 54 62
CLOVER    31 31 11 11
;
PROC DISCRIM DATA=CROPS POOL=YES LIST OUT=CROPCAL;
    CLASS CROP; ID XVALUES; VAR X1-X4;
    TITLE2 'CLASSIFICATION OF CROP DATA';
```

```
DATA TEST;
    INPUT CROP $ 1-10 X1-X4 XVALUES $ 11-21; CARDS;
CORN      16 27 31 33
SOYBEANS  21 25 23 24
COTTON    29 24 26 28
SUGARBEETS54 23 21 54
CLOVER    32 32 62 16
;
PROC DISCRIM DATA=CROPCAL TESTDATA=TEST TESTLIST; CLASS CROP; TESTCLASS CROP;
    TESTID XVALUES; VAR X1-X4; TITLE2 'CLASSIFICATION OF TEST DATA'; RUN;
```

结　果

分类鉴别分析的结果显示：玉米最易被辨认，所以没有任何错分的观察体。黄豆类中有三个观察体错分成苜蓿、玉米，或棉花。棉花类中也有三株被错分成黄豆 (两次) 或甜菜。甜菜类中有三株均被误认为黄豆。最后，苜蓿类，有两株被分派到棉花或甜菜类中。

第二次分类鉴别分析也再次证明：棉花与苜蓿的分类有问题。然而，玉米、黄豆、与甜菜的分类则无问题。这个结论与旧版程序 NEIGHBOR 的分析结果相近。

报表 39.4　杂粮的分类鉴别分析

REMOTE SENSING DATA ON FIVE CROPS

CLASSIFICATION OF CROP DATA

DISCRIMINANT ANALYSIS

30 Observations	29 DF Total
4 Variables	25 DF Within Classes
5 Classes	4 DF Between Classes

Class Level Information

CROP	Output SAS Name	Frequency	Weight	Proportion	Prior Probability
CLOVER	CLOVER	6	6.0000	0.200000	0.200000
CORN	CORN	6	6.0000	0.200000	0.200000
COTTON	COTTON	6	6.0000	0.200000	0.200000
SOYBEANS	SOYBEANS	6	6.0000	0.200000	0.200000
SUGARBEETS	SUGARBEE	6	6.0000	0.200000	0.200000

DISCRIMINANT ANALYSIS　POOLED COVARIANCE MATRIX INFORMATION

Covariance Matrix Rank	Natural Log of Determinant of the Covariance Matrix

4　20.2020782

Pairwise Generalized Squared Distances Between Groups

$$D^2(I|j) = (\bar{X}_i - \bar{X}_j)' \, COV^{-1} \, (\bar{X}_i - \bar{X}_j)$$

Generalized Squared Distance to CROP

From CROP	CLOVER	CORN	COTTON	SOYBEANS	SUGARBEETS
CLOVER	0	9.67112	3.27636	5.51437	2.66130
CORN	9.67112	0	2.14355	1.04981	5.19053
COTTON	3.27636	2.14355	0	1.10269	1.89695
SOYBEANS	5.51437	1.04981	1.10269	0	1.70143
SUGARBEETS	2.66130	5.19053	1.89695	1.70143	0

DISCRIMINANT ANALYSIS　LINEAR DISCRIMINANT FUNCTION

$$\text{Constant} = -.5 \, \bar{X}_j' \, COV^{-1} \, \bar{X}_j \qquad \text{Coefficient Vector} = COV^{-1} \, \bar{X}_j$$

CROP

	CLOVER	CORN	COTTON	SOYBEANS	SUGARBEETS
CONSTANT	-13.40052	-5.67592	-9.07274	-5.56509	-8.36779
X1	0.10408	-0.07978	-0.00248	-0.01844	0.05069
X2	0.37732	0.15532	0.23822	0.21107	0.28186
X3	0.01337	0.18755	0.13365	0.08930	-0.00669
X4	0.11869	0.12777	0.15391	0.12540	0.15386

Classification Results for Calibration Data: WORK.CROPS

Resubstitution Results using Linear Discriminant Function

Generalized Squared Distance Function:

$$D_j^2(X) = (X-\bar{X}_j)' \, COV^{-1} \, (X-\bar{X}_j)$$

Posterior Probability of Membership in each CROP:

$$Pr(j|X) = \exp(-.5 \, D_j^2(X)) \, / \, SUM_k \exp(-.5 \, D_k^2(X))$$

Classification Results for Calibration Data: WORK.CROPS

Resubstitution Results using Linear Discriminant Function

Posterior Probability of Membership in CROP:

XVALUES	From CROP	Classified into CROP		CLOVER	CORN	COTTON	SOYBEANS	SUGARBEETS
16 27 31 33	CORN	CORN		0.0052	0.4589	0.2211	0.2710	0.0438
15 23 30 30	CORN	CORN		0.0025	0.5354	0.1673	0.2644	0.0304
16 27 27 26	CORN	CORN		0.0093	0.3869	0.1926	0.3442	0.0670
18 20 25 23	CORN	CORN		0.0046	0.4336	0.1452	0.3618	0.0548
15 15 31 32	CORN	CORN		0.0004	0.6829	0.1097	0.1948	0.0122
15 32 32 15	CORN	CORN		0.0127	0.4539	0.1813	0.3151	0.0369
20 23 23 25	SOYBEANS	SOYBEANS		0.0125	0.2982	0.1754	0.4028	0.1111
24 24 25 32	SOYBEANS	SOYBEANS		0.0177	0.2459	0.2309	0.3628	0.1427
21 25 23 24	SOYBEANS	SOYBEANS		0.0200	0.2530	0.1849	0.4072	0.1348
27 45 24 12	SOYBEANS	CLOVER	*	0.4781	0.0253	0.1067	0.1671	0.2228
12 13 15 42	SOYBEANS	CORN	*	0.0014	0.3836	0.1274	0.3829	0.1047
22 32 31 43	SOYBEANS	COTTON	*	0.0223	0.2374	0.3576	0.2614	0.1213
31 32 33 34	COTTON	COTTON		0.0626	0.1664	0.3565	0.2671	0.1475
29 24 26 28	COTTON	SOYBEANS	*	0.0315	0.2001	0.2360	0.3670	0.1653
34 32 28 45	COTTON	COTTON		0.0953	0.0675	0.3184	0.2074	0.3114
26 25 23 24	COTTON	SOYBEANS	*	0.0361	0.1823	0.1962	0.3988	0.1865
53 48 75 26	COTTON	COTTON		0.0967	0.1805	0.6730	0.0448	0.0050
34 35 25 68	COTTON	SUGARBEETS	*	0.0848	0.0225	0.2924	0.1041	0.4961
22 23 25 42	SUGARBEETS	SOYBEANS	*	0.0098	0.2692	0.2589	0.3245	0.1376
25 25 24 26	SUGARBEETS	SOYBEANS	*	0.0286	0.2099	0.2086	0.3895	0.1635
34 25 16 52	SUGARBEETS	SUGARBEETS		0.0574	0.0253	0.1534	0.1685	0.5953
54 23 21 54	SUGARBEETS	SUGARBEETS		0.1521	0.0064	0.1248	0.0796	0.6371
25 43 32 15	SUGARBEETS	SOYBEANS	*	0.2317	0.1142	0.2460	0.2704	0.1377
26 54 2 54	SUGARBEETS	SUGARBEETS		0.4061	0.0001	0.0090	0.0090	0.5759
12 45 32 54	CLOVER	COTTON	*	0.0359	0.1767	0.4560	0.1921	0.1392
24 58 25 34	CLOVER	CLOVER		0.6495	0.0048	0.0801	0.0472	0.2183
87 54 61 41	CLOVER	CLOVER		0.9642	0.0001	0.0245	0.0010	0.0102
51 31 31 16	CLOVER	CLOVER		0.4550	0.0229	0.1475	0.1506	0.2241
96 48 54 62	CLOVER	CLOVER		0.9534	0.0000	0.0193	0.0006	0.0267
31 31 11 11	CLOVER	SUGARBEETS	*	0.2582	0.0151	0.0533	0.2099	0.4635

* Misclassified observation

Classification Summary for Calibration Data: WORK.CROPS

Resubstitution Summary using Linear Discriminant Function

Generalized Squared Distance Function:

$$D_j^2(X) = (X-\bar{X}_j)' \, COV^{-1} \, (X-\bar{X}_j)$$

Posterior Probability of Membership in each CROP:

$$Pr(j|X) = \exp(-.5\, D_j^2(X)) \,/\, SUM_k \exp(-.5\, D_k^2(X))$$

Number of Observations and Percent Classified into CROP:

From CROP	CLOVER	CORN	COTTON	SOYBEANS	SUGARBEETS	Total
CLOVER	4	0	1	0	1	6
	66.67	0.00	16.67	0.00	16.67	100.00
CORN	0	6	0	0	0	6
	0.00	0.00	0.00	0.00	0.00	100.00
COTTON	0	0	3	2	1	6
	0.00	0.00	50.00	33.33	16.67	100.00
SOYBEANS	1	1	1	3	0	6
	16.67	16.67	16.67	50.00	0.00	100.00
SUGARBEETS	0	0	0	3	3	6
	0.00	0.00	0.00	50.00	50.00	100.00
Total	5	7	5	8	5	30
Percent	16.67	23.33	16.67	26.67	16.67	100.00
Priors	0.2000	0.2000	0.2000	0.2000	0.2000	

Error Count Estimates for CROP:

	CLOVER	CORN	COTTON	SOYBEANS	SUGARBEETS	Total
Rate	0.3333	0.0000	0.5000	0.5000	0.5000	0.3667
Priors	0.2000	0.2000	0.2000	0.2000	0.2000	

30 Observations　29 DF Total
4 Variables　25 DF Within Classes
5 Classes　4 DF Between Classes

Class Level Information

CROP	Frequency	Weight	Proportion	Prior Probability
CLOVER	6	6.0000	0.200000	0.200000

CORN	6	6.0000	0.200000	0.200000
COTTON	6	6.0000	0.200000	0.200000
SOYBEANS	6	6.0000	0.200000	0.200000
SUGARBEETS	6	6.0000	0.200000	0.200000

DISCRIMINANT ANALYSIS POOLED COVARIANCE MATRIX INFORMATION

Covariance Matrix Rank	Natural Log of Determinant of the Covariance Matrix
4	20.2020782

Pairwise Generalized Squared Distances Between Groups

$$D^2(i|j) = (\bar{X}_i - \bar{X}_j)' \, COV^{-1} \, (\bar{X}_i - \bar{X}_j)$$

Generalized Squared Distance to CROP

FromCROP	CLOVER	CORN	COTTON	SOYBEANS	SUGARBEETS
CLOVER	0	9.67112	3.27636	5.51437	2.66130
CORN	9.67112	0	2.14355	1.04981	5.19053
COTTON	3.27636	2.14355	0	1.10269	1.89695
SOYBEANS	5.51437	1.04981	1.10269	0	1.70143
SUGARBEETS	2.66130	5.19053	1.89695	1.70143	0

DISCRIMINANT ANALYSIS LINEAR DISCRIMINANT FUNCTION

$$Constant = -.5 \, \bar{X}_j' \, COV^{-1} \, \bar{X}_j \qquad Coefficient\ Vector = COV^{-1} \, \bar{X}_j$$

	CLOVER	CORN	COTTON	SOYBEANS	SUGARBEETS
CONSTANT	-13.40052	-5.67592	-9.07274	-5.56509	-8.36779
X1	0.10408	-0.07978	-0.00248	-0.01844	0.05069
X2	0.37732	0.15532	0.23822	0.21107	0.28186
X3	0.01337	0.18755	0.13365	0.08930	-0.00669
X4	0.11869	0.12777	0.15391	0.12540	0.15386

Classification Summary for Calibration Data: WORK.CROPCAL

Resubstitution Summary using Linear Discriminant Function

Generalized Squared Distance Function:

$$D_j^2(X) = (X-\bar{X}_j)' \, COV^{-1} \, (X-\bar{X}_j)$$

Posterior Probability of Membership in each CROP:

$$Pr(j|X) = \exp(-.5\, D_j^2(X)) \,/\, SUM_k \exp(-.5\, D_k^2(X))$$

Classification Summary for Calibration Data: WORK.CROPCAL

Resubstitution Summary using Linear Discriminant Function

Number of Observations and Percent Classified into CROP:

From CROP	CLOVER	CORN	COTTON	SOYBEANS	SUGARBEETS
CLOVER	4	0	1	0	1
	66.67	0.00	16.67	0.00	16.67
CORN	0	6	0	0	0
	0.00	100.00	0.00	0.00	0.00
COTTON	0	0	3	2	1
	0.00	0.00	50.00	33.33	16.67
SOYBEANS	1	1	1	3	0
	16.67	16.67	16.67	50.00	0.00
SUGARBEETS	0	0	0	3	3
	0.00	0.00	0.00	50.00	50.00
Total	5	7	5	8	5
Percent	16.67	23.33	16.67	26.67	16.67
Priors	0.2000	0.2000	0.2000	0.2000	0.2000

Error Count Estimates for CROP:

	CLOVER	CORN	COTTON	SOYBEANS	SUGARBEETS	Total
Rate	0.3333	0.0000	0.5000	0.5000	0.5000	0.3667
Priors	0.2000	0.2000	0.2000	0.2000	0.2000	

Classification Results using Linear Discriminant Function

Generalized Squared Distance Function:

$$D_j^2(X) = (X-\bar{X}_j)' \, COV^{-1} \, (X-\bar{X}_j)$$

Posterior Probability of Membership in each CROP:

$$Pr(j|X) = \exp(-.5\, D_j^2(X)) \,/\, SUM_k \exp(-.5\, D_k^2(X))$$

Classification Results using Linear Discriminant Function

XVALUES	From CROP	Classified into CROP		Posterior Probability of Membership in CROP: CLOVER	CORN	COTTON
				SOYBEANS	SUGARBEETS	
16 27 31 33	CORN	CORN		0.0052	0.4589	0.2211
				0.2710	0.0438	
21 25 23 24	SOYBEANS	SOYBEANS		0.0200	0.2530	0.1849
				0.4072	0.1348	
29 24 26 28	COTTON	SOYBEANS	*	0.0315	0.2001	0.2360
				0.3670	0.1653	
54 23 21 54	SUGARBEETS	SUGARBEETS		0.1521	0.0064	0.1248
				0.0796	0.6371	
32 32 62 16	CLOVER	CORN	*	0.0024	0.7084	0.2146
				0.0730	0.0016	

* Misclassified observation

REMOTE SENSING DATA ON FIVE CROPS

CLASSIFICATION OF TEST DATA

DISCRIMINANT ANALYSIS CLASSIFICATION SUMMARY FOR TEST DATA: WORK.TEST

Classification Summary using Linear Discriminant Function

Generalized Squared Distance Function:

$$D_j(X) = (X-\bar{X}_j)' \, COV^{-1} \, (X-\bar{X}_j)$$

Posterior Probability of Membership in each CROP:

$$Pr(j|X) = \exp(-.5\, D_j(X)) \,/\, SUM_k \exp(-.5\, D_k(X))$$

Number of Observations and Percent Classified into CROP:

From CROP	CLOVER	CORN	COTTON	SOYBEANS	SUGARBEETS	Total
CLOVER	0	1	0	0	0	1
	0.00	100.00	0.00	0.00	0.00	100.00
CORN	0	1	0	0	0	1
	0.00	100.00	0.00	0.00	0.00	100.00
COTTON	0	0	0	1	0	1
	0.00	0.00	0.00	100.00	0.00	100.00
SOYBEANS	0	0	0	1	0	1
	0.00	0.00	0.00	100.00	0.00	100.00

SUGARBEETS	0	0	0	0	1	1
	0.00	0.00	0.00	0.00	100.00	100.00
Total	0	2	0	2	1	5
Percent	0.00	40.00	0.00	40.00	20.00	100.00
Priors	0.2000	0.2000	0.2000	0.2000	0.2000	

Error Count Estimates for CROP:

	CLOVER	CORN	COTTON	SOYBEANS	SUGARBEETS	Total
Rate	1.0000	0.0000	1.0000	0.0000	0.0000	0.4000
Priors	0.2000	0.2000	0.2000	0.2000	0.2000	

第 40 章　典型鉴别分析：统计程序 PROC CANDISC

40.1　专有名词简介

■ 连续变量与名义变量

典型判别分析 (Canonical Discriminant Analysis) 是典型相关的特例。在此特例中，一组变量是连续变量 (如：身高，体重，智商等)，另一组变量是名义变量 (Nominal Variable)，其值不具任何实质的意义 (如：男=1，女=2)。

■ 典型变量、典型系数、典型相关系数

在连续变量与名义变量之间，我们试着找出连续变量之间的各种线性组合 (这些线性组合就称为典型变量)，使其能尽量区分名义变量类别之间的不同。构成典型变量的系数称为典型系数或典型加权值。典型变量与名义变量之间的相关系数称为典型相关系数。至于典型变量，按其前后形成的顺序，有第一，第二等分别。第一典型变量较第二典型变量重要 (或说第一典型变量与各类别之间的典型相关系数最高)。第二典型变量较第三典型变量重要，以此类推。每一典型变量都会经过 F 检定，以决定它是否显著。典型变量之间是线性独立的，但典型系数之间则不一定是线性独立。

40.2　PROC CANDISC 程序概述

CANDISC 程序执行典型判别分析，计算玛氏距离的平方 [Mahalanobis Distance，亦即经过标准化后的欧几里得距离 (Euclidean Distance)]，并且执行单变量与多变量的一因子变异数分析。输出的资料文件包括 (未经过) 标准化的典型系数、典型相关系数、典型变量在各类别内的平均数、以及典型变量值等。

CANDISC 程序的分析步骤如下：

一、将连续变量标准化，使其在名义变量下各组内共变异数的平均矩阵是一个单位矩阵。

二、计算标准化后各变量的组内平均数。

三、将这些平均数乘以该组观察体的总数，这一个步骤称为「加权调整」。然后针对这些经过加权后的平均数做主成份分析。

四、将所得的主成份再转变回原变量的单位，就得到典型变量。

40.3　如何撰写 PROC CANDISC 程序

PROC CANDISC 含六道指令，它们的格式如下：

PROC CANDISC	选项串；
VAR	变量名称串；
CLASS	变量名称；
FREQ	变量名称；
WEIGHT	变量名称；
BY	变量名称串；

读者必须选用 PROC CANDISC 与 CLASS，其余则是附加的指令。

指令 #1　PROC CANDISC 选项串；

PROC CANDISC 的选项可分下列四大类讨论：第一类选项与资料文件的界定有关，第二类选项与典型变量有关，第三类选项可用来控制报表的打印，第四类选项界定其它有关事宜。

第一类选项　下列三个选项与资料文件的界定有关：

(1) 输入资料文件名称

为输入资料文件命名。这个资料文件可以含原始的变量数据，它也可以是其它程序的输出资料文件。例如：一个含 BY 指令的 CORR 程序的输出资料文件 (亦即一个相关系数矩阵，其 TYPE=CORR)，或是前一个 CANDISC 程序的 OUTSTAT 输出资料文件 (亦即一个共变异数矩阵，其 TYPE=COV)。此外，TYPE=SSCP 或 CSSCP 的资料文件也可被 CANDISC 程序接受。若省略此选项，则 SAS 会自动找出在此程序之前最后形成的 SAS 资料文件，对它执行典型判别分析。

(2) OUT=第一个输出资料文件名称

这一个输出资料文件包括原输入资料文件的数据以及典型变量值 (Canonical Variable Score)。

(3) OUTSTAT=第二个输出资料文件名称

这一个输出资料文件包括典型判别分析的各式结果。请参考下页的表以了解各结果的代号与定义：

代号 (_TYPE_=)	定　　义
CSSCP	整个资料文件内经过平均数矫正过后的 SSCP 矩阵 (亦即平方和与内乘积的矩阵) 或各组内的 SSCP 矩阵 (与 CLASS 指令合用)
BSSCP	组间的 SSCP 矩阵
PSSCP	各组内 SSCP 矩阵的平均
COV	整个资料文件的共变异数矩阵
BCOV	组间的共变异数矩阵
PCOV	各组内共变异数矩阵的平均
CORR	整个资料文件的相关系数矩阵
BCORR	组间的相关系数矩阵
PCORR	各组内相关系数矩阵的平均

MEAN	若此程序中包含 CLASS 指令，则 MEAN 代表各组在某一变量上的平均数。若此程序中包含 PROB 指令，则 MEAN 代表整个资料文件在某一变量上的平均数。
STD	整个资料文件的标准差
BSTD	组间的标准差
PSTD	各组内标准差的平均
RSQUARED	相关平方
N	若此程序中包含 CLASS 指令，则 N 代表各组内观察体的个数。若此程序中不包含 CLASS 指令，则 N 代表整个资料文件内观察体的总数。
STDMEAN	整个资料文件在标准化变量上的平均数
PSTDMEAN	各组内在标准化变量上的平均数
CANCORR	典型相关系数
STRUCTUR	典型结构
SCORE	标准化后的典型系数
RAWSCORE	未经标准化的典型系数
CANMEAN	组内典型变量的平均数
SUMWGT	是 WEIGHT 指定的结果，代表各组内加权值的总和 (与 CLASS 指令合用) 或代表整个资料文件内加权值的总和 (若省略 CLASS 指令)。

第二类选项　下列两个选项与典型变量有关：

(1) NCAN=整数

告诉 PROC CANDISC 要找出几个典型变量。内设值是输入资料文件中连续变量的数目。如果读者定 NCAN=0，则 PROC CANDISC 只印出典型相关系数。如果读者指派一个负整数，则 PROC CANDISC 会抑止整个典型相关分析的进行。

(2) PREFIX=典型变量的名字

如果读者订 PREFIX=ABC，则第一，第二...等典型变量会被称为 ABC1，ABC2...。内设值是 CAN。典型变量的名字 (包括名字及编号) 不可超过八个字母。

第三类选项　下列十九个选项可用来控制报表的打印：

(1) SIMPLE

以各组或整个样本为单位，计算且打印各 (连续) 变量的描述性统计值。

(2) ANOVA

印出单变量变异数分析的 F 检定值及其显著度。

(3) STDMEAN

印出经过标准化变量的平均数。

(4) TCORR

印出整个资料文件的相关系数矩阵。

(5) PCORR

印出各组内相关系数矩阵的平均。

(6) BCORR

印出组间相关系数矩阵。

(7) TCOV

印出整个资料文件的共变异数矩阵。

(8) PCOV

印出各组内共变异数矩阵的平均。

(9) BCOV

印出组间共变异数矩阵。

(10) TSSCP

印出整个资料文件的 SSCP 矩阵 (亦即平方和与内乘积的矩阵)。

(11) PSSCP

印出各组内 SSCP 矩阵的平均。

(12) BSSCP

印出组间的 SSCP 矩阵。

(13) DISTANCE

印出任何两组之间的玛氏距离平方 (亦即经过标准化的欧几里得距离平方)。

(14) ALL

印出上述所有的统计值。

(15) SHORT

只印出典型相关系数表及多变量统计检定的值。

(16) NOPRINT

不印出分析的结果。

(17) WCORR

印出各组内相关系数矩阵。

(18) WCOV

印出各组内共变异数矩阵。

(19) WSSCP

印出各组内经平均数矫正后的 SSCP 矩阵。

下列一个选项界定其它有关事宜：

(1)　SINGULAR(或 SIN)=　小的正小数(P)

检查各矩阵是否为满秩矩阵。这个标准是用来检定整个资料文件的相关系数矩阵。此值必须小于 1 且大于 0 。内设值是 10 的 -8 次方。若某一变量与其它变量的复相关平方大于或等于 1-P，则此变量将自动从典型相关分析中剔出，该变量的典型系数便等于 0。

指令 #2　VAR 变量名称串；

指示 PROC CANDISC 应该对哪些数值变量执行典型的判别分析。若省略此指令，则 PROCCANDISC 自动对在本程序内其它指令未曾提到的所有数值变量执行典型判别分析。

指令 #3　CLASS 变量名称；

此指令界定一个分类变量。这个变量可以是数值的 (如：男=1，女=2) ，也可以是文字的 (如：男=M，女=F)。

指令 #4　FREQ 变量名称；

此变量的值代表资料文件内各观察体重复出现的次数。

指令 #5　WEIGHT 变量名称；

当输入资料文件内各观察体的变异数不等时，读者可依据观察体变异数的倒数指派不同的加权值以区分各观察体的重要性，这些加权值就是 WEIGHT 变量的值。

指令 #6　BY 变量名称串；

SAS 依据此指令所列举的变量将资料文件分成几个小的资料文件，然后对每一个小的资料文件分别执行典型判别分析。当读者选用此指令时，资料文件内的数据必须先依照 BY 变量串的值做由小到大的重新排列，这个步骤可藉 PROC SORT 达成。

40.4　范　　例

例一：费氏紫罗兰的典型判别分析

有关这个资料文件 (IRIS) 的详细介绍，请参阅第 38 章分类鉴别分析的例一。分析的过程里，数据首先以典型判别分析处理，找出两个典型变量 (分别命名为 CAN1，CAN2)。接下来，一百五十株紫罗兰在这两个典型变量上的值以 PROC PLOT 绘制成一个平面图。图上，紫罗兰的识别代号是它们原来的属种 (1=SETOSA，2=VERSICOLOR，3=VIRGINICA)。根据图形的显示，我们可以下结论说：第一种紫罗兰在 CAN1 上的值最低，第二种居中，第三种紫罗兰在 CAN1 上的值最高。在鉴别三种紫罗兰的过程中，第一典型变量 (CAN1) 远比第二典型变量 (CAN2) 更有效!

程　序

```
DATA IRIS;
    TITLE 'FISHER (1936) IRIS DATA';
    INPUT SEPALLEN SEPALWID PETALLEN PETALWID SPEC_NO @@;
    IF SPEC_NO=1 THEN SPECIES='SETOSA';
    ELSE IF SPEC_NO=2 THEN SPECIES='VERSICOLOR';
    ELSE SPECIES='VIRGINICA';
    LABEL SEPALLEN='SEPAL LENGTH IN MM.'
          SEPALWID='SEPAL WIDTH  IN MM.'
          PETALLEN='PETAL LENGTH IN MM.'
          PETALWID='PETAL WIDTH  IN MM.';
```

```
        CARDS;
        (原数据请见第 39 章例一)
    ;
    PROC CANDISC ALL  OUT=DISC;
        CLASS SPECIES;
        VAR SEPALLEN SEPALWID PETALLEN PETALWID;
    PROC PLOT VPERCENT=300;
        PLOT CAN2*CAN1=SPEC_NO / VAXIS=-3 TO 3 BY 1 VSPACE=7;
        TITLE2 'PLOT OF CANONICAL DISCRIMINANT FUNCTIONS';
    RUN;
```

结　果

报表 40.1　费氏紫罗兰的典型判别分析

FISHER (1936) IRIS DATA

CANONICAL DISCRIMINANT ANALYSIS

150 Observations　　149 DF Total

4 Variables　　147 DF Within Classes

3 Classes　　2 DF Between Classes

Class Level Information

SPECIES	Frequency	Weight	Proportion
SETOSA	50	50.0000	0.333333
VERSIC	50	50.0000	0.333333
VIRGIN	50	50.0000	0.333333

SPECIES = SETOSA

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID	
SEPALLEN	608.8200000	486.1600000	80.1400000	50.6200000	SEPAL LENGTH IN MM.
SEPALWID	486.1600000	704.0800000	57.3200000	45.5600000	SEPAL WIDTH IN MM.
PETALLEN	80.1400000	57.3200000	147.7800000	29.7400000	PETAL LENGTH IN MM.
PETALWID	50.6200000	45.5600000	29.7400000	54.4200000	PETAL WIDTH IN MM.

SPECIES = VERSIC

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID	
SEPALLEN	1305.520000	417.400000	896.200000	273.320000	SEPAL LENGTH IN MM.
SEPALWID	417.400000	482.500000	405.000000	201.900000	SEPAL WIDTH IN MM.
PETALLEN	896.200000	405.000000	1082.000000	358.200000	PETAL LENGTH IN MM.
PETALWID	273.320000	201.900000	358.200000	191.620000	PETAL WIDTH IN MM.

SPECIES = VIRGIN

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID	
SEPALLEN	1981.280000	459.440000	1486.120000	240.560000	SEPAL LENGTH IN MM.
SEPALWID	459.440000	509.620000	349.760000	233.380000	SEPAL WIDTH IN MM.
PETALLEN	1486.120000	349.760000	1492.480000	239.240000	PETAL LENGTH IN MM.
PETALWID	240.560000	233.380000	239.240000	369.620000	PETAL WIDTH IN MM.

CANONICAL DISCRIMINANT ANALYSIS

Pooled Within-Class SSCP Matrix

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID	
SEPALLEN	3895.620000	1363.000000	2462.460000	564.500000	SEPAL LENGTH IN MM.
SEPALWID	1363.000000	1696.200000	812.080000	480.840000	SEPAL WIDTH IN MM.
PETALLEN	2462.460000	812.080000	2722.260000	627.180000	PETAL LENGTH IN MM.
PETALWID	564.500000	480.840000	627.180000	615.660000	PETAL WIDTH IN MM.

Between-Class SSCP Matrix

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID	
SEPALLEN	6321.21333	-1995.26667	16524.84000	7127.93333	SEPAL LENGTH IN MM.
SEPALWID	-1995.26667	1134.49333	-5723.96000	-2293.26667	SEPAL WIDTH IN MM.
PETALLEN	16524.84000	-5723.96000	43710.28000	18677.40000	PETAL LENGTH IN MM.
PETALWID	7127.93333	-2293.26667	18677.40000	8041.33333	PETAL WIDTH IN MM.

Total-Sample SSCP Matrix

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID	
SEPALLEN	10216.83333	-632.26667	18987.30000	7692.43333	SEPAL LENGTH IN MM.
SEPALWID	-632.26667	2830.69333	-4911.88000	-1812.42667	SEPAL WIDTH IN MM.
PETALLEN	18987.30000	-4911.88000	46432.54000	19304.58000	PETAL LENGTH IN MM.
PETALWID	7692.43333	-1812.42667	19304.58000	8656.99333	PETAL WIDTH IN MM.

CANONICAL DISCRIMINANT ANALYSIS　WITHIN-CLASS COVARIANCE MATRICES

SPECIES = SETOSA　DF = 49

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID	
SEPALLEN	12.42489796	9.92163265	1.63551020	1.03306122	SEPAL LENGTH IN MM.
SEPALWID	9.92163265	14.36897959	1.16979592	0.92979592	SEPAL WIDTH IN MM.
PETALLEN	1.63551020	1.16979592	3.01591837	0.60693878	PETAL LENGTH IN MM.
PETALWID	1.03306122	0.92979592	0.60693878	1.11061224	PETAL WIDTH IN MM.

SPECIES = VERSIC　DF = 49

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID	
SEPALLEN	26.64326531	8.51836735	18.28979592	5.57795918	SEPAL LENGTH IN MM.

SEPALWID	8.51836735	9.84693878	8.26530612	4.12040816	SEPAL WIDTH IN MM.
PETALLEN	18.28979592	8.26530612	22.08163265	7.31020408	PETAL LENGTH IN MM.
PETALWID	5.57795918	4.12040816	7.31020408	3.91061224	PETAL WIDTH IN MM.

SPECIES = VIRGIN　DF = 49

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID	
SEPALLEN	40.43428571	9.37632653	30.32897959	4.90938776	SEPAL LENGTH IN MM.
SEPALWID	9.37632653	10.40040816	7.13795918	4.76285714	SEPAL WIDTH IN MM.
PETALLEN	30.32897959	7.13795918	30.45877551	4.88244898	PETAL LENGTH IN MM.
PETALWID	4.90938776	4.76285714	4.88244898	7.54326531	PETAL WIDTH IN MM.

Pooled Within-Class Covariance Matrix　DF = 147

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID	
SEPALLEN	26.50081633	9.27210884	16.75142857	3.84013605	SEPAL LENGTH IN MM.
SEPALWID	9.27210884	11.53877551	5.52435374	3.27102041	SEPAL WIDTH IN MM.
PETALLEN	16.75142857	5.52435374	18.51877551	4.26653061	PETAL LENGTH IN MM.
PETALWID	3.84013605	3.27102041	4.26653061	4.18816327	PETAL WIDTH IN MM.

Between-Class Covariance Matrix　DF = 2

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID	
SEPALLEN	63.2121333	-19.9526667	165.2484000	71.2793333	SEPAL LENGTH IN MM.
SEPALWID	-19.9526667	11.3449333	-57.2396000	-22.9326667	SEPAL WIDTH IN MM.
PETALLEN	165.2484000	-57.2396000	437.1028000	186.7740000	PETAL LENGTH IN MM.
PETALWID	71.2793333	-22.9326667	186.7740000	80.4133333	PETAL WIDTH IN MM.

Total-Sample Covariance Matrix　DF = 149

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID	
SEPALLEN	68.5693512	-4.2434004	127.4315436	51.6270694	SEPAL LENGTH IN MM.
SEPALWID	-4.2434004	18.9979418	-32.9656376	-12.1639374	SEPAL WIDTH IN MM.
PETALLEN	127.4315436	-32.9656376	311.6277852	129.5609396	PETAL LENGTH IN MM.
PETALWID	51.6270694	-12.1639374	129.5609396	58.1006264	PETAL WIDTH IN MM.

Within-Class Correlation Coefficients / Prob > |R|

SPECIES = SETOSA

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.00000	0.74255	0.26718	0.27810
SEPAL LENGTH IN MM.	0.0	0.0001	0.0607	0.0505
SEPALWID	0.74255	1.00000	0.17770	0.23275
SEPAL WIDTH IN MM.	0.0001	0.0	0.2170	0.1038

PETALLEN	0.26718	0.17770	1.00000	0.33163
PETAL LENGTH IN MM.	0.0607	0.2170	0.0	0.0186
PETALWID	0.27810	0.23275	0.33163	1.00000
PETAL WIDTH IN MM.	0.0505	0.1038	0.0186	0.0

SPECIES = VERSIC

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.00000	0.52591	0.75405	0.54646
SEPAL LENGTH IN MM.	0.0	0.0001	0.0001	0.0001
SEPALWID	0.52591	1.00000	0.56052	0.66400
SEPAL WIDTH IN MM.	0.0001	0.0	0.0001	0.0001
PETALLEN	0.75405	0.56052	1.00000	0.78667
PETAL LENGTH IN MM.	0.0001	0.0001	0.0	0.0001
PETALWID	0.54646	0.66400	0.78667	1.00000
PETAL WIDTH IN MM.	0.0001	0.0001	0.0001	0.0

SPECIES = VIRGIN

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.00000	0.45723	0.86422	0.28111
SEPAL LENGTH IN MM.	0.0	0.0008	0.0001	0.0480
SEPALWID	0.45723	1.00000	0.40104	0.53773
SEPAL WIDTH IN MM.	0.0008	0.0	0.0039	0.0001
PETALLEN	0.86422	0.40104	1.00000	0.32211
PETAL LENGTH IN MM.	0.0001	0.0039	0.0	0.0225
PETALWID	0.28111	0.53773	0.32211	1.00000
PETAL WIDTH IN MM.	0.0480	0.0001	0.0225	0.0

Pooled Within-Class Correlation Coefficients / Prob > |R|

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.00000	0.53024	0.75616	0.36451
SEPAL LENGTH IN MM.	0.0	0.0001	0.0001	0.0001
SEPALWID	0.53024	1.00000	0.37792	0.47053
SEPAL WIDTH IN MM.	0.0001	0.0	0.0001	0.0001
PETALLEN	0.75616	0.37792	1.00000	0.48446
PETAL LENGTH IN MM.	0.0001	0.0001	0.0	0.0001
PETALWID	0.36451	0.47053	0.48446	1.00000
PETAL WIDTH IN MM.	0.0001	0.0001	0.0001	0.0

Between-Class Correlation Coefficients / Prob > |R|

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID

SEPALLEN	1.00000	-0.74507	0.99413	0.99977
SEPAL LENGTH IN MM.	0.0	0.4648	0.0690	0.0137
SEPALWID	-0.74507	1.00000	-0.81284	-0.75926
SEPAL WIDTH IN MM.	0.4648	0.0	0.3958	0.4511
PETALLEN	0.99413	-0.81284	1.00000	0.99623
PETAL LENGTH IN MM.	0.0690	0.3958	0.0	0.0553
PETALWID	0.99977	-0.75926	0.99623	1.00000
PETAL WIDTH IN MM.	0.0137	0.4511	0.0553	0.0

Total-Sample Correlation Coefficients / Prob > |R|

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.00000	-0.11757	0.87175	0.81794
SEPAL LENGTH IN MM.	0.0	0.1519	0.0001	0.0001
SEPALWID	-0.11757	1.00000	-0.42844	-0.36613
SEPAL WIDTH IN MM.	0.1519	0.0	0.0001	0.0001
PETALLEN	0.87175	-0.42844	1.00000	0.96287
PETAL LENGTH IN MM.	0.0001	0.0001	0.0	0.0001
PETALWID	0.81794	-0.36613	0.96287	1.00000
PETAL WIDTH IN MM.	0.0001	0.0001	0.0001	0.0

SIMPLE STATISTICS　　Total-Sample

Variable	N	Sum	Mean	Variance	Std Dev	Label
SEPALLEN	150	8765	58.43333	68.56935	8.28066	SEPAL LENGTH IN MM.
SEPALWID	150	4586	30.57333	18.99794	4.35866	SEPAL WIDTH IN MM.
PETALLEN	150	5637	37.58000	311.62779	17.65298	PETAL LENGTH IN MM.
PETALWID	150	1799	11.99333	58.10063	7.62238	PETAL WIDTH IN MM.

SPECIES = SETOSA

Variable	N	Sum	Mean	Variance	Std Dev	Label
SEPALLEN	50	2503	50.06000	12.42490	3.52490	SEPAL LENGTH IN MM.
SEPALWID	50	1714	34.28000	14.36898	3.79064	SEPAL WIDTH IN MM.
PETALLEN	50	731.00000	14.62000	3.01592	1.73664	PETAL LENGTH IN MM.
PETALWID	50	123.00000	2.46000	1.11061	1.05386	PETAL WIDTH IN MM.

SPECIES = VERSIC

Variable	N	Sum	Mean	Variance	Std Dev	Label
SEPALLEN	50	2968	59.36000	26.64327	5.16171	SEPAL LENGTH IN MM.
SEPALWID	50	1385	27.70000	9.84694	3.13798	SEPAL WIDTH IN MM.
PETALLEN	50	2130	42.60000	22.08163	4.69911	PETAL LENGTH IN MM.
PETALWID	50	663.00000	13.26000	3.91061	1.97753	PETAL WIDTH IN MM.

SPECIES = VIRGIN

Variable	N	Sum	Mean	Variance	Std Dev	Label
SEPALLEN	50	3294	65.88000	40.43429	6.35880	SEPAL LENGTH IN MM.
SEPALWID	50	1487	29.74000	10.40041	3.22497	SEPAL WIDTH IN MM.
PETALLEN	50	2776	55.52000	30.45878	5.51895	PETAL LENGTH IN MM.
PETALWID	50	1013	20.26000	7.54327	2.74650	PETAL WIDTH IN MM.

	Total-Sample Standardized Class Means			Pooled Within-Class Standardized Class Means		
Variable	SETOSA	VERSIC	VIRGIN	SETOSA	VERSIC	VIRGIN
SEPALLEN	-1.011191383	0.111907327	0.899284057	-1.626555005	0.180008874	1.446546131
SEPALWID	0.850413715	-0.659223581	-0.191190134	1.091198274	-0.845874921	-0.245323353
PETALLEN	-1.300630090	0.284371213	1.016258877	-5.335384837	1.166534490	4.168850347
PETALWID	-1.250703517	0.166177390	1.084526127	-4.658359237	0.618942836	4.039416401

Canonical Discriminant Analysis

$$D^2(i|j) = (\bar{X}_i - \bar{X}_j)' \, COV^{-1} \, (\bar{X}_i - \bar{X}_j)$$

Pairwise Squared Distances Between Groups

Squared Distance to SPECIES

From SPECIES	SETOSA	VERSIC	VIRGIN
SETOSA	0	89.86419	179.38471
VERSIC	89.86419	0	17.20107
VIRGIN	179.38471	17.20107	0

	F Statistics, NDF=4, DDF=144 for Squared Distance to SPECIES			Prob > Mahalanobis Distance for Squared Distance to SPECIES		
From SPECIES	SETOSA	VERSIC	VIRGIN	SETOSA	VERSIC	VIRGIN
SETOSA	0	550.18889	1098	1.0000	0.0001	0.0001
VERSIC	550.18889	0	105.31265	0.0001	1.0000	0.0001
VIRGIN	1098	105.31265	0	0.0001	0.0001	1.0000

Univariate Test Statistics

F Statistics, Num DF= 2 Den DF= 147

Variable	Total STD	Pooled STD	Between STD	R-Squared	RSQ/ (1-RSQ)	F	Pr > F
SEPALLEN	8.2807	5.1479	7.9506	0.618706	1.6226	119.2645	0.0001
SEPALWID	4.3587	3.3969	3.3682	0.400783	0.6688	49.1600	0.0001
PETALLEN	17.6530	4.3033	20.9070	0.941372	16.0566	1180.1612	0.0001
PETALWID	7.6224	2.0465	8.9673	0.928883	13.0613	960.0071	0.0001

Average R-Squared: Unweighted = 0.7224358　　Weighted by Variance = 0.8689444

Multivariate Statistics and F Approximations

S=2　M=0.5　N=71

Statistic	Value	F	Num DF	Den DF	Pr > F
Wilks' Lambda	0.02343863	199.1453	8	288	0.0001
Pillai's Trace	1.19189883	53.4665	8	290	0.0001
Hotelling-Lawley Trace	32.47732024	580.5321	8	286	0.0001
Roy's Greatest Root	32.19192920	1166.957	4	145	0.0001

NOTE: F Statistic for Roy's Greatest Root is an upper bound.

NOTE: F Statistic for Wilks' Lambda is exact.

	Canonical Correlation	Adjusted Canonical Correlation	Approx Standard Error	Squared Canonical Correlation
1	0.984821	0.984508	0.002468	0.969872
2	0.471197	0.461445	0.063734	0.222027

Eigenvalues of INV(E)*H
= CanRsq/(1-CanRsq)

	Eigenvalue	Difference	Proportion	Cumulative
1	32.1919	31.9065	0.9912	0.9912
2	0.2854	.	0.0088	1.0000

Test of H0: The canonical correlations in the current row and all that follow are zero

	Likelihood Ratio	Approx F	Num DF	Den DF	Pr > F
1	0.02343863	199.1453	8	288	0.0001
2	0.77797337	13.7939	3	145	0.0001

Total Canonical Structure

	CAN1	CAN2	
SEPALLEN	0.791888	0.217593	SEPAL LENGTH IN MM.
SEPALWID	-0.530759	0.757989	SEPAL WIDTH IN MM.
PETALLEN	0.984951	0.046037	PETAL LENGTH IN MM.
PETALWID	0.972812	0.222902	PETAL WIDTH IN MM.

Between Canonical Structure

	CAN1	CAN2	
SEPALLEN	0.991468	0.130348	SEPAL LENGTH IN MM.
SEPALWID	-0.825658	0.564171	SEPAL WIDTH IN MM.
PETALLEN	0.999750	0.022358	PETAL LENGTH IN MM.
PETALWID	0.994044	0.108977	PETAL WIDTH IN MM.

Pooled Within Canonical Structure

	CAN1	CAN2	
SEPALLEN	0.222596	0.310812	SEPAL LENGTH IN MM.
SEPALWID	-0.119012	0.863681	SEPAL WIDTH IN MM.
PETALLEN	0.706065	0.167701	PETAL LENGTH IN MM.
PETALWID	0.633178	0.737242	PETAL WIDTH IN MM.

Pooled Within-Class Standardized Canonical Coefficients

	CAN1	CAN2	
SEPALLEN	-.4269548486	0.0124075316	SEPAL LENGTH IN MM.
SEPALWID	-.5212416758	0.7352613085	SEPAL WIDTH IN MM.
PETALLEN	0.9472572487	-.4010378190	PETAL LENGTH IN MM.
PETALWID	0.5751607719	0.5810398645	PETAL WIDTH IN MM.

Raw Canonical Coefficients

	CAN1	CAN2	
SEPALLEN	-.0829377642	0.0024102149	SEPAL LENGTH IN MM.
SEPALWID	-.1534473068	0.2164521235	SEPAL WIDTH IN MM.
PETALLEN	0.2201211656	-.0931921210	PETAL LENGTH IN MM.
PETALWID	0.2810460309	0.2839187853	PETAL WIDTH IN MM.

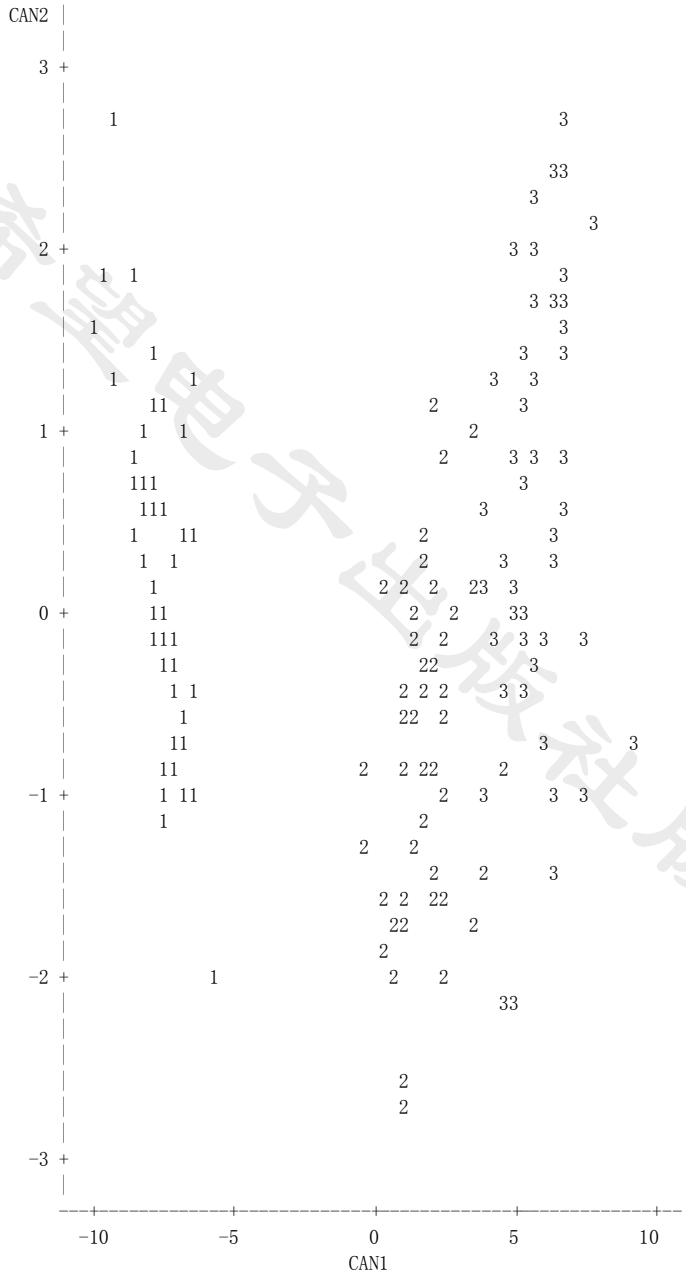
Class Means on Canonical Variables

SPECIES	CAN1	CAN2
SETOSA	-7.607599927	0.215133017
VERSIC	1.825049490	-0.727899622
VIRGIN	5.782550437	0.512766605

FISHER (1936) IRIS DATA

PLOT OF CANONICAL DISCRIMINANT FUNCTIONS

Plot of CAN2*CAN1. Symbol is value of SPEC_NO.



NOTE: 18 obs hidden.

第 41 章　回归鉴别分析：统计程序 PROC STEPDISC

41.1　PROC STEPDISC 程序中三种挑选变量的方法

STEPDISC 程序对输入资料文件执行回归鉴别分析，选出一组最适合区分类别的数值变量，然后从这些变量中导出一个鉴别函数 (亦即这些变量的线性组合)。根据第 18 章所介绍的回归分析，有三种方法可用来挑选构成鉴别函数 (亦称鉴别模型) 的变量。这三种方法是：向淘汰法及逐步排除法。有兴趣的读者可参阅克拉卡 (Klecka，1980) 的著作。下面简单地介绍这三种方法的原理：

■ 顺向选择法

根据这个方法，鉴别模型首先不含任何变量。 在每一次选择的步骤中，PROC STEPDISC 选出一个变量， 其对鉴别函数的鉴别力贡献最大。 这个贡献以威尔克斯 (Wilk's) 的 Lambda 衡量之。这个 Lambda 其实也就是可能比的指标 (Likelihood Ratio Criterion)。如此重复，直到所余的变量都不能达到预定的 Lambda 标准时，顺向选择法便停止。

■ 反向淘汰法

反向淘汰法与上述的顺向选择法正好相反。首先，鉴别模型包含所有的变量。在每一次淘汰的步骤中，PROC STEPDISC 剔除一个变量，其对鉴别模型的鉴别力贡献最小。对鉴别力的贡献也是以威尔克斯的 Lambda 来衡量。如此重复，直到留在模型中的变量都达到预定的 Lambda 标准时，反向淘汰法便停止。

■ 逐步排除法

逐步排除法是顺向选择法与反向淘汰法的综合。首先，鉴别模型中不包含任何变量。然后采用顺向选择法，根据 Lambda 挑选变量进入鉴别模型中。在每一步骤中，已经被纳入模型的变量必须再经过反向淘汰法的审核，以决定该变量要被淘汰还是要留下。

每当某一变量加入 (或退出) 鉴别模型时，必须符合下列的条件之一：

1. 以此变量为因变量，以其它在鉴别模型内的变量串为共变量串，进行共变量分析。此共变量分析的 F 值达到显著水准；

2. 此变量与 CLASS 指令的变量之间的净相关平方 (排除掉其它已在模型内的变量串)达到预定的标准。

请读者注意，STEPDISC 程序假定各类别内的观察体来自多元常态分配。其次，当变量逐一地加入或退出模型时，每一次的 F 检定都有可能导致第一类型的错误 (Type Ⅰ Error)。犯第一类型错误的概率会累积，因此，经过一系列的 F 检定之后，读者所可能犯的第一类型错误的总概率已经远超过 5% (或 1%，视每一个 F 检定的显著度而定)。所

以，最好选择一个较保守的显著度。

41.2　如何撰写 PROC STEPDISC 程序

PROC STEPDISC 含六道指令，它们的格式如下：

```
PROC STEPDISC     选项串;
    VAR           变量名称串;
    CLASS         变量名称;
    FREQ          变量名称;
    WEIGHT        变量名称;
    BY            变量名称串;
```

读者必须选用 PROC STEPDISC 与 CLASS 两指令，其余则是附加的指令。

指令 #1　PROC STEPDISC 选项串;

PROC STEPDISC 的选项可分四大类来讨论：第一类选项与资料文件的界定有关，第二类选项界定挑选变量的方法，第三类选项界定挑选变量的有关事宜，第四类选项控制报表的打印。

第一类选项　下列选项与资料文件的界定有关：

(1) DATA= 输入资料文件名称;

为输入资料文件命名。这个资料文件可以是原始的数据，它也可以是其它程序的输出资料文件。例如：一个含 BY 指令的 CORR 程序的输出资料文件，亦即一个相关系数矩阵 (TYPE=CORR)，或前一个 CANDISC 程序的 OUTSTAT 输出资料文件 [即一个共变异数矩阵 (TYPE=COV)]。此外，TYPE=SSCP 或 TYPE=CSSCP 的资料文件也是合法的输入资料文件 。若省略此选项，则 SAS 会自动找出在此程序之前最后形成的 SAS 资料文件，对它执行回归鉴别分析。由于分析过程极为复杂，PROC STEPDISC 的结果无法纳入任何一个输出资料文件。

第二类选项　下列的选项界定挑选变量的方法：

(1) METHOD=STEPWISE (或 FORWARD 或 BACKWARD)

现将这三种方法分别解说如下：

STEPWISE (或 SW)　要求逐步排除法，这是本指令的内设值。

FORWARD (或 FW)　要求顺向选择法。

BACKWARD (或 BW) 要求反向淘汰法。

第三类选项　下列九个选项界定挑选变量的有关事宜：

(1) SLENTRY(或 SLE)=概率

在顺向选择法中，某一变量被纳入鉴别模型时所必须达到的统计显著度。此概率的值介于 0 与 1 之间，内设值是 .15。

(2) SLSTAY(或 SLS)=概率

在反向淘汰法中，某一变量留在鉴别模型内所必须达到的统计显著度。此概率的

值介于 0 与 1 之间，内设值是 .15。

(3) PR2ENTRY(或 PR2E)=正小数

在顺向选择法中，某一变量被纳入鉴别模型中所必须达到的净相关平方。

(4) PR2STAY(或 PR2S)=正小数

在反向淘汰法中，某一变量被留在鉴别模型中所必须达到的净相关平方。

(5) SINGULAR(或 SING)=极小的正小数(P)

检查是否为满秩矩阵。如果一个变量与所有其它已被纳入鉴别模型的变量之间的复相关平方超过 1-P 时，则 PROC STEPDISC 会认为此变量是多余的，而不将其纳入模型里。

内设值等于 10 的 -8 次方。

(6) INCLUDE=正整数 (如 5)

要求 VAR 指令中前几 (5) 个变量被纳入鉴别模型里，内设值是 0。

(7) MAXSTEP=正整数 (如 7)

界定挑选变量时，最多只进行几 (7) 个步骤，内设值是 VAR 指令中所列举之变量个数的两倍。

(8) START= 正整数 (如 2)

此选项要求前几个 (如 2 个) CLASS 变量必须纳入第一个回归鉴别函数据，否则任何一种选择法都无法启动。当 METHOD=FORWARD 或 STEPWISE 时，内设值是 0。若 METHOD=BACKWARD，则内设值是 VAR 变量之个数。

(9) STOP= 正整数 (如 4)

界定最后一个回归模型里所含自变量的数目，此选项只可与 METHOD=FORWARD 或 METHOD=BACKWARD 联用。

第四类选项　下列十六个选项可用来控制报表的打印：

(1) SIMPLE

印出变量的平均数与标准差。

(2) STDMEAN

印出各组内与整个资料文件在各标准化变量上的平均数。

(3) TCORR

印出整个资料文件的相关系数矩阵。

(4) WCORR

印出组内 (如：男、女两组) 的相关系数矩阵。

(5) PCORR

印出组内相关系数矩阵的平均。

(6) BCORR

印出组间之相关系数矩阵。

(7) TCOV

印出整个资料文件的共变异数矩阵。

(8) WCOV

印出组内 (如：男、女两组) 的共变异数矩阵。

(9) PCOV

印出组内共变异数矩阵的平均。

(10) BCOV

印出组间之共变异数矩阵。

(11) TSSCP

印出整个资料文件内经过平均数矫正过后的 SSCP (平方和与内乘积的矩阵)。

(12) WSSCP

印出组内 (如男、女两组) 内经过平均数矫正过后的 SSCP 矩阵。

(13) PSSCP

印出组内的 SSCP 矩阵之平均。

(14) BSSCP

印出组间的 SSCP 矩阵。

(15) ALL

促成所有选项的效果。

(16) SHORT

只印出鉴别分析的最后结果。

指令 #2　VAR 变量名称串;

告诉 PROC STEPDISC 有那些数值变量可供挑选。若省略此指令，则 PROC STEPDISC 会自动从本程序内其它指令未曾提到的所有数值变量里加以挑选。

指令 #3　CLASS 变量名称;

此指令指明一个分类变量，这个变量可以是数值的 (如：男=1，女=2)，也可以是文字的 (如：男=M，女=F)。

指令 #4　FREQ 变量名称;

FREQ 变量的值代表资料文件中各观察体重复出现的次数。

指令 #5　WEIGHT 变量名称;

当输入资料文件内各观察体的变异数不等时，读者可根据观察体变异数的倒数指派不同的加权值以区分各个体的重要性。这些加权值就是 WEIGHT 变量的值。WEIGHT 的意义与上述 FREQ 类似。然而，WEIGHT 变量的值不会影响 F 检定的自由度，但是 FREQ 变量的值则会。

指令 #6　BY 变量名称串;

SAS 依据此指令所列举的变量将资料文件分成几个小的资料文件，然后对每一个小的资料文件分别执行回归鉴别分析。当读者选用此指令时，资料文件内的数据必须先按照 BY 变量串的值做由小到大的重新排列。这个步骤可藉 PROC SORT 达成。

41.3　范　　例

例一：费氏紫罗兰的回归鉴别分析

有关这个资料文件 (IRIS) 的详细介绍，请参阅第 38 章的分类鉴别分析的例一。资料文件输入后，STEPDISC 程序对三种不同属种的紫罗兰执行回归鉴别分析。分析采逐步排除法由下列四变量中挑选，即：花萼的长与宽及花瓣的长与宽。

程　序

```
DATA IRIS;
     TITLE 'FISHER (1936) IRIS DATA';
     NPUT SEPALLEN SEPALWID PETALLEN PETALWID SPEC_NO @@;
     F SPEC_NO=1 THEN SPECIES='SETOSA';
     ELSE IF SPEC_NO=2 THEN SPECIES='VERSICOLOR';
     ELSE SPECIES='VIRGINICA';
     LABEL SEPALLEN='SEPAL LENGTH IN MM.'
        SEPALWID='SEPAL WIDTH  IN MM.'
        PETALLEN='PETAL LENGTH IN MM.'
        PETALWID='PETAL WIDTH  IN MM.';
   CARDS;
  (原始数据请见第 39 章的例一)
;
PROC STEPDISC STEPWISE STDMEAN TCORR WCORR;
     CLASS SPECIES;
     VAR SEPALLEN SEPALWID PETALLEN PETALWID;
RUN;
```

结　果

从逐步排除法的分析结果看来，最佳的鉴别模型是由所有的变量所形成的。然而，四个变量的重要性不等。依据这些变量先后进入鉴别模型的顺序，我们可将它们的重要性排列如下：

花瓣长，花萼宽，花瓣宽，花萼长。

这四个变量与原来紫罗兰属种的典型相关平方达到 0.5959。

报表 41.1　费氏紫罗兰的回归鉴别分析

FISHER (1936) IRIS DATA

STEPWISE DISCRIMINANT ANALYSIS

150 Observations	4 Variable(s) in the Analysis
3 Class Levels	0 Variable(s) will be included

The Method for Selecting Variables will be: STEPWISE

Significance Level to Enter = 0.1500

Significance Level to Stay = 0.1500

Class Level Information

SPECIES	Frequency	Weight	Proportion
SETOSA	50	50.0000	0.333333
VERSIC	50	50.0000	0.333333
VIRGIN	50	50.0000	0.333333

Within-Class Correlation Coefficients / Prob > |R|

SPECIES = SETOSA

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.00000	0.74255	0.26718	0.27810
SEPAL LENGTH IN MM.	0.0	0.0001	0.0607	0.0505
SEPALWID	0.74255	1.00000	0.17770	0.23275
SEPAL WIDTH IN MM.	0.0001	0.0	0.2170	0.1038
PETALLEN	0.26718	0.17770	1.00000	0.33163
PETAL LENGTH IN MM.	0.0607	0.2170	0.0	0.0186
PETALWID	0.27810	0.23275	0.33163	1.00000
PETAL WIDTH IN MM.	0.0505	0.1038	0.0186	0.0

Within-Class Correlation Coefficients / Prob > |R|

SPECIES = VERSIC

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.00000	0.52591	0.75405	0.54646
SEPAL LENGTH IN MM.	0.0	0.0001	0.0001	0.0001
SEPALWID	0.52591	1.00000	0.56052	0.66400
SEPAL WIDTH IN MM.	0.0001	0.0	0.0001	0.0001
PETALLEN	0.75405	0.56052	1.00000	0.78667
PETAL LENGTH IN MM.	0.0001	0.0001	0.0	0.0001
PETALWID	0.54646	0.66400	0.78667	1.00000
PETAL WIDTH IN MM.	0.0001	0.0001	0.0001	0.0

Within-Class Correlation Coefficients / Prob > |R|

SPECIES = VIRGIN

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.00000	0.45723	0.86422	0.28111
SEPAL LENGTH IN MM.	0.0	0.0008	0.0001	0.0480
SEPALWID	0.45723	1.00000	0.40104	0.53773
SEPAL WIDTH IN MM.	0.0008	0.0	0.0039	0.0001
PETALLEN	0.86422	0.40104	1.00000	0.32211
PETAL LENGTH IN MM.	0.0001	0.0039	0.0	0.0225
PETALWID	0.28111	0.53773	0.32211	1.00000
PETAL WIDTH IN MM.	0.0480	0.0001	0.0225	0.0

Total-Sample Correlation Coefficients / Prob > |R|

Variable	SEPALLEN	SEPALWID	PETALLEN	PETALWID
SEPALLEN	1.00000	-0.11757	0.87175	0.81794
SEPAL LENGTH IN MM.	0.0	0.1519	0.0001	0.0001
SEPALWID	-0.11757	1.00000	-0.42844	-0.36613
SEPAL WIDTH IN MM.	0.1519	0.0	0.0001	0.0001
PETALLEN	0.87175	-0.42844	1.00000	0.96287
PETAL LENGTH IN MM.	0.0001	0.0001	0.0	0.0001
PETALWID	0.81794	-0.36613	0.96287	1.00000
PETAL WIDTH IN MM.	0.0001	0.0001	0.0001	0.0

Total-Sample Standardized Class Means　　Pooled Within-Class Standardized Class Means

Variable	SETOSA	VERSIC	VIRGIN	SETOSA	VERSIC	VIRGIN
SEPALLEN	-1.011191383	0.111907327	0.899284057	-1.626555005	0.180008874	1.446546131
SEPALWID	0.850413715	-0.659223581	-0.191190134	1.091198274	-0.845874921	-0.245323353
PETALLEN	-1.300630090	0.284371213	1.016258877	-5.335384837	1.166534490	4.168850347
PETALWID	-1.250703517	0.166177390	1.084526127	-4.658359237	0.618942836	4.039416401

Stepwise Selection: Step 1　　Statistics for Entry, DF = 2, 147

Variable	R**2	F	Prob > F	Tolerance	Label
SEPALLEN	0.6187	119.265	0.0001	1.0000	SEPAL LENGTH IN MM.
SEPALWID	0.4008	49.160	0.0001	1.0000	SEPAL WIDTH IN MM.
PETALLEN	0.9414	1180.161	0.0001	1.0000	PETAL LENGTH IN MM.
PETALWID	0.9289	960.007	0.0001	1.0000	PETAL WIDTH IN MM.

Variable PETALLEN will be entered

Multivariate Statistics

Wilks' Lambda = 0.05862828　　F(2, 147) = 1180.161　　Prob > F = 0.0001

Pillai's Trace = 0.941372　　F(2, 147) = 1180.161　　Prob > F = 0.0001

Average Squared Canonical Correlation = 0.47068586

Stepwise Selection: Step 2 Statistics for Removal, DF = 2, 147

Variable	R**2	F	Prob > F	Label
PETALLEN	0.9414	1180.161	0.0001	PETAL LENGTH IN MM.

No variables can be removed

Stepwise Selection: Step 2 Statistics for Entry, DF = 2, 146

Variable	Partial R**2	F	Prob > F	Tolerance	Label
SEPALLEN	0.3198	34.323	0.0001	0.2400	SEPAL LENGTH IN MM.
SEPALWID	0.3709	43.035	0.0001	0.8164	SEPAL WIDTH IN MM.
PETALWID	0.2533	24.766	0.0001	0.0729	PETAL WIDTH IN MM.

Variable SEPALWID will be entered The following variable(s) have been entered:
SEPALWID PETALLEN

Multivariate Statistics

Wilks' Lambda = 0.03688411 F(4, 292) = 307.105 Prob > F = 0.0001

Pillai's Trace = 1.119908 F(4, 294) = 93.528 Prob > F = 0.0001

Average Squared Canonical Correlation = 0.55995394

Stepwise Selection: Step 3 Statistics for Removal, DF = 2, 146

Variable	Partial R**2	F	Prob > F	Label
SEPALWID	0.3709	43.035	0.0001	SEPAL WIDTH IN MM.
PETALLEN	0.9384	1112.954	0.0001	PETAL LENGTH IN MM.

No variables can be removed

Stepwise Selection: Step 3 Statistics for Entry, DF = 2, 145

Variable	Partial R**2	F	Prob > F	Tolerance	Label
SEPALLEN	0.1447	12.268	0.0001	0.1323	SEPAL LENGTH IN MM.

```
PETALWID      0.3229        34.569      0.0001       0.0662    PETAL WIDTH  IN MM.
Variable PETALWID will be entered        The following variable(s) have been entered:
                                                  SEPALWID PETALLEN PETALWID

                          Multivariate Statistics

 Wilks' Lambda  = 0.02497554    F( 6, 290) =  257.503      Prob > F = 0.0001
 Pillai's Trace =   1.189914    F( 6, 292) =   71.485      Prob > F = 0.0001

              Average Squared Canonical Correlation = 0.59495691
Stepwise Selection:  Step 4       Statistics for Removal,  DF = 2, 145
                Partial
  Variable         R**2          F        Prob > F     Label

  SEPALWID       0.4295         54.577       0.0001     SEPAL WIDTH  IN MM.
  PETALLEN       0.3482         38.724       0.0001     PETAL LENGTH IN MM.
  PETALWID       0.3229         34.569       0.0001     PETAL WIDTH  IN MM.

                        No variables can be removed
Stepwise Selection:  Step 4       Statistics for Entry, DF = 2, 144
             Partial
Variable       R**2          F      Prob > F   Tolerance   Label
SEPALLEN     0.0615        4.721      0.0103      0.0320   SEPAL LENGTH IN MM.
Variable SEPALLEN will be entered        All variables have been entered

                          Multivariate Statistics

 Wilks' Lambda  = 0.02343863    F( 8, 288) =  199.145      Prob > F = 0.0001
 Pillai's Trace =   1.191899    F( 8, 290) =   53.466      Prob > F = 0.0001

              Average Squared Canonical Correlation = 0.59594941
Stepwise Selection:  Step 5      Statistics for Removal,  DF = 2, 144
                Partial
  Variable         R**2          F        Prob > F     Label
  SEPALLEN       0.0615         4.721       0.0103     SEPAL LENGTH IN MM.
  SEPALWID       0.2335        21.936       0.0001     SEPAL WIDTH  IN MM.
  PETALLEN       0.3308        35.590       0.0001     PETAL LENGTH IN MM.
  PETALWID       0.2570        24.904       0.0001     PETAL WIDTH  IN MM.
```

No variables can be removed　　No further steps are possible

Stepwise Selection: Summary

Step	Variable Entered	Removed	Number In	Partial R**2	F Statistic	Prob > F	Wilks' Lambda	Prob < Lambda
1	PETALLEN		1	0.9414	1180.161	0.0001	0.05862828	0.0001
2	SEPALWID		2	0.3709	43.035	0.0001	0.03688411	0.0001
3	PETALWID		3	0.3229	34.569	0.0001	0.02497554	0.0001
4	SEPALLEN		4	0.0615	4.721	0.0103	0.02343863	0.0001

Step	Variable Entered	Removed	Number In	Average Squared Canonical Correlation	Prob > ASCC	Label
1	PETALLEN		1	0.47068586	0.0001	PETAL LENGTH IN MM.
2	SEPALWID		2	0.55995394	0.0001	SEPAL WIDTH IN MM.
3	PETALWID		3	0.59495691	0.0001	PETAL WIDTH IN MM.
4	SEPALLEN		4	0.59594941	0.0001	SEPAL LENGTH IN MM.