

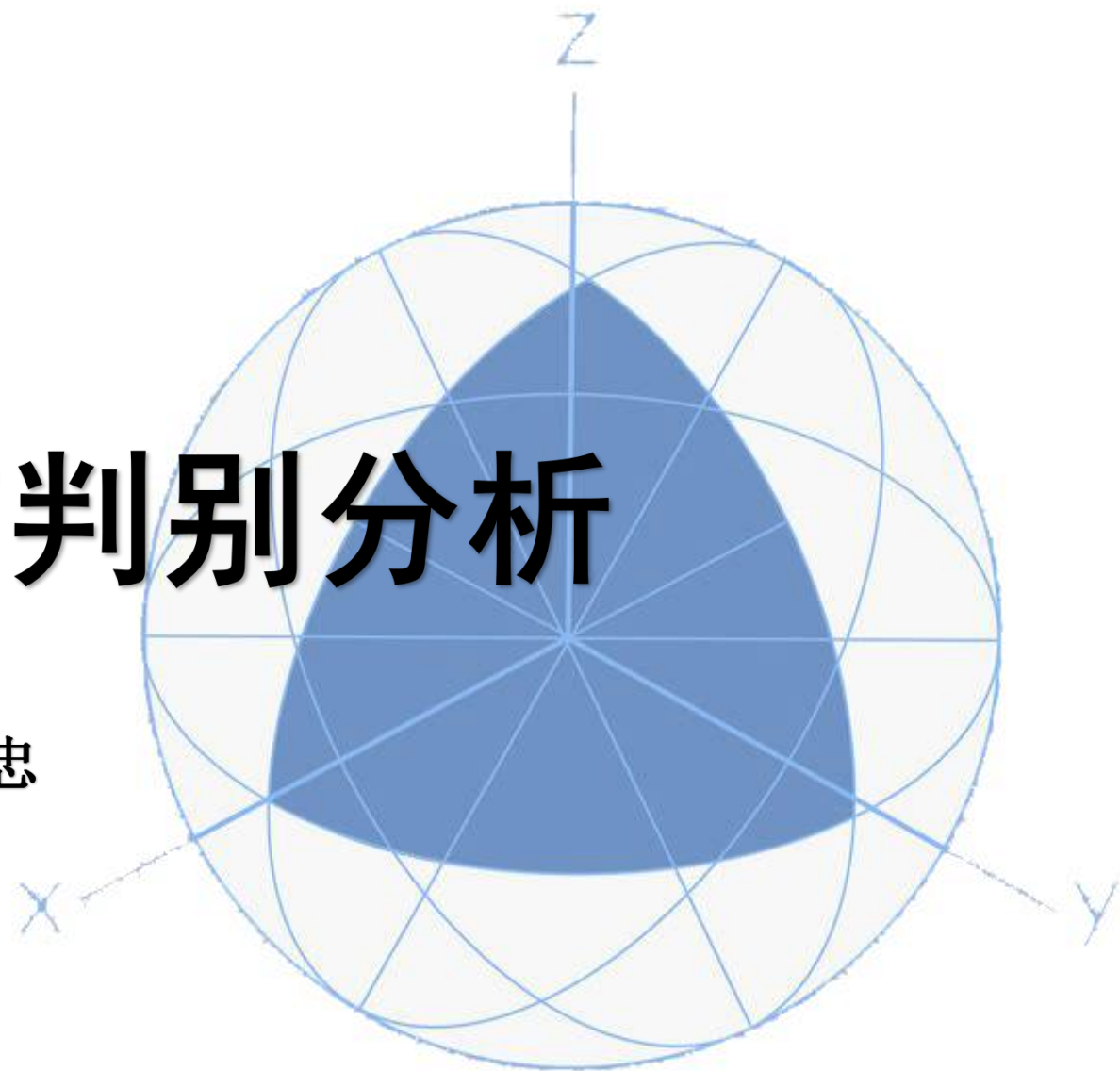


廈門大學

XIAMEN UNIVERSITY

# 聚类分析与判别分析

谭 忠





厦门大学  
XIAMEN UNIVERSITY



案例分析



廈門大學  
XIAMEN UNIVERSITY

Part 1

# 源头问题与当今应用





## 7.1 源头问题与当今应用

“物以类聚，人以群分”，在现实生活中存在大量的分类问题，常见的分类方法有聚类分析和判别分析，它们在生物学、经济学、人口学、生态学、电子商务等很多方面有着非常广泛的应用。

如银行在进行个人贷款业务时，会对贷款者的资格进行审核，包括贷款者的收入水平，抵押状况、有无不良信用记录等信息。那么怎样根据这些信息对贷款者进行分类评价，并给予相应的贷款额度呢？





廈門大學  
XIAMEN UNIVERSITY

Part 2

# 聚类与判别思想 及其建模方法





聚类与判别都是分类问题，那么什么是分类？分类就是将一个观测对象指定到某一类 (组).

分类的问题可以分成两种：一种是对当前所研究的问题已知它的类别数目，且知道各类的特征 (如分布规律，或知道来自各类的训练样本)，我们要将另一些未知类别的个体正确归属于其中某一类，这就是判别分析所要解决的问题. 另一种是事先不知道观测个体的具体分类情况，我们要选定一种度量个体接近程度的标准，并按这种亲近标准把观测对象合理分类. 这种问题正是聚类分析所要解决的问题.



## 7.2 聚类分析

常用聚类分析的方法可分为以下几种：

- (1) 系统聚类法；
- (2) K-均值聚类法 (动态聚类法)；
- (3) 模糊聚类法；

聚类分析根据分类对象的不同分为 R 型和 Q 型两大类，R 型是对变量 (指标) 进行分类处理，Q 型是对样品进行分类处理。

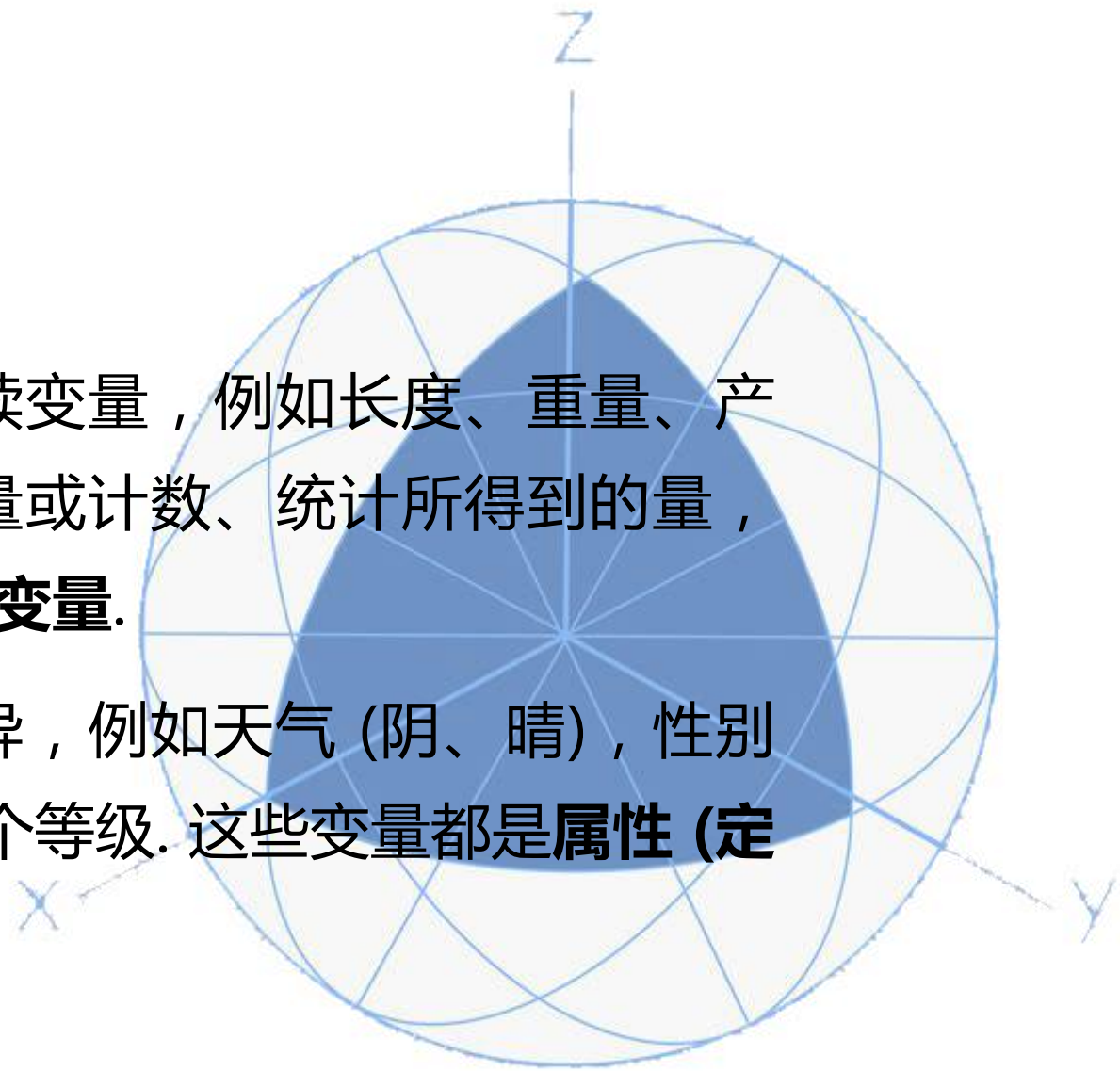




## 1、定量变量和定性 (属性)变量

定量变量就是我们通常所说的连续变量，例如长度、重量、产量、人口、温度等，它们是由测量或计数、统计所得到的量，这类变量具有数值特征，称为**定量变量**.

定性 (属性)变量只有性质上的差异，例如天气 (阴、晴)，性别 (男、女)，产品质量分为上中下三个等级. 这些变量都是**属性 (定性)变量**.



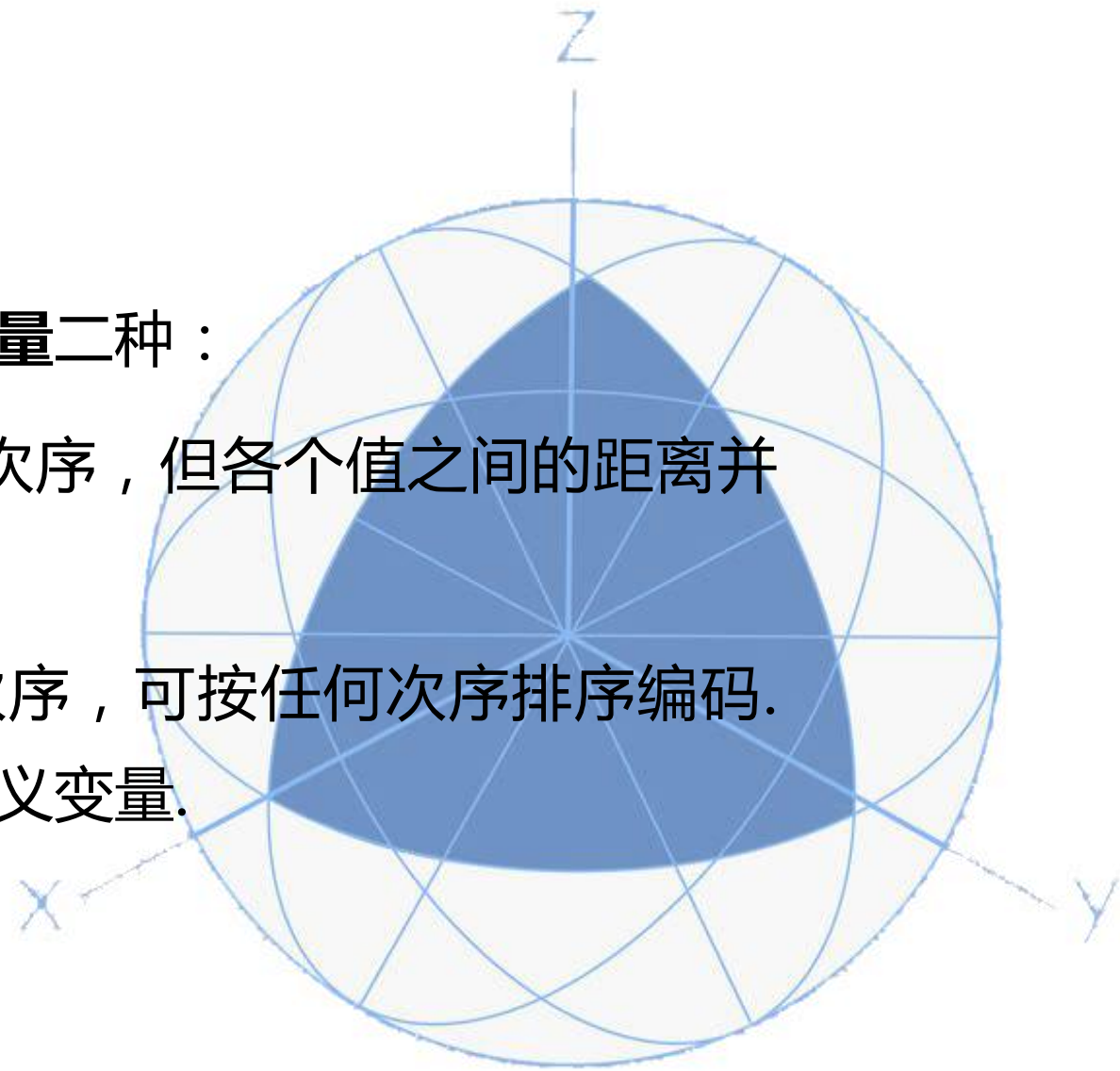




属性变量又分为**名义变量**和**有序变量**二种：

(1)有序变量：其值有明确的逻辑次序，但各个值之间的距离并不清楚.

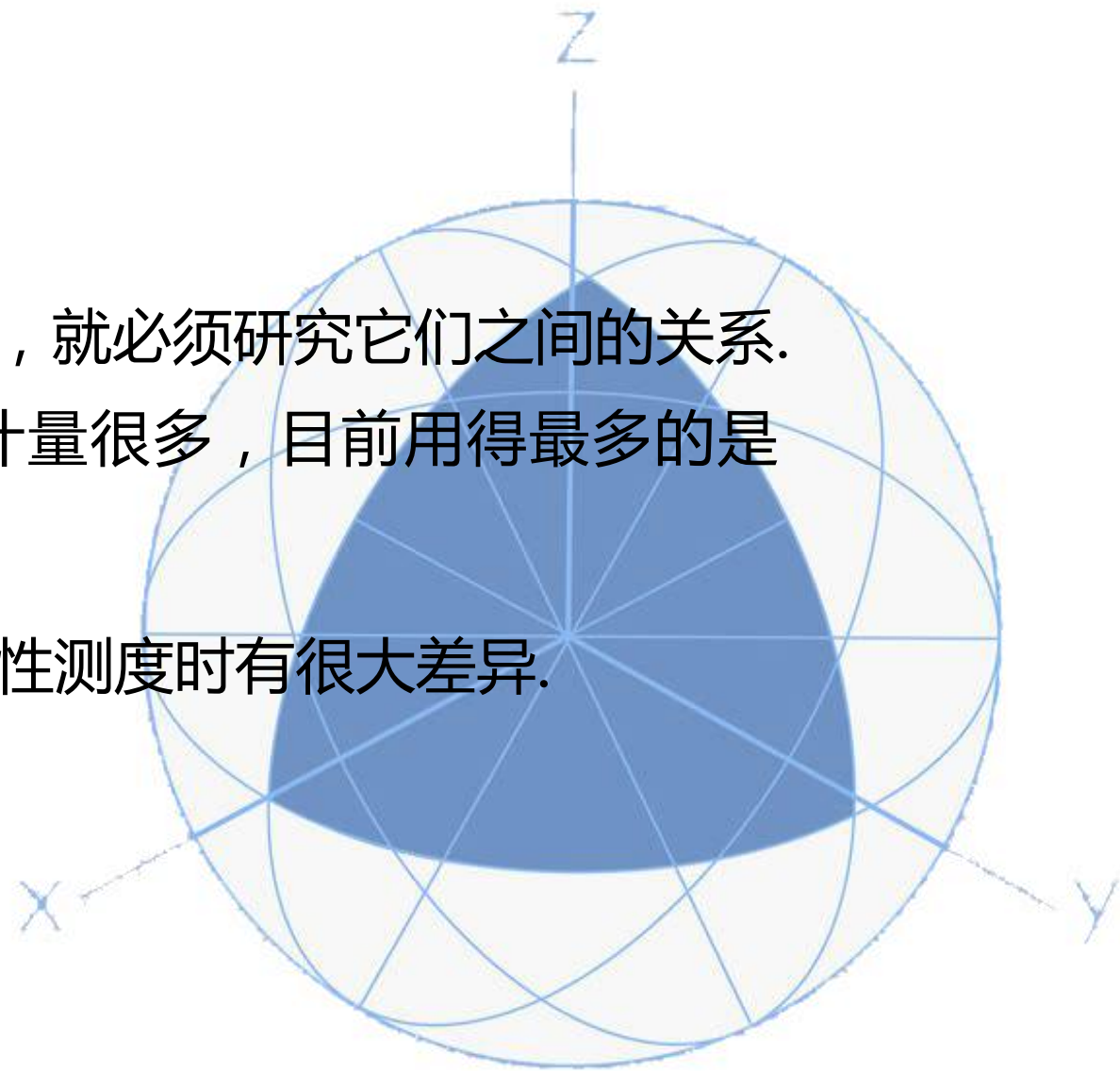
(2) 名义变量：其值之间无逻辑次序，可按任何次序排序编码.  
例如性别，职业，地区等等都是名义变量.





为了对观测样品 (或变量)进行分类，就必须研究它们之间的关系. 描述样品之间亲疏相似程度的统计量很多，目前用得最多的是距离和相似系数.

不同类型的变量在定义距离或相似性测度时有很大差异.



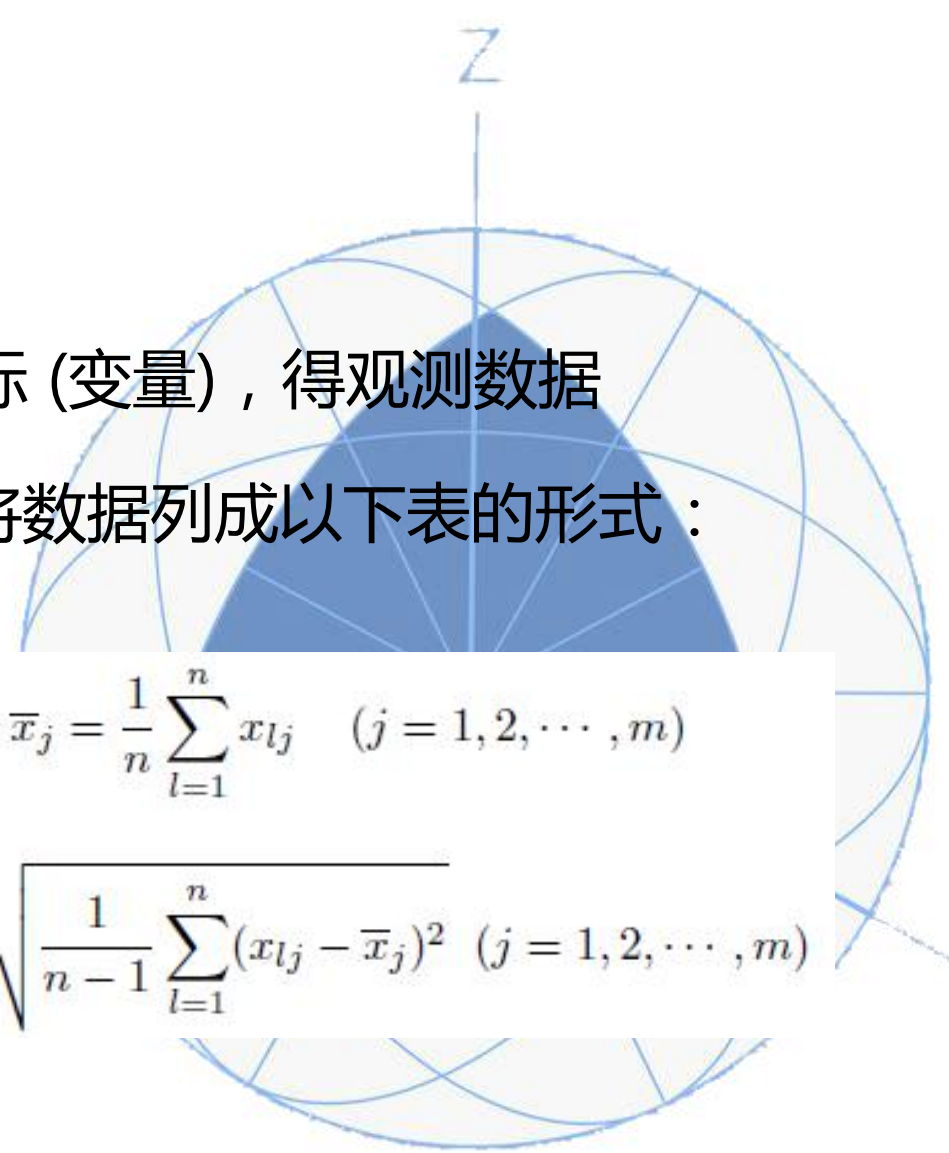


## 2、数据的变换方法

设有  $n$  个样品，每个样品测得  $m$  项指标 (变量)，得观测数据

$x_{ij}$  ( $i = 1, \dots, n, j = 1, \dots, m$ ). 通常将数据列成以下表的形式：

	$X_1$	$\dots$	$X_j$	$\dots$	$X_m$
$X_{(1)}$	$x_{11}$	$\dots$	$x_{1j}$	$\dots$	$x_{1m}$
$\dots$	$\dots$		$\dots$		$\dots$
$X_{(i)}$	$x_{i1}$	$\dots$	$x_{ij}$	$\dots$	$x_{im}$
$\dots$	$\dots$		$\dots$		$\dots$
$X_{(n)}$	$x_{n1}$	$\dots$	$x_{nj}$	$\dots$	$x_{nm}$
均值	$\bar{x}_1$	$\dots$	$\bar{x}_j$	$\dots$	$\bar{x}_m$
标准差	$s_1$	$\dots$	$s_j$	$\dots$	$s_m$
极差	$R_1$	$\dots$	$R_j$	$\dots$	$R_m$


$$\bar{x}_j = \frac{1}{n} \sum_{l=1}^n x_{lj} \quad (j = 1, 2, \dots, m)$$

$$s_j = \sqrt{\frac{1}{n-1} \sum_{l=1}^n (x_{lj} - \bar{x}_j)^2} \quad (j = 1, 2, \dots, m)$$



## (1) 中心化变换

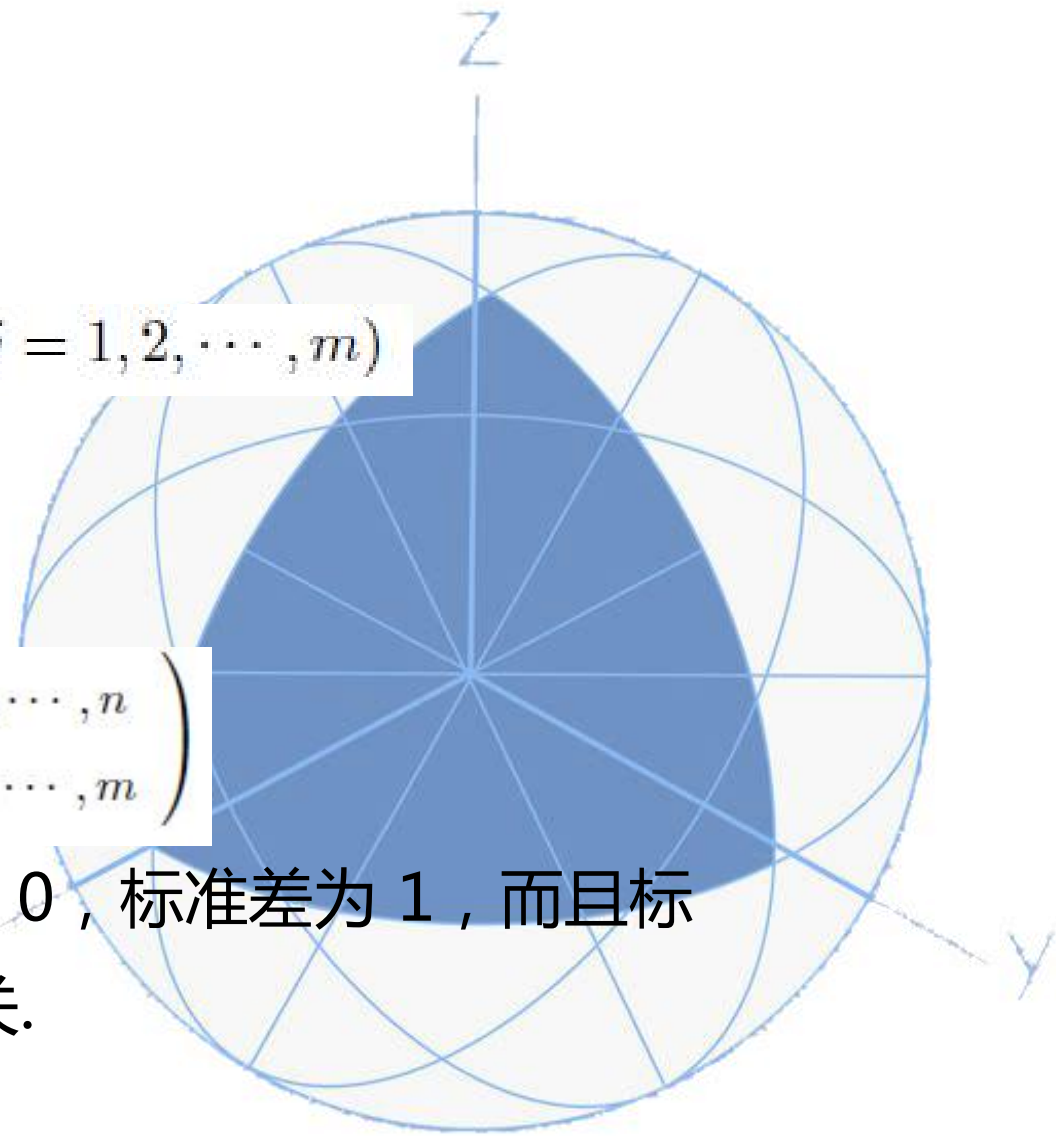
$$x_{ij}^* = x_{ij} - \bar{x}_j (i = 1, 2, \dots, n; j = 1, 2, \dots, m)$$

变换后数据的均值为 0，而协差阵不变.

## (2) 标准化变换

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{s_j} \begin{pmatrix} i = 1, 2, \dots, n \\ j = 1, 2, \dots, m \end{pmatrix}$$

变换后的数据，每个变量的样本均值为 0，标准差为 1，而且标准化变换后的数据  $x_{ij}$  与变量的量纲无关.







### (3) 极差标准化变换

$$x_{ij}^* = \frac{x_{ij} - \bar{x}_j}{R_j} \quad \begin{pmatrix} i = 1, 2, \dots \\ j = 1, 2, \dots, m \end{pmatrix}$$

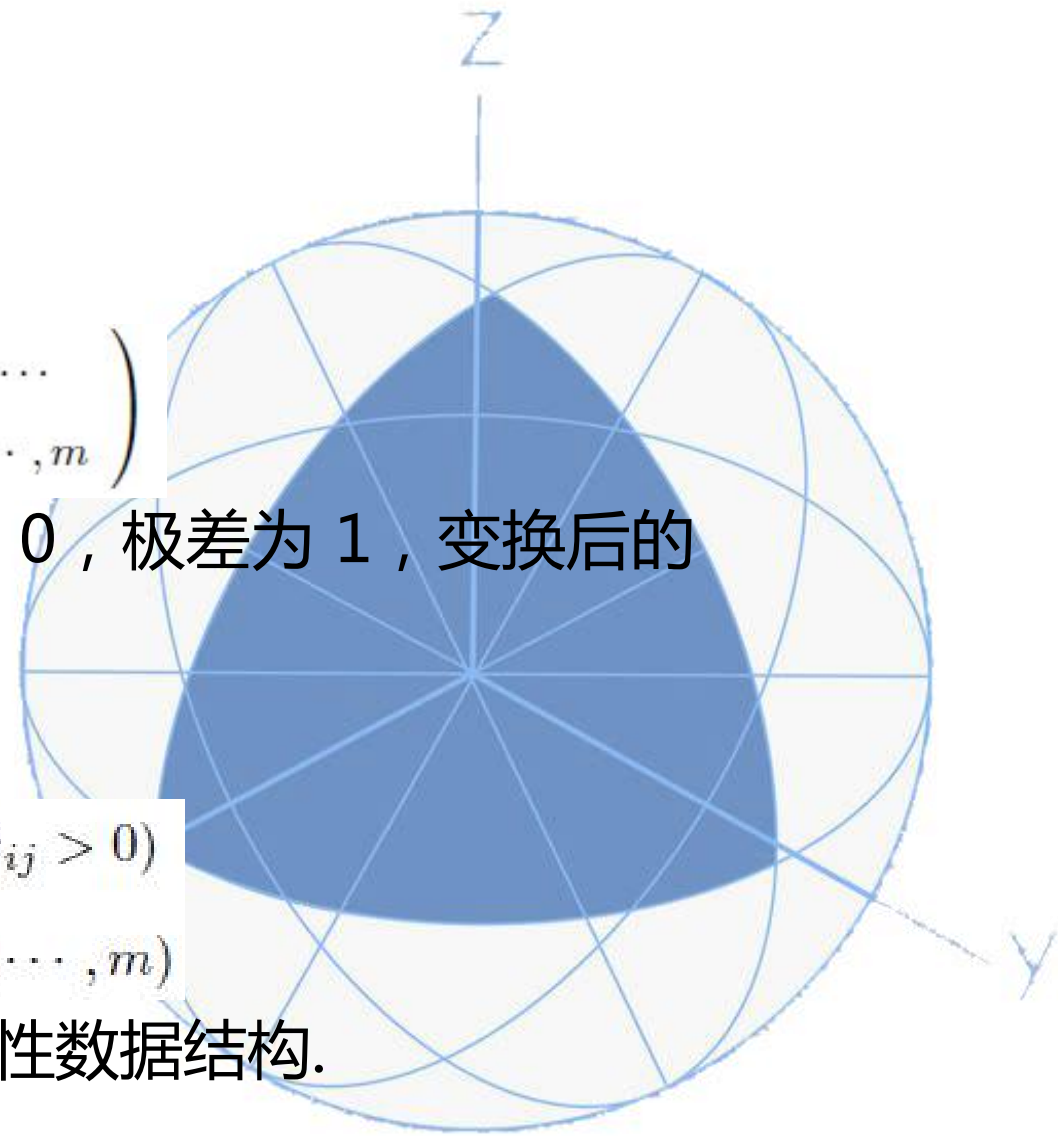
变换后的数据，每个变量的样本均值为 0，极差为 1，变换后的数据也是无量纲的量。

### (4) 对数变换

$$x_{ij}^* = \log(x_{ij}) \quad (\text{要求 } x_{ij} > 0)$$

$$(i = 1, 2, \dots, n; j = 1, 2, \dots, m)$$

它可将具有指数特征的数据结构化为线性数据结构。



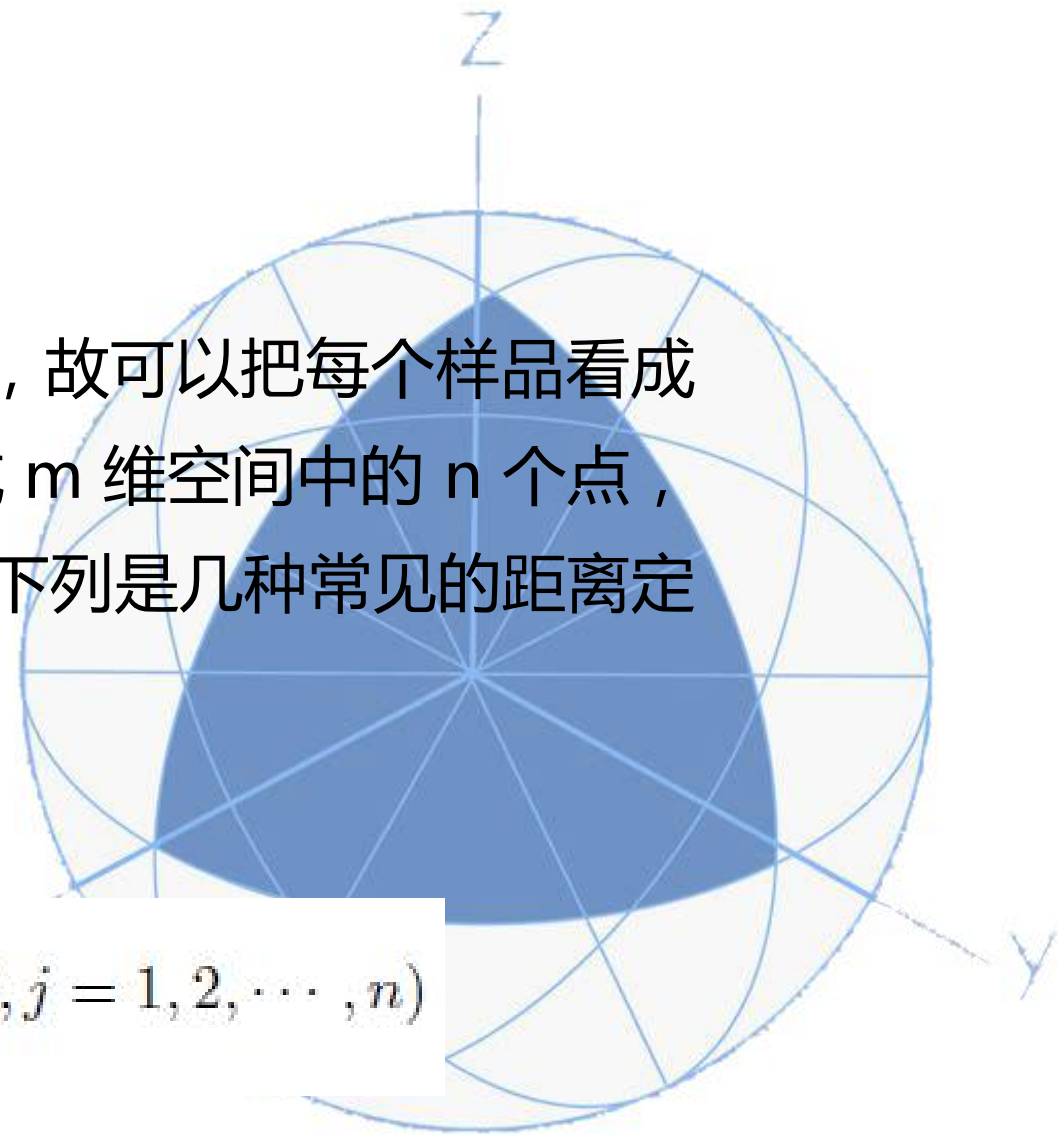


### 3、样品间的距离

在上述表格中，每个样品有  $m$  个指标，故可以把每个样品看成  $m$  维空间中的一个点， $n$  个样品就组成  $m$  维空间中的  $n$  个点，用  $d_{ij}$  表示样品  $X_{(i)}$  和  $X_{(j)}$  之间的距离，下列是几种常见的距离定义式。

#### (1) 闵科夫斯基 (Minkowski) 距离

$$d_{ij}(q) = \left[ \sum_{k=1}^m |x_{ik} - x_{jk}|^q \right]^{\frac{1}{q}} (i, j = 1, 2, \dots, n)$$





当  $q = 1$  时的一阶 Minkowski 度量就称为绝对值距离

$$d_{ij}(1) = \sum_{k=1}^m |x_{ik} - x_{jk}| (i, j = 1, 2, \dots, n)$$

当  $q = 2$  时的二阶 Minkowski 度量就称为欧氏距离. 欧氏距离是聚类分析中用得最广泛的距离

$$d_{ij}(2) = \sqrt{\sum_{k=1}^m |x_{ik} - x_{jk}|^2} (i, j = 1, 2, \dots, n)$$

当  $q = \infty$  时的 Minkowski 度量称为切比雪夫距离

$$d_{ij}(\infty) = \max_{k=1}^m |x_{ik} - x_{jk}| (i, j = 1, 2, \dots, n)$$



缺点：

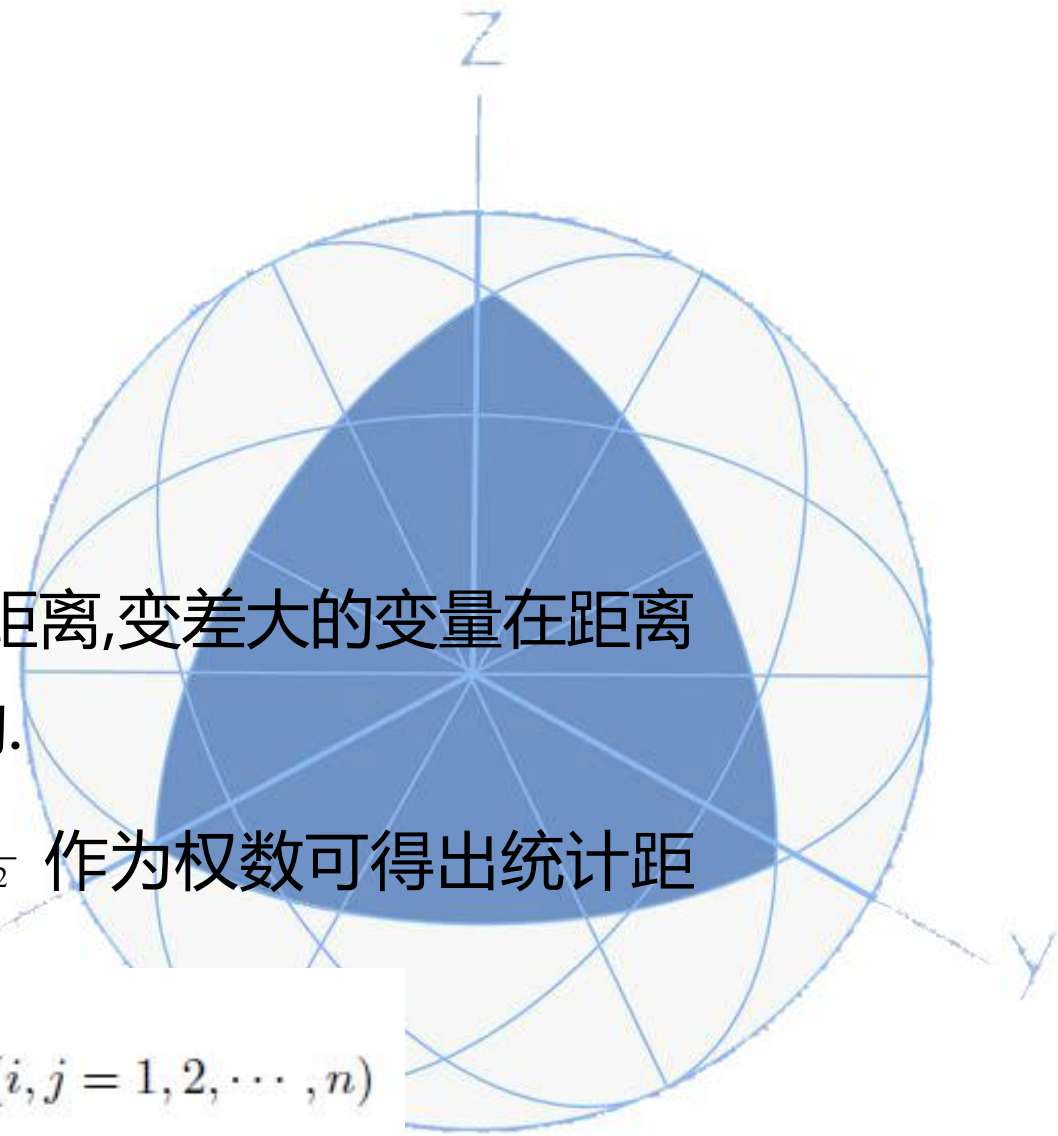
(i)与各变量的量纲有关；

(ii)没有考虑指标的相关性；

(iii)没有考虑各变量方差的不同.如欧式距离,变差大的变量在距离中的作用 (贡献) 就会大，这是不合适的.

合理的方法就是对各变量加权，如用  $\frac{1}{s^2}$  作为权数可得出统计距离：

$$d_{ij}^* = \sqrt{\sum_{k=1}^p \left( \frac{x_{ik} - x_{jk}}{s_k} \right)^2} \quad (i, j = 1, 2, \dots, n)$$







对  $n$  个样品计算两两间的距离  $d_{ij}$  后，可排成矩阵  $D$

$$D = \begin{pmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ \vdots & \vdots & & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{pmatrix}_{n \times n}$$

$d_{ij}$  值越小，表示两个样品越接近， $d_{ij}$  值越大，表示两个样品越不接近。



## (2) 兰氏距离 (要求 $x_{ij} > 0$ )

兰氏距离是由 Lance 和 Williams 最早提出的,故称为兰氏距离

$$d_{ij}(L) = \frac{1}{m} \sum_{k=1}^P \frac{|x_{ik} - x_{jk}|}{(x_{ik} + x_{jk})} (i, j = 1, 2, \dots, n)$$

这是一个无量纲的量, 克服了闵氏距离与各指标的量纲有关的缺点. 且兰氏距离对大的奇异值不敏感, 这样使得它特别适合高度偏倚的数据. 但兰氏距离也没有考虑变量间的相关性.



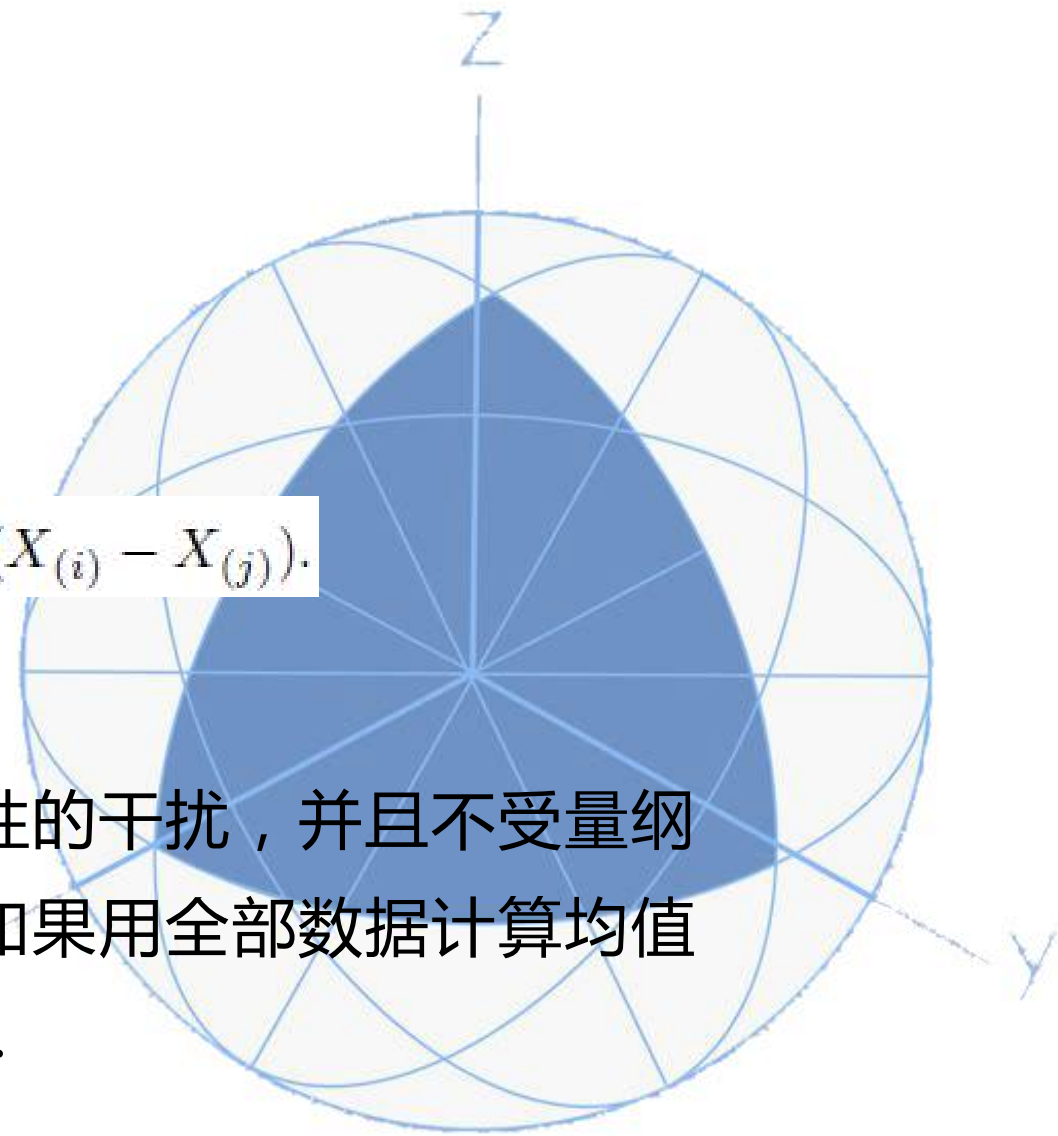
### (3) 马氏距离 (Mahalanobis)

样品  $X_{(i)}$  和  $X_{(j)}$  的马氏距离为

$$d_{ij}(M) = (X_{(i)} - X_{(j)})^T S^{-1} (X_{(i)} - X_{(j)}).$$

其中  $S^{-1}$  为样本协差阵的逆矩阵.

马氏距离虽然可以排除变量之间相关性的干扰，并且不受量纲的影响，但是在聚类分析处理之前，如果用全部数据计算均值和协差阵来求马氏距离，效果不是很好。



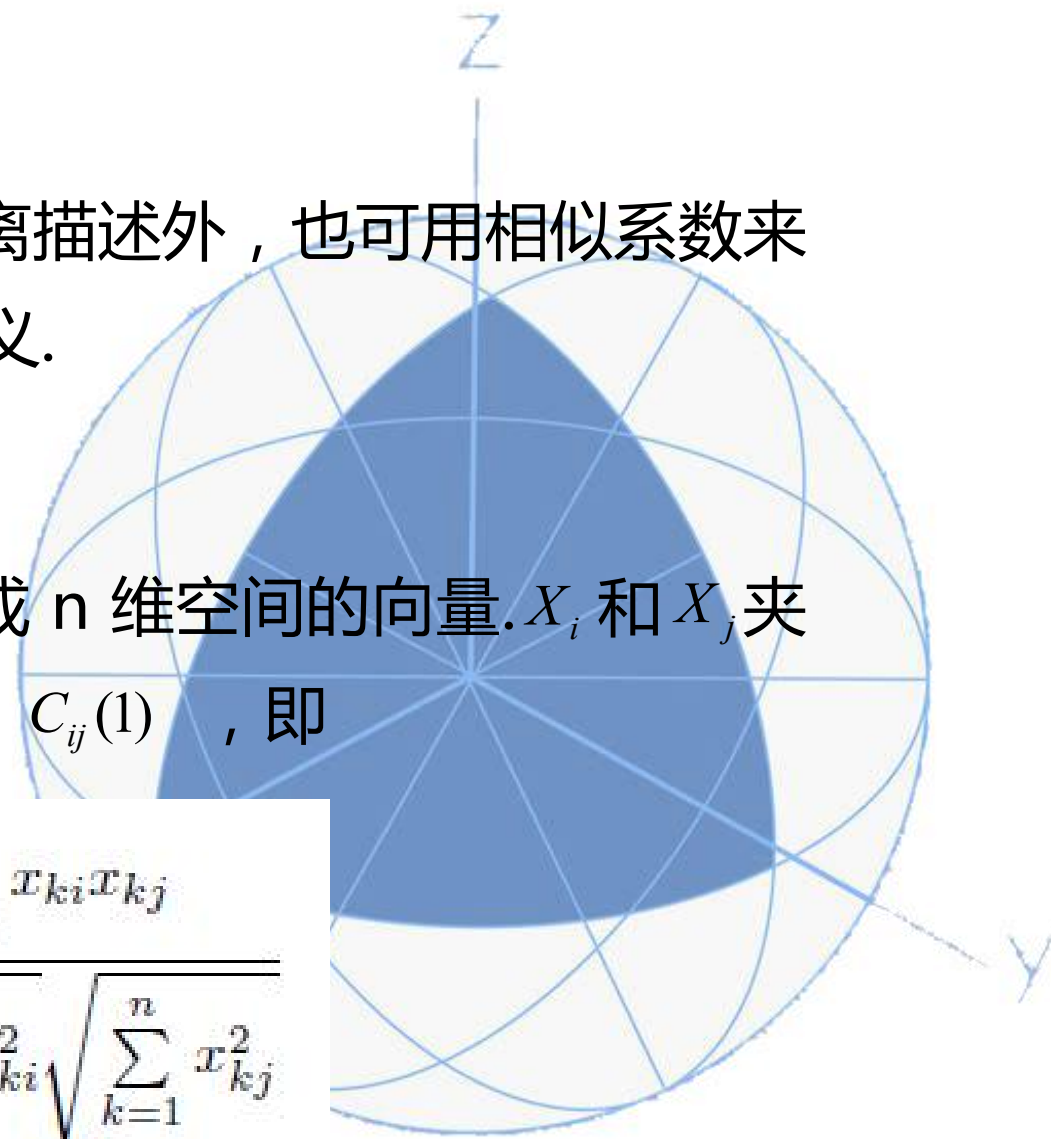


(4)相似系数：样品间的亲疏程度除了用距离描述外，也可用相似系数来表示. 参见以下“变量间的相似系数”的定义.

(a)夹角余弦

变量  $X_i$  的  $n$  次观测值  $(x_{1i}, x_{2i}, \dots, x_{ni})$  看成  $n$  维空间的向量.  $X_i$  和  $X_j$  夹角  $\alpha_{ij}$  的余弦称为两向量的相似系数，记为  $C_{ij}(1)$ ，即

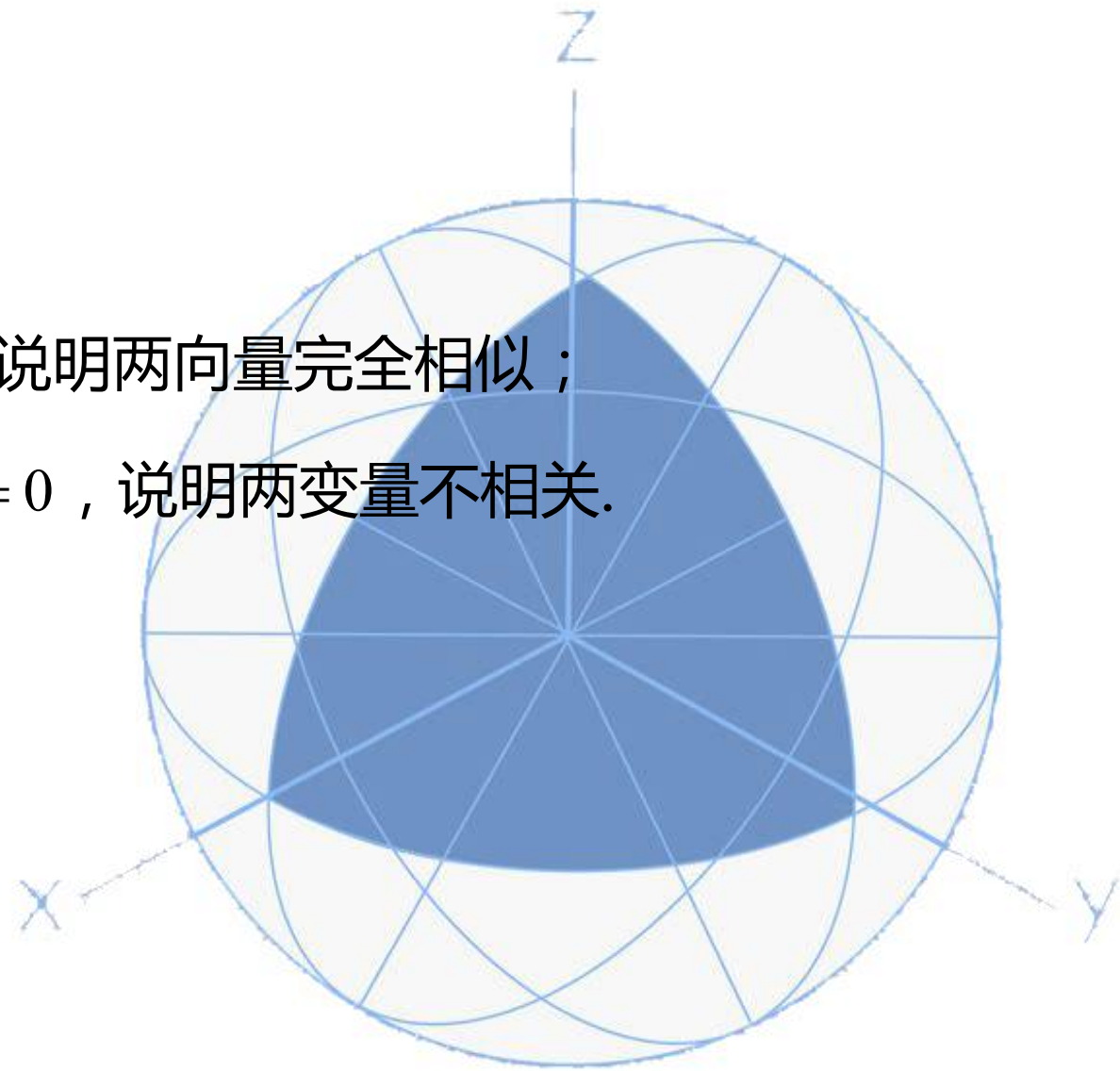
$$C_{ij}(1) = \cos[\alpha_{ij}] = \frac{\sum_{k=1}^n x_{ki} x_{kj}}{\sqrt{\sum_{k=1}^n x_{ki}^2} \sqrt{\sum_{k=1}^n x_{kj}^2}}$$







当  $i = j$  时, 夹角  $\alpha_{ij} = 0, C_{ij}(1) = 1$ , 说明两向量完全相似;  
当  $X_i$  和  $X_j$  正交时,  $\alpha_{ij} = 90, C_{ij}(1) = 0$ , 说明两变量不相关.



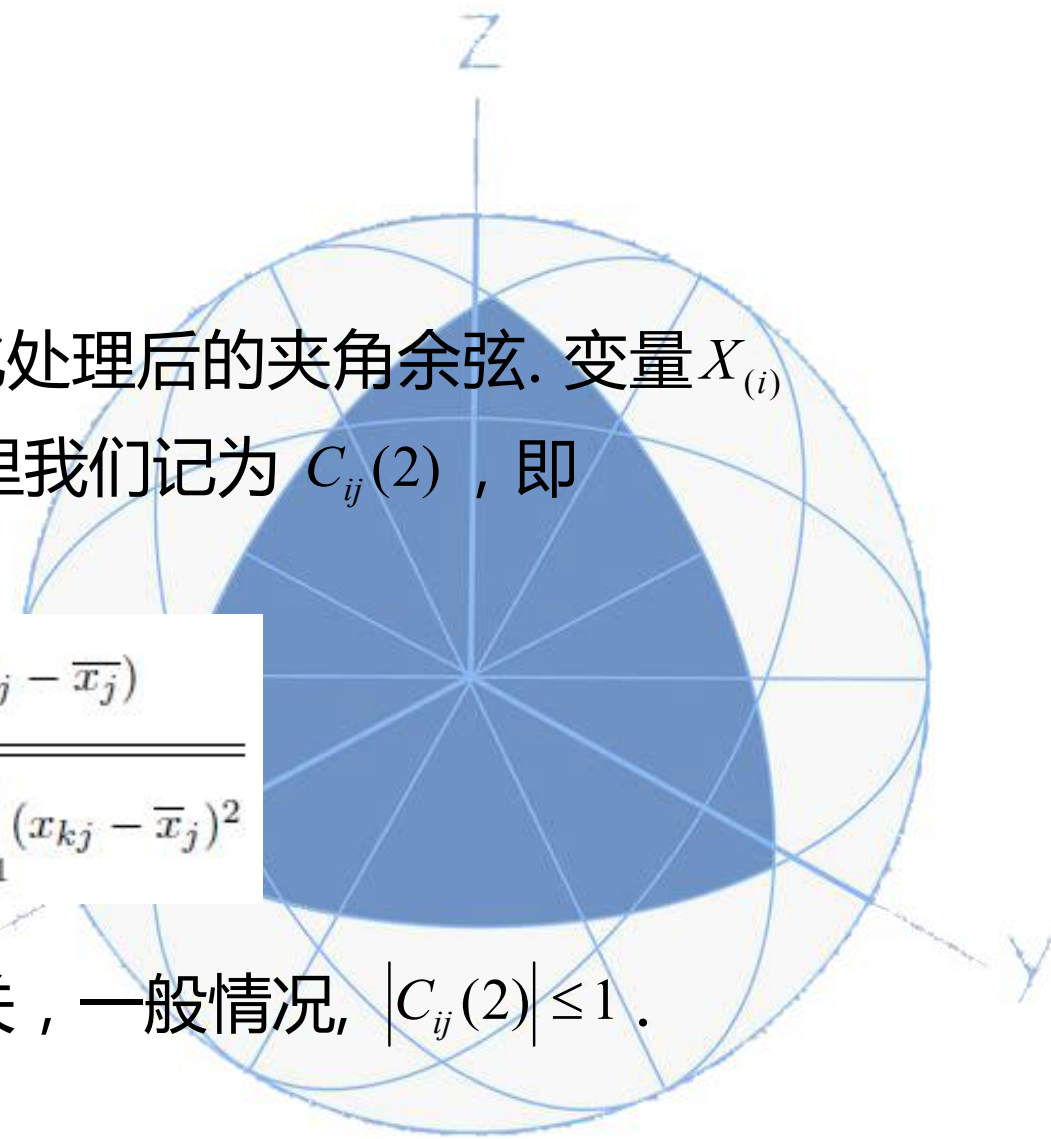


## (b) 相关系数

相关系数就是对数据作中心化或标准化处理后的夹角余弦. 变量  $X_{(i)}$  和  $X_{(j)}$  的相关系数常用  $r_{ij}$  表示, 在这里我们记为  $C_{ij}(2)$ , 即

$$C_{ij}(2) = \frac{\sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^n (x_{ki} - \bar{x}_i)^2} \sqrt{\sum_{k=1}^n (x_{kj} - \bar{x}_j)^2}}$$

当  $i = j$  时,  $C_{ij} = 1$  表示两变量线性相关, 一般情况,  $|C_{ij}(2)| \leq 1$ .





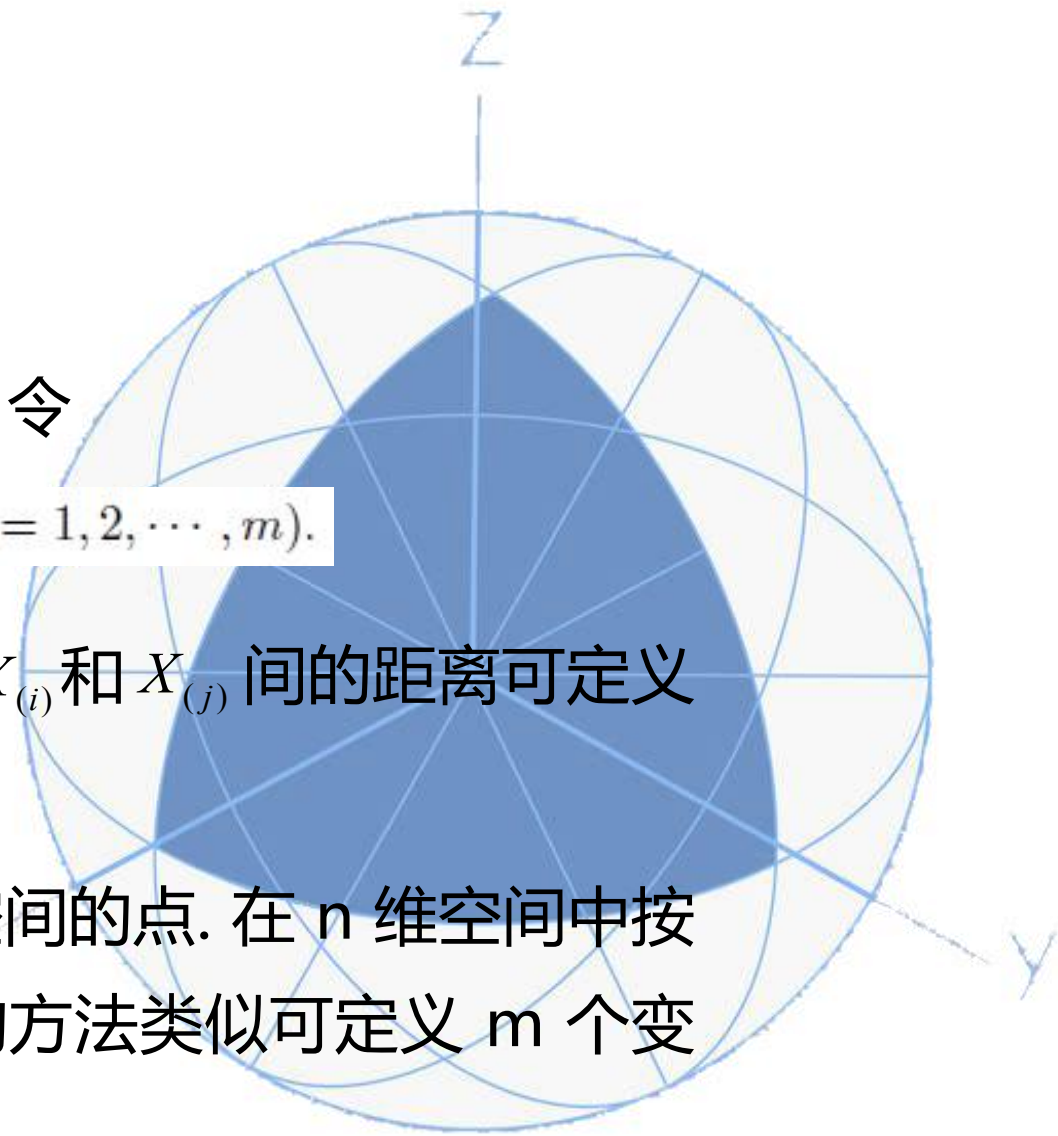
### (c)变量间的距离

(1)利用相似系数来定义变量间的距离，令

$$d_{ij} = 1 - |C_{ij}| \text{ 或 } d_{ij}^2 = 1 - C_{ij}^2 (i, j = 1, 2, \dots, m).$$

(2)利用样本协方差阵  $S = (s_{ij}) > 0$ ，变量  $X_{(i)}$  和  $X_{(j)}$  间的距离可定义为  $d_{ij} = s_{ii} + s_{jj} - 2s_{ij}$

(3)把变量  $X_i$  的  $n$  次观测值看成  $n$  维空间的点. 在  $n$  维空间中按“样品间的距离和相似系数”中介绍的方法类似可定义  $m$  个变量间的种种距离.

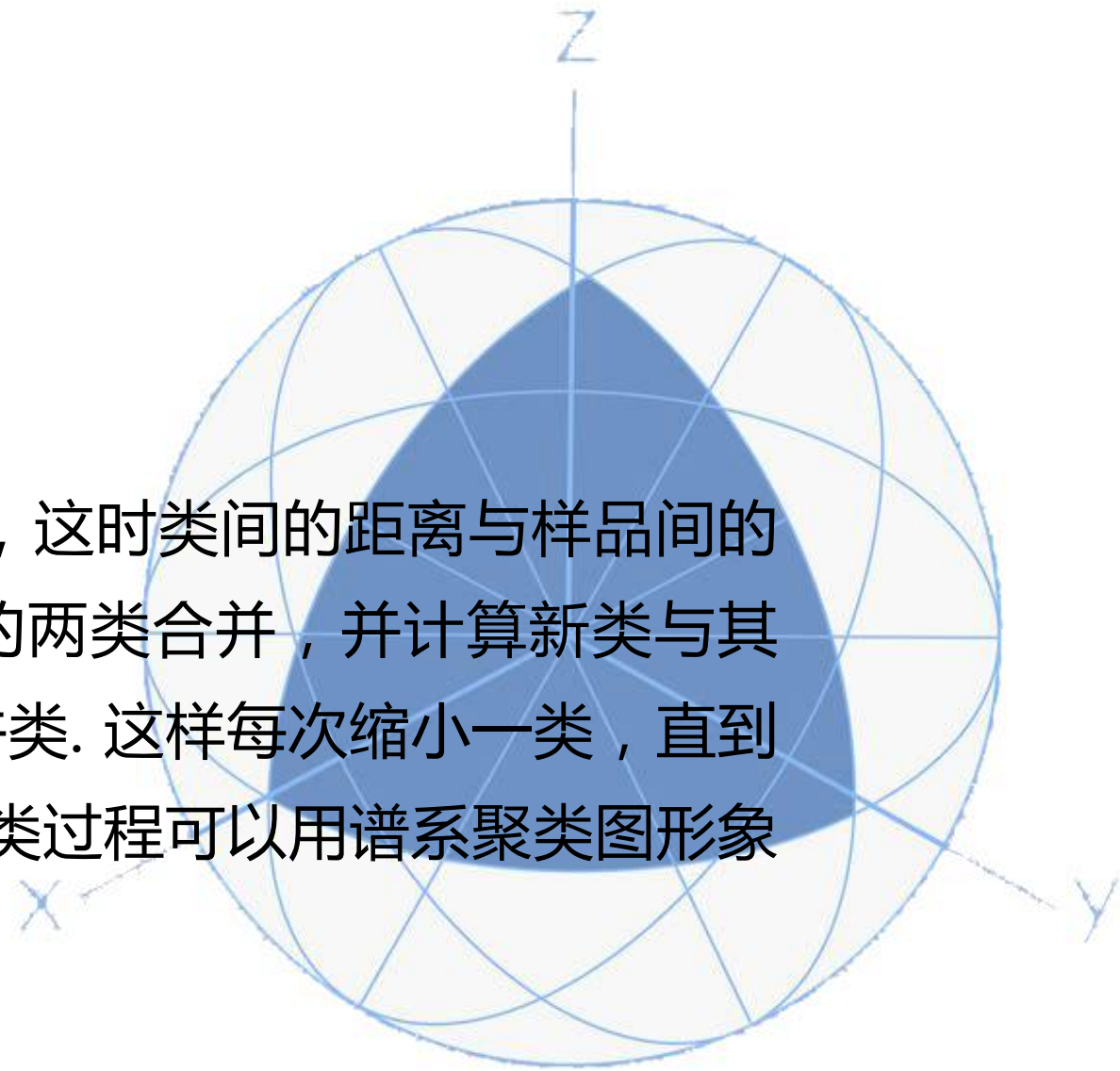




## 4、系统聚类方法

### 4.1 系统聚类方法的基本思想是：

一开始将  $n$  个样品各自自成一类，这时类间的距离与样品间的距离是等价的；然后将距离最近的两类合并，并计算新类与其他类的类间距离，再按最小距离并类. 这样每次缩小一类，直到所有的样品都成一类为止. 这个并类过程可以用谱系聚类图形象地表达出来.







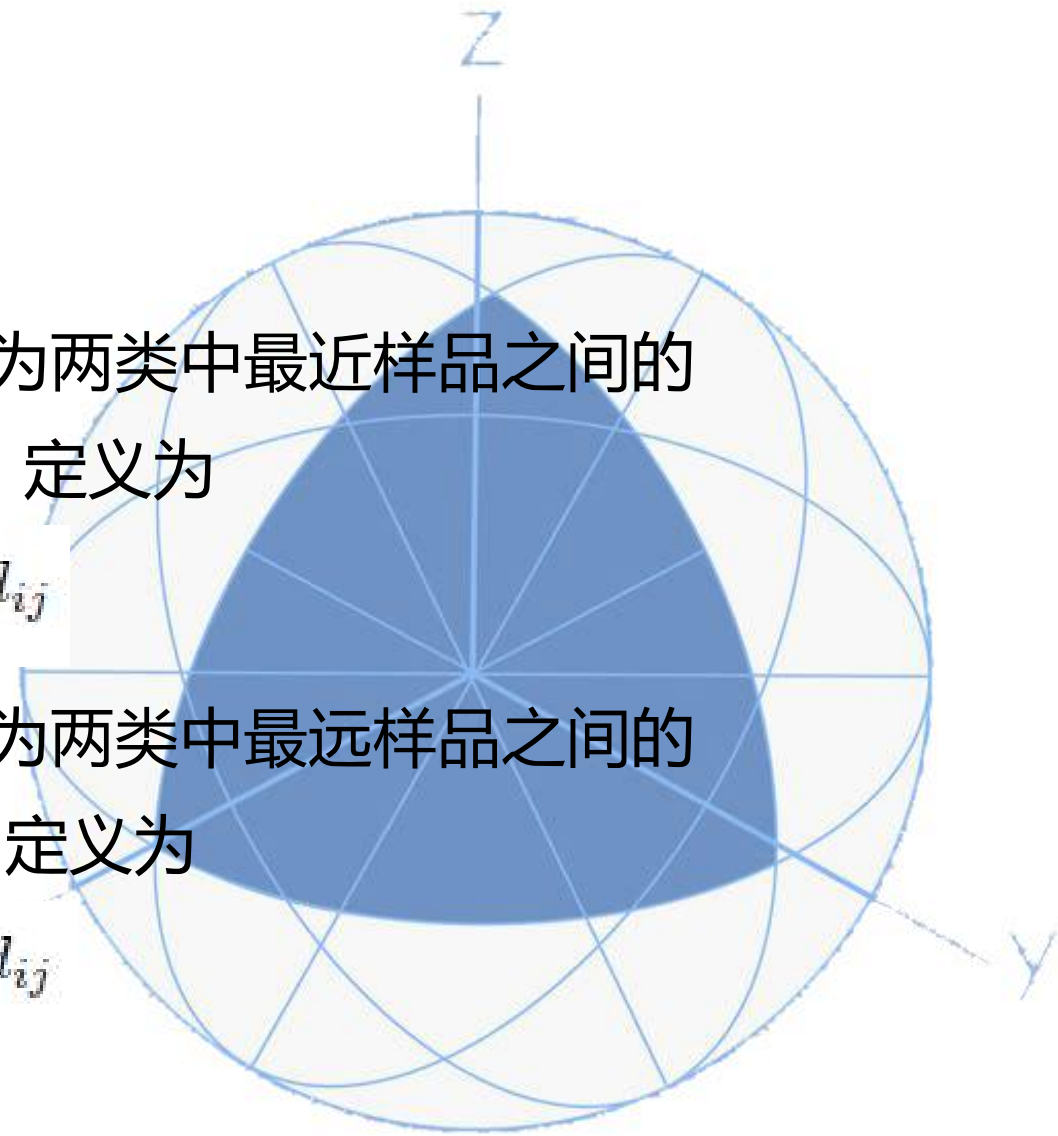
## 4.2 类与类间距离

(1)最短距离法 :类与类之间的距离定义为两类中最近样品之间的距离 , 即类  $G_p$  和  $G_q$  之间的距离  $G_{pq}$  定义为

$$D_{pq} = \min_{i \in G_p, j \in G_q} d_{ij}$$

(2)最长距离法 :类与类之间的距离定义为两类中最远样品之间的距离 , 即类  $G_p$  和  $G_q$  之间的距离  $G_{pq}$  定义为

$$D_{pq} = \max_{i \in G_p, j \in G_q} d_{ij}$$

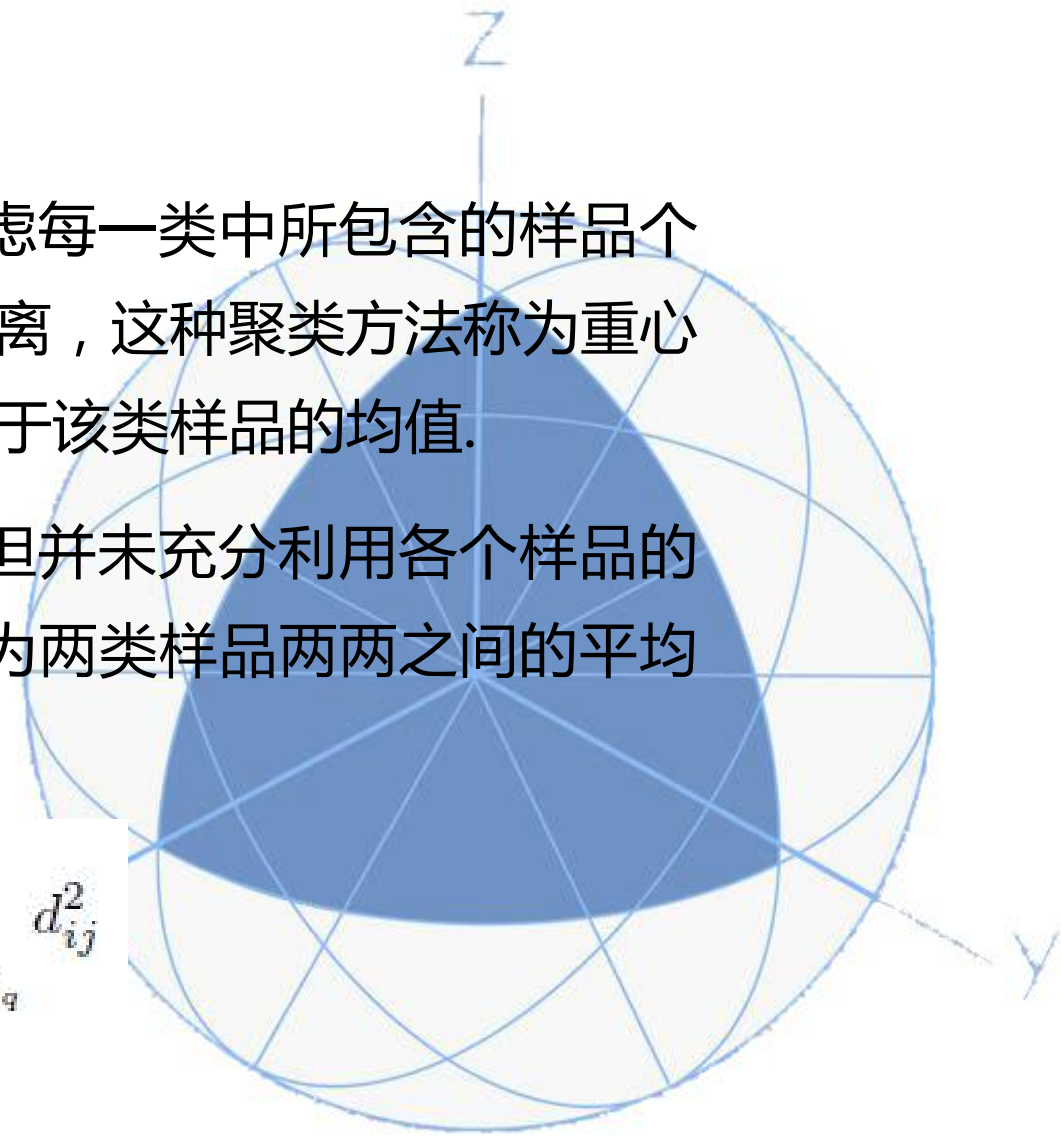




(3)重心法 :在定义类与类之间距离时，考虑每一类中所包含的样品个数. 将两类间的距离定义为两类重心间的距离，这种聚类方法称为重心法. 对样品分类而言，每一类的重心就是属于该类样品的均值.

(4)类平均法 :重心法虽有较好的代表性，但并未充分利用各个样品的信息. 类平均法把类与类之间的距离定义为两类样品两两之间的平均平方距离，即

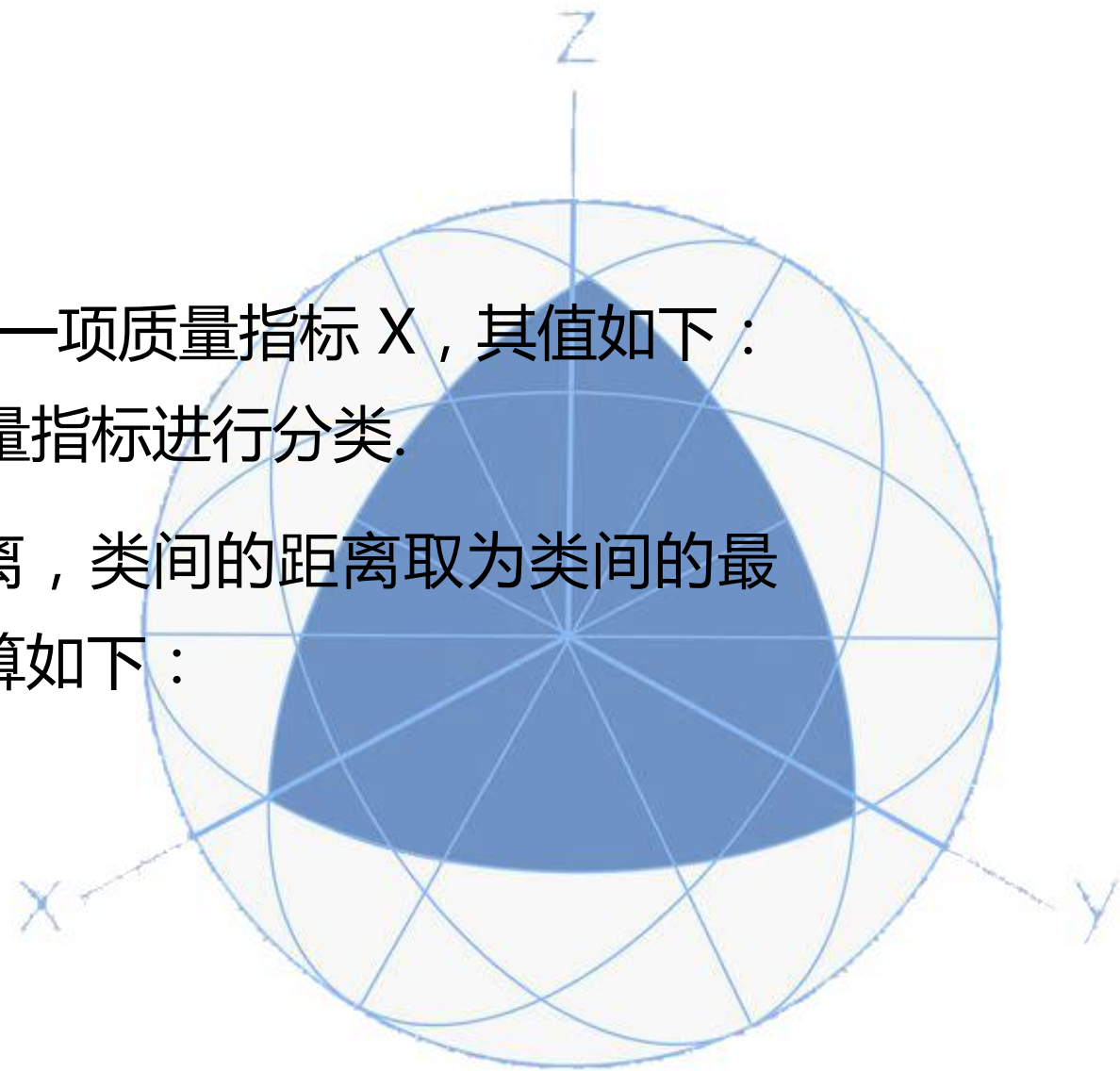
$$D_{pq}^2 = \frac{1}{n_p n_q} \sum_{i \in G_p, j \in G_q} d_{ij}^2$$





**例** 设有 5 个产品，每个产品测得一项质量指标  $X$ ，其值如下：  
1, 2, 4.5, 6, 8. 试对 5 个产品按质量指标进行分类.

**解：**设样品间的距离取为欧氏距离，类间的距离取为类间的最短距离. 根据上面介绍的步骤，计算如下：





(1) 计算 5 个样品  $X(1), X(2), X(3), X(4), X(5)$  两两间的距离，得初始的类间距离矩阵  $D^{(1)}$ ：

表 7.3 距离阵  $D^{(1)}$

	$G_1 = \{X_{(1)}\}$	$G_2 = \{X_{(2)}\}$	$G_3 = \{X_{(3)}\}$	$G_4 = \{X_{(4)}\}$	$G_5 = \{X_{(5)}\}$
$G_1 = \{X_{(1)}\}$	0	[1]	3.5	5	7
$G_2 = \{X_{(2)}\}$		0	2.5	4	6
$G_3 = \{X_{(3)}\}$			0	1.5	3.5
$G_4 = \{X_{(4)}\}$				0	2
$G_5 = \{X_{(5)}\}$					0





表 7.4 距离阵  $D^{(2)}$

	$G_3 = \{X_{(3)}\}$	$G_4 = \{X_{(4)}\}$	$G_5 = \{X_{(5)}\}$	$G_6 = \{X_{(1)}, X_{(2)}\}$
$G_3 = \{X_{(3)}\}$	0	[1.5]	3.5	2.5
$G_4 = \{X_{(4)}\}$		0	2	4
$G_5 = \{X_{(5)}\}$			0	6
$G_6 = \{X_{(1)}, X_{(2)}\}$				0

表 7.5 距离阵  $D^{(3)}$

	$G_5 = \{X_{(5)}\}$	$G_6 = \{X_{(1)}, X_{(2)}\}$	$G_7 = \{X_{(3)}, X_{(4)}\}$
$G_5 = \{X_{(5)}\}$	0	6	[2]
$G_6 = \{X_{(1)}, X_{(2)}\}$		0	2.5
$G_7 = \{X_{(3)}, X_{(4)}\}$			0

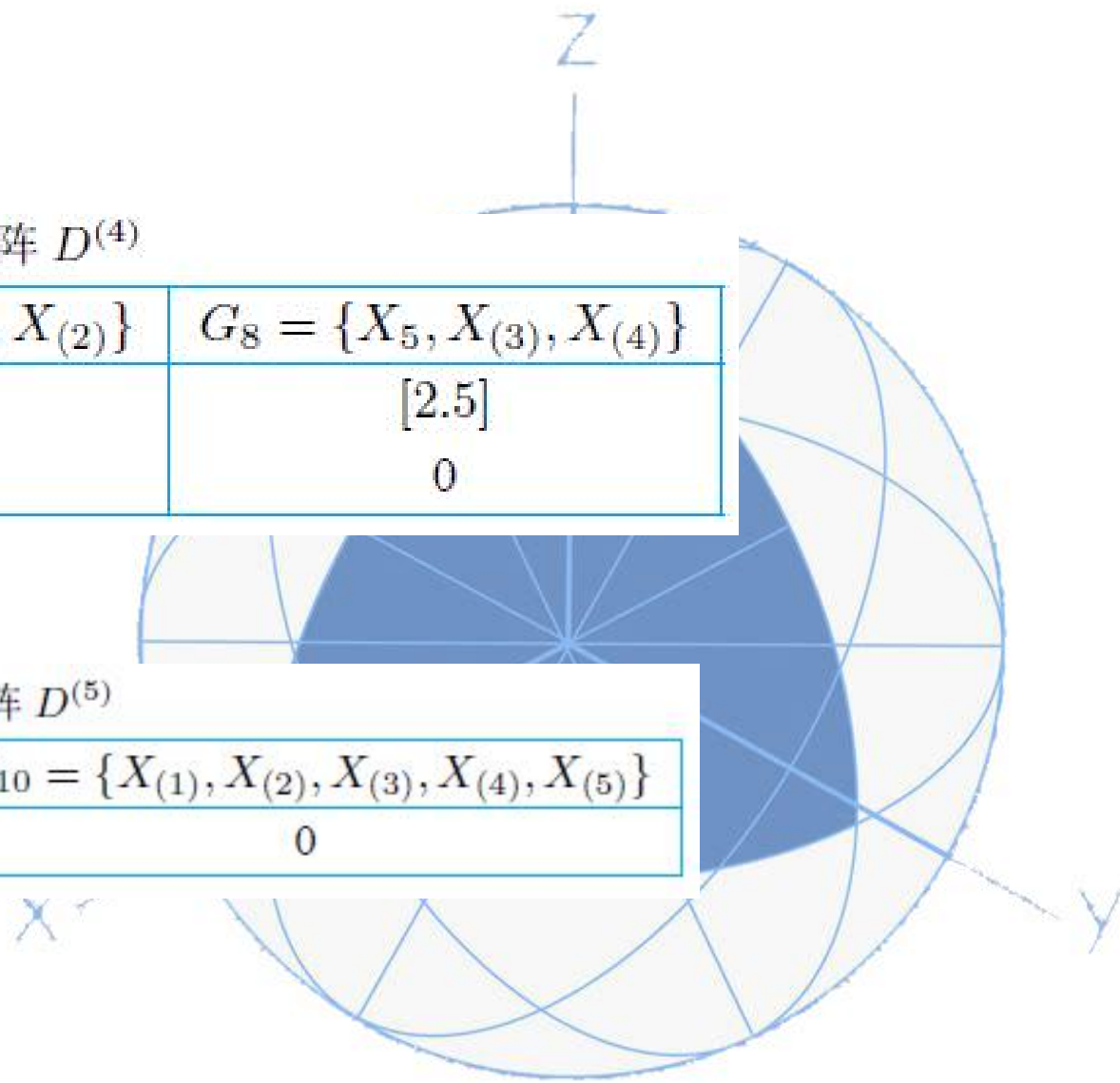


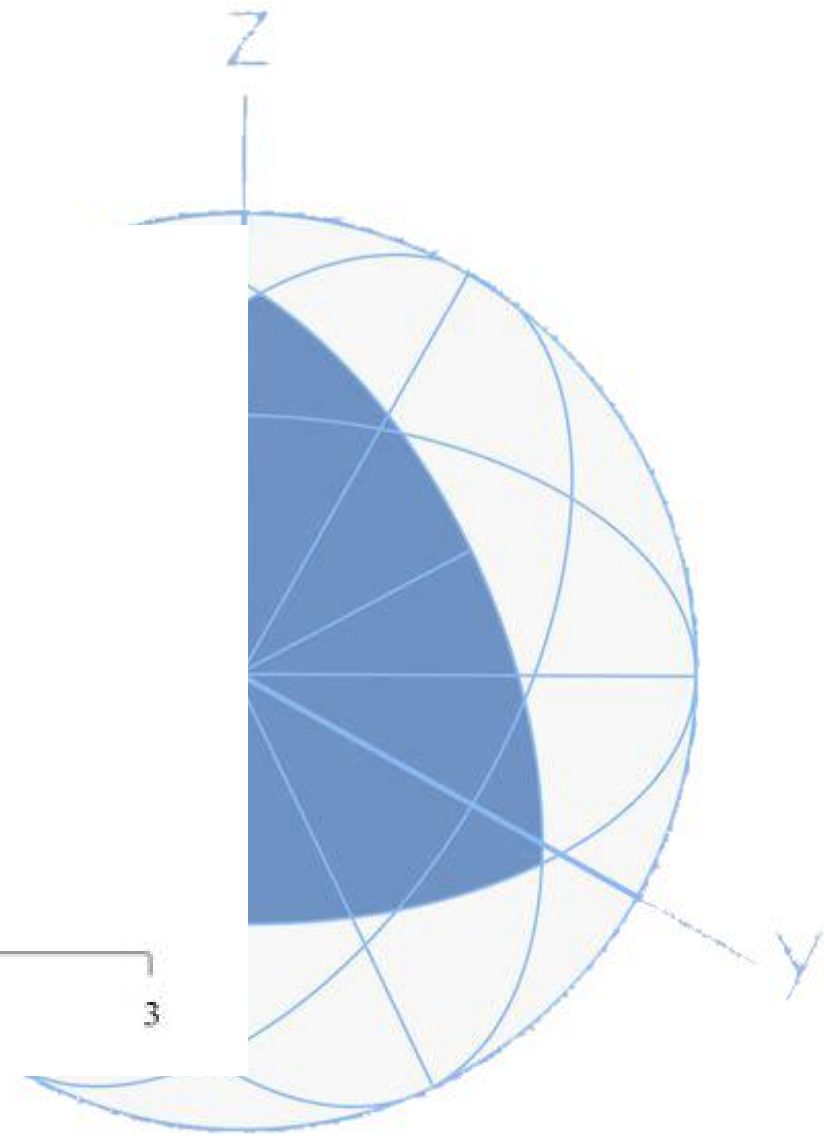
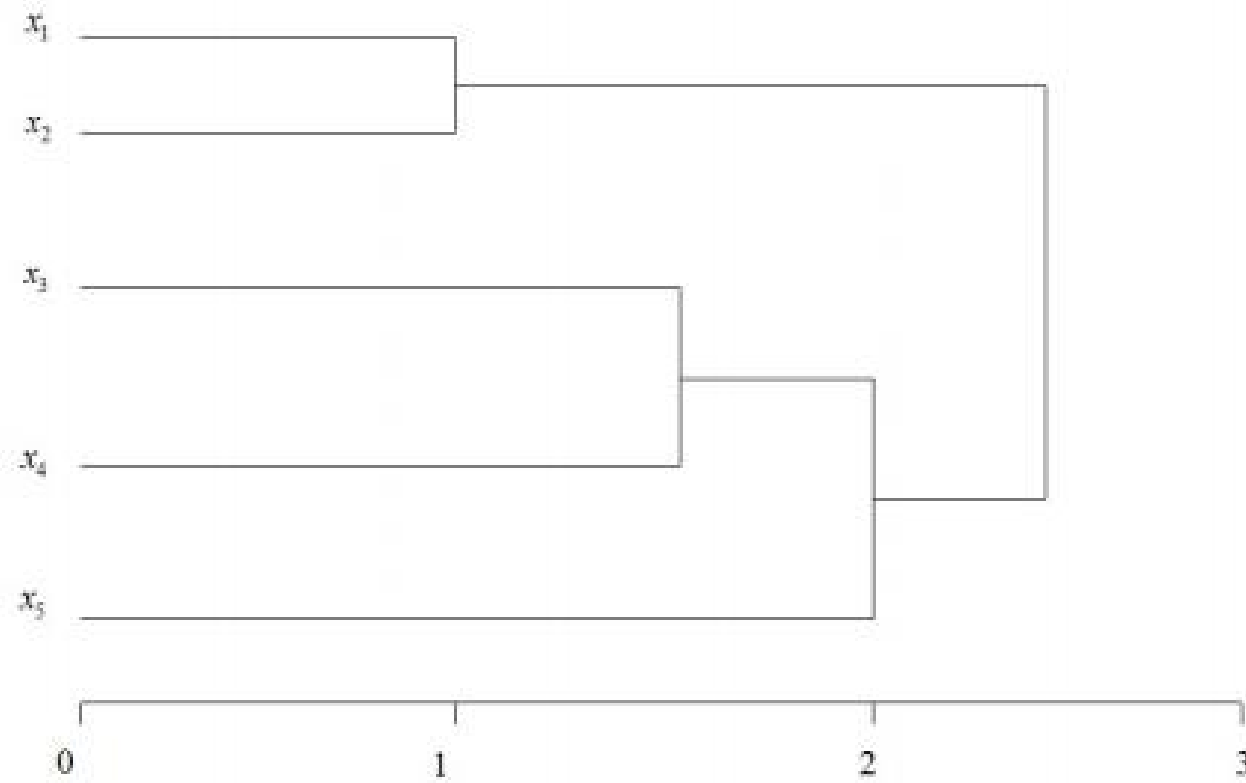
表 7.6 距离阵  $D^{(4)}$

	$G_6 = \{X_{(1)}, X_{(2)}\}$	$G_8 = \{X_5, X_{(3)}, X_{(4)}\}$
$G_6 = \{X_{(1)}, X_{(2)}\}$	0	[2.5]
$G_8 = \{X_5, X_{(3)}, X_{(4)}\}$		0

表 7.7 距离阵  $D^{(5)}$

	$G_{10} = \{X_{(1)}, X_{(2)}, X_{(3)}, X_{(4)}, X_{(5)}\}$
$G_{10} = \{X_{(1)}, X_{(2)}, X_{(3)}, X_{(4)}, X_{(5)}\}$	0







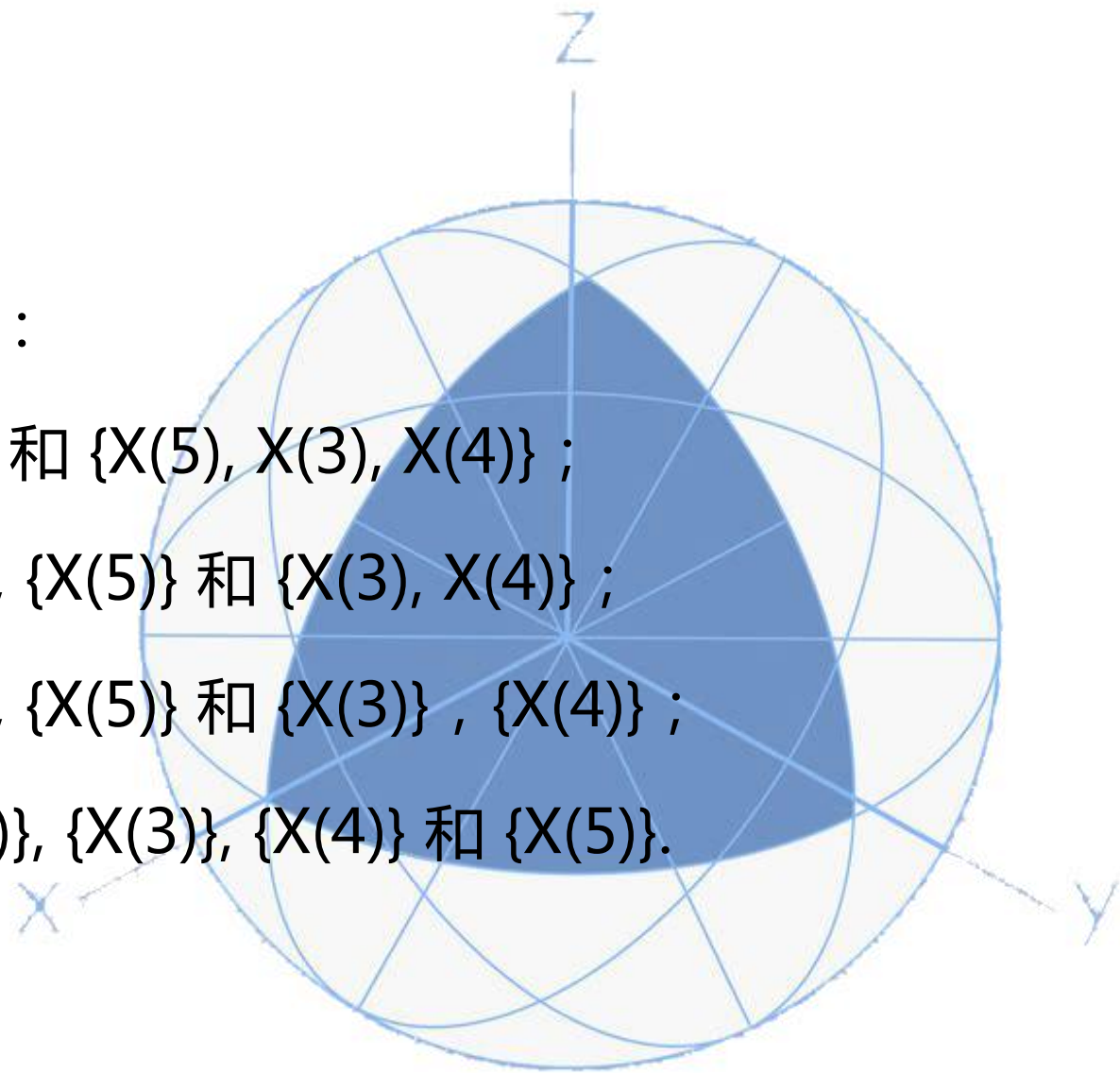
根据谱系聚类图可得到分类的结果：

若分为两类，则分为： $\{X(1), X(2)\}$  和  $\{X(5), X(3), X(4)\}$ ；

若分为三类，则分为： $\{X(1), X(2)\}$ ,  $\{X(5)\}$  和  $\{X(3), X(4)\}$ ；

若分为四类，则分为： $\{X(1), X(2)\}$ ,  $\{X(5)\}$  和  $\{X(3)\}$ ,  $\{X(4)\}$ ；

若分为五类，则分为： $\{X(1)\}$ ,  $\{X(2)\}$ ,  $\{X(3)\}$ ,  $\{X(4)\}$  和  $\{X(5)\}$ .



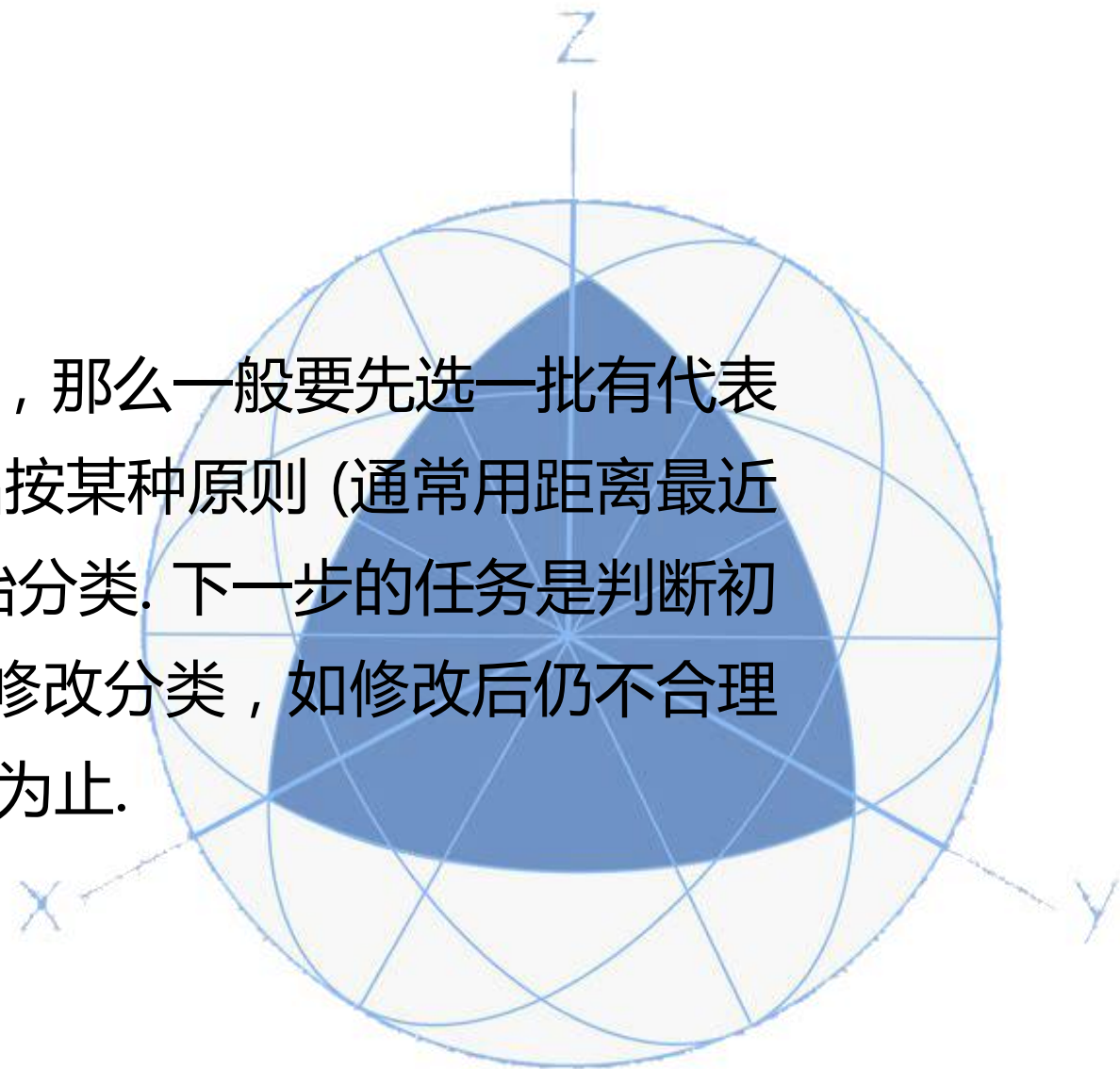




## 5、动态聚类法

它先粗糙的进行预分类 (初始分类), 那么一般要先选一批有代表性的样品点当凝聚点, 然后让样品按某种原则 (通常用距离最近原则) 向凝聚点会聚, 从而得到初始分类. 下一步的任务是判断初始分类是否合理. 如果不合理, 就修改分类, 如修改后仍不合理时, 就再一次修改分类, 直至合理为止.

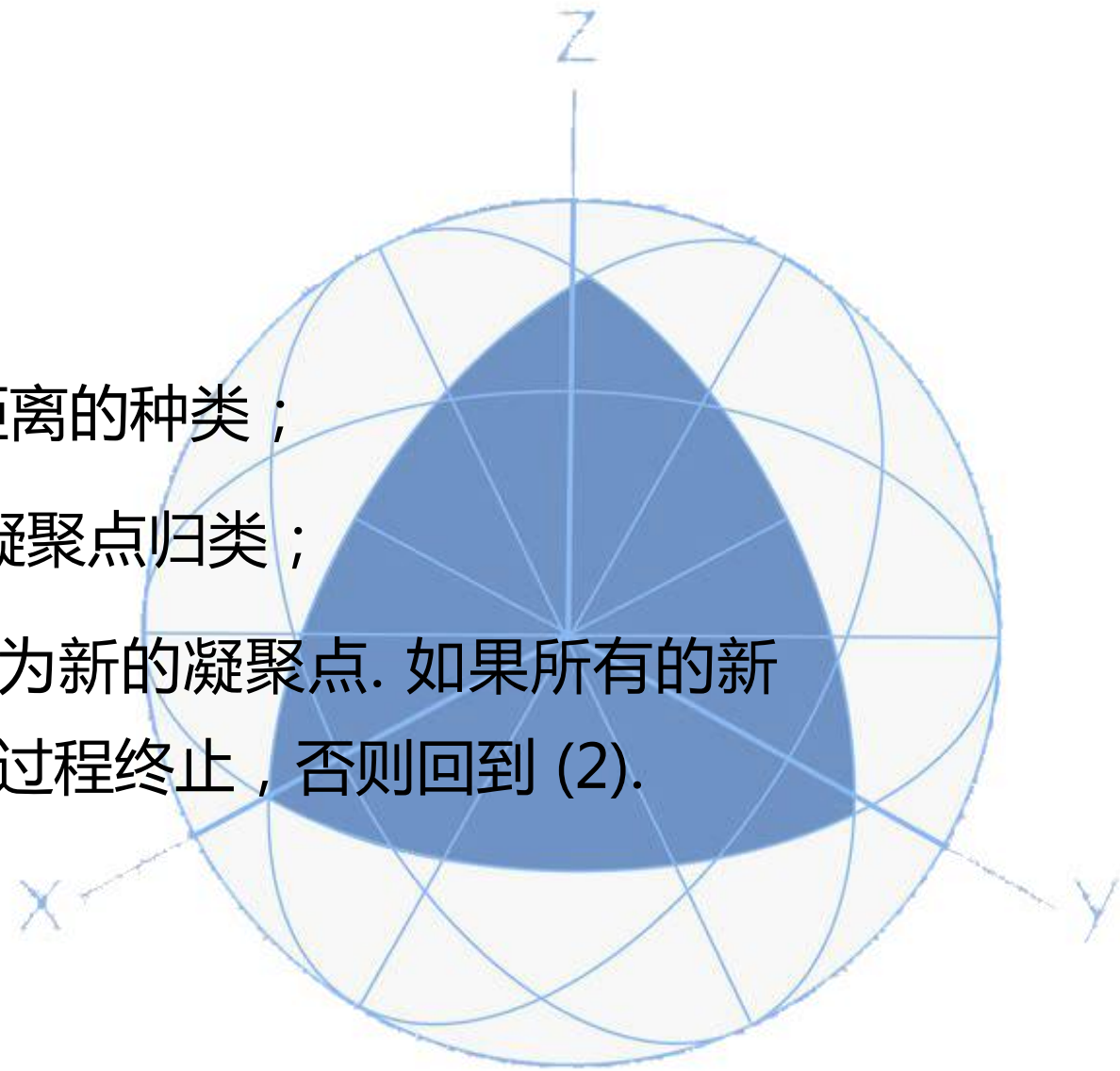
常见的聚类方法有 K-均值法.





按 K-均值法步骤:

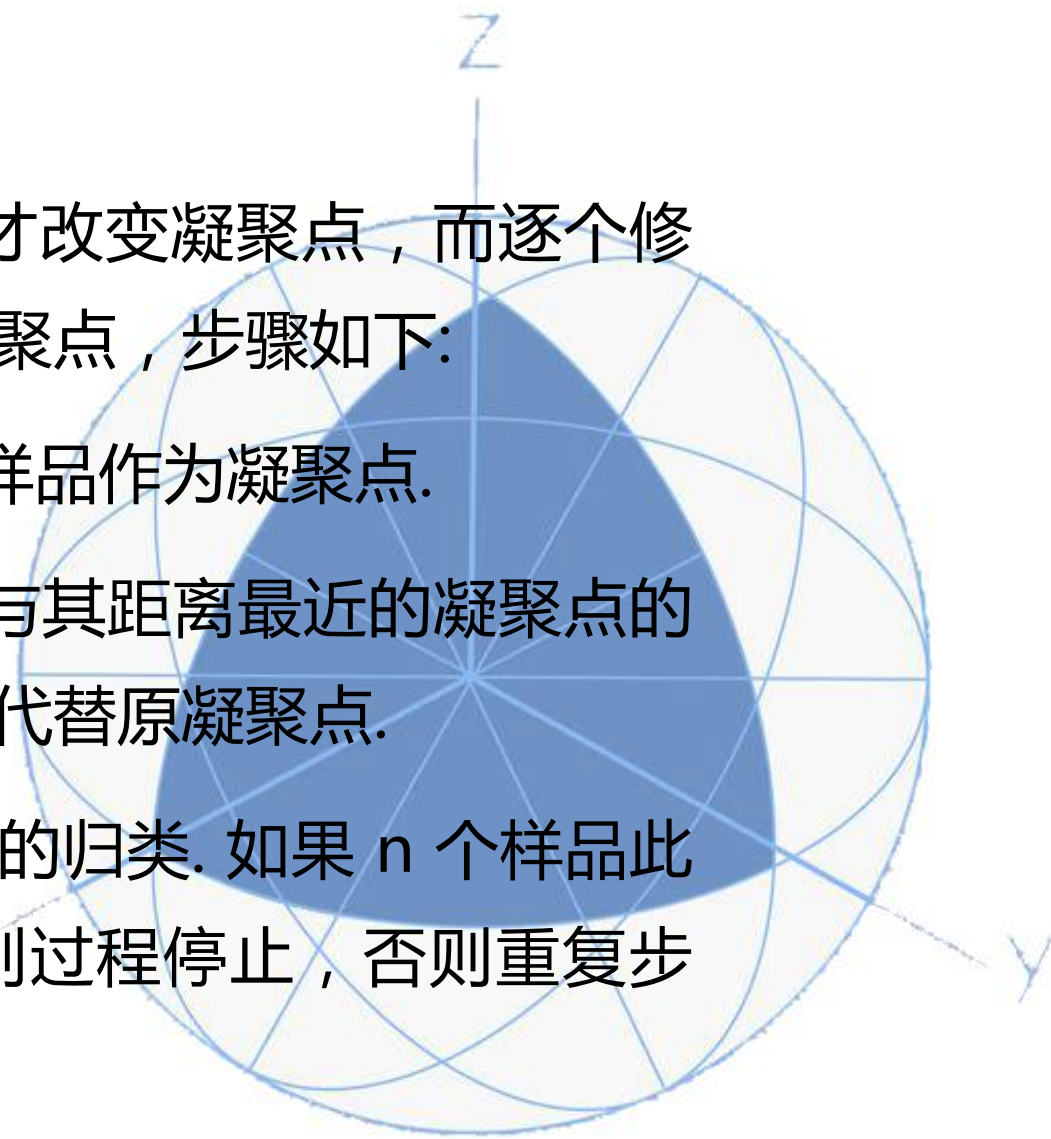
- (1) 选择一批凝聚点，并选定所用距离的种类；
- (2) 将所有样品按与其距离最近的凝聚点归类；
- (3) 计算每一类的重心，将重心作为新的凝聚点. 如果所有的新凝聚点与前一次的老凝聚点重合，过程终止，否则回到 (2).





按 K-均值法是等样品全部调整完毕后才改变凝聚点，而逐个修改法是每个样品一旦调整后立即改变凝聚点，步骤如下：

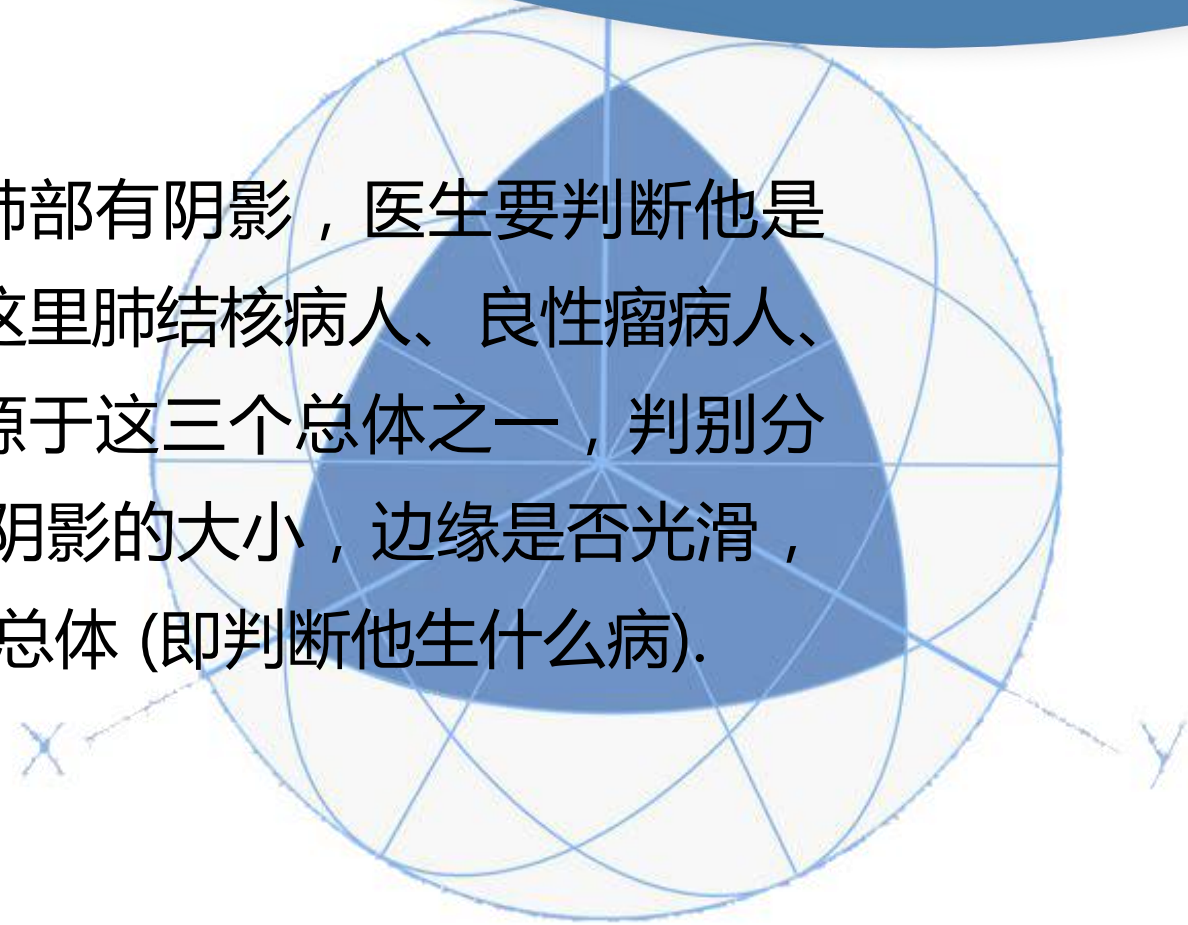
- (1) 人为地定出分类数目  $K$ ，取前  $K$  个样品作为凝聚点。
- (2) 将剩下的  $n - K$  个样品逐个的归入与其距离最近的凝聚点的那一类，随即计算该类重心，并用重心代替原凝聚点。
- (3) 将  $n$  个样品重新的按照步骤 2 逐个的归类。如果  $n$  个样品此时所属的类与原来的归类完全一样，则过程停止，否则重复步骤 3。





## 7.3 判别分析

例如：在医学诊断中，一个病人肺部有阴影，医生要判断他是肺结核、肺部良性肿瘤还是肺癌. 这里肺结核病人、良性瘤病人、肺癌病人组成三个总体，病人来源于这三个总体之一，判别分析的目的是通过测得病人的指标（阴影的大小，边缘是否光滑，体温多少……）来判断他应该属哪个总体（即判断他生什么病）.



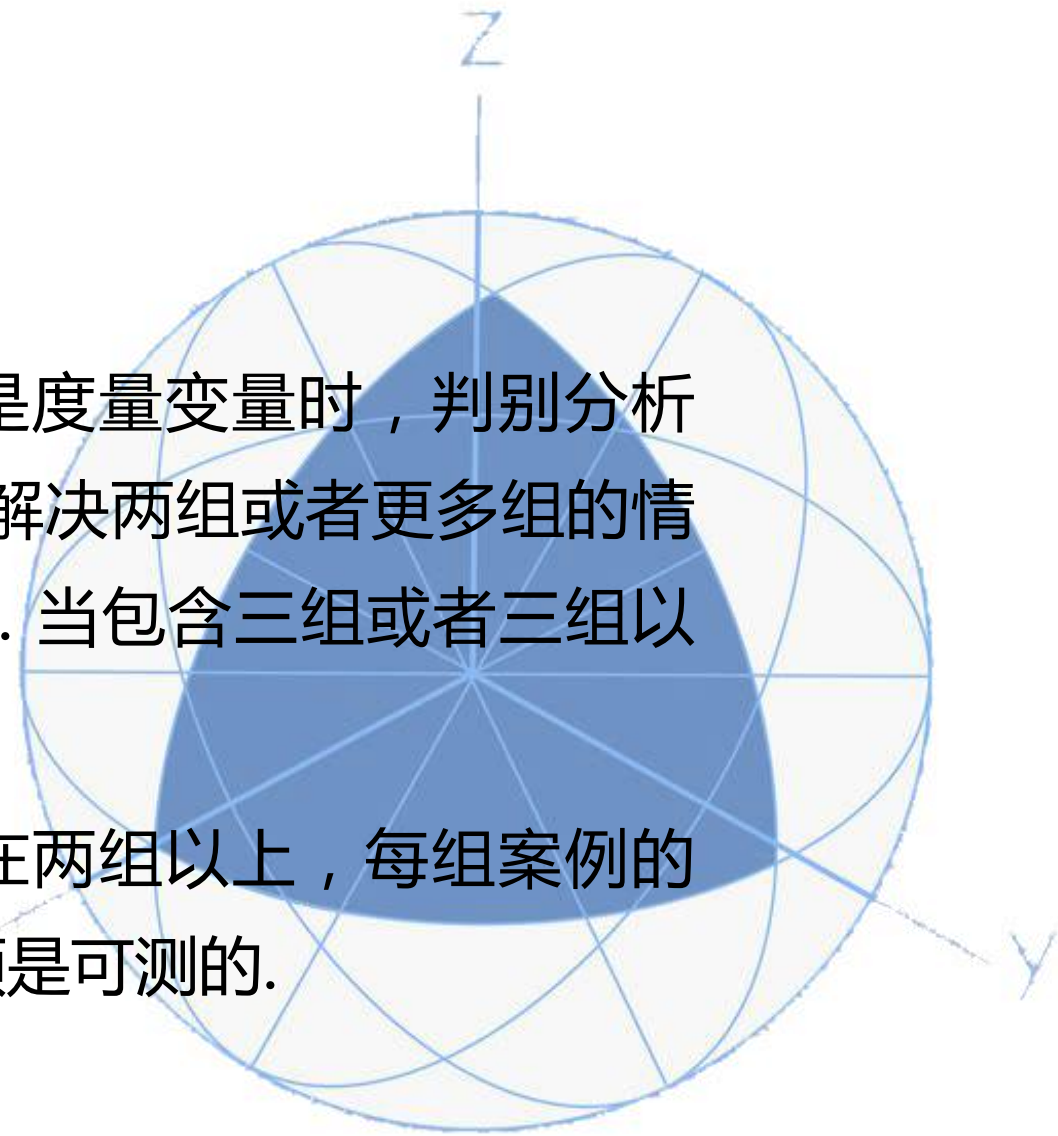




## 一、判别分析的基本思想

当被解释变量是属性变量而解释变量是度量变量时，判别分析是合适的统计分析方法. 判别分析能够解决两组或者更多组的情况. 当包含两组时，称作两组判别分析. 当包含三组或者三组以上时称作多组判别分析.

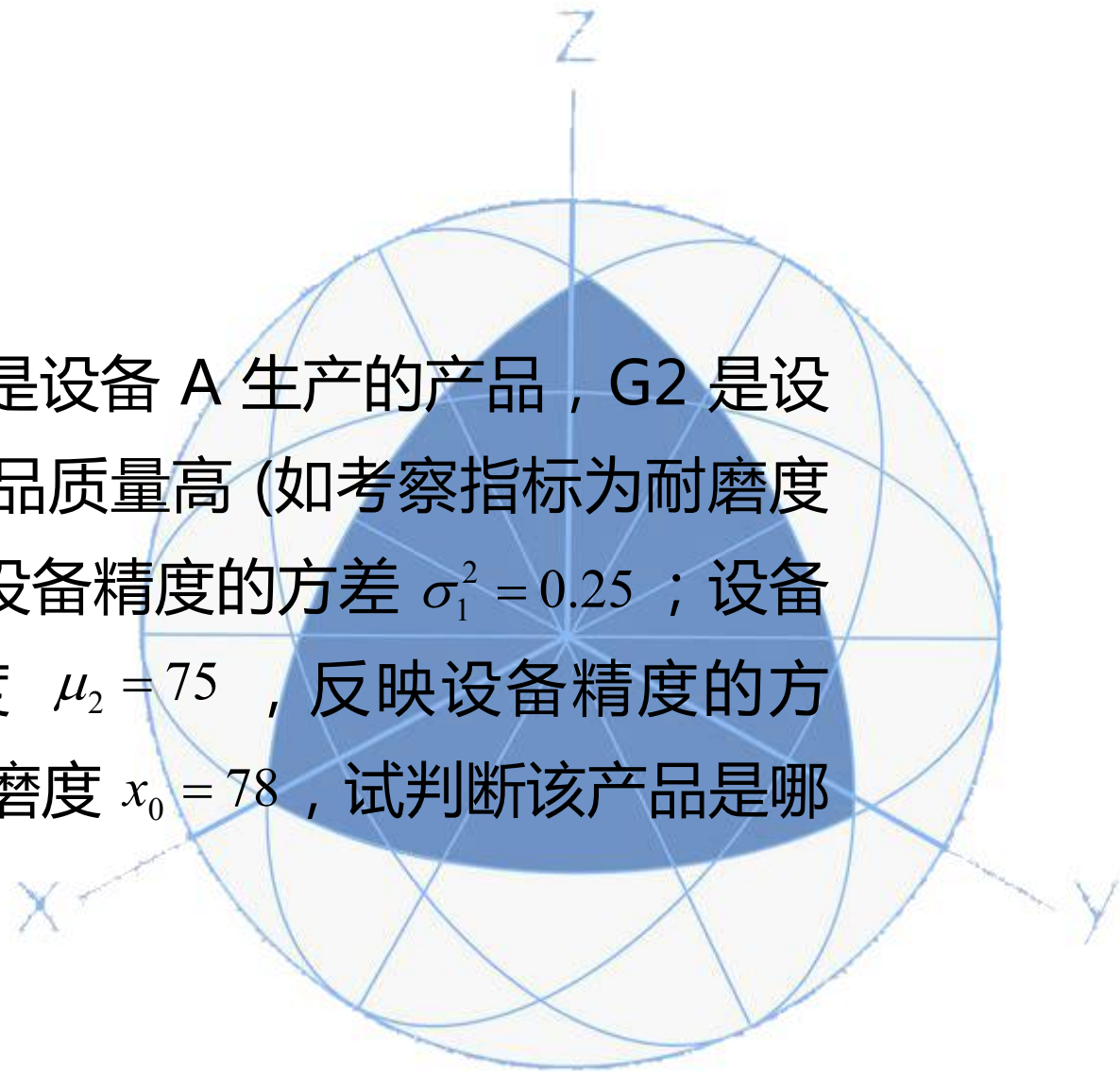
判别分析的最基本要求是：分组类型在两组以上，每组案例的规模必须至少在一个以上. 解释变量必须是可测的.





## 二、距离判别法

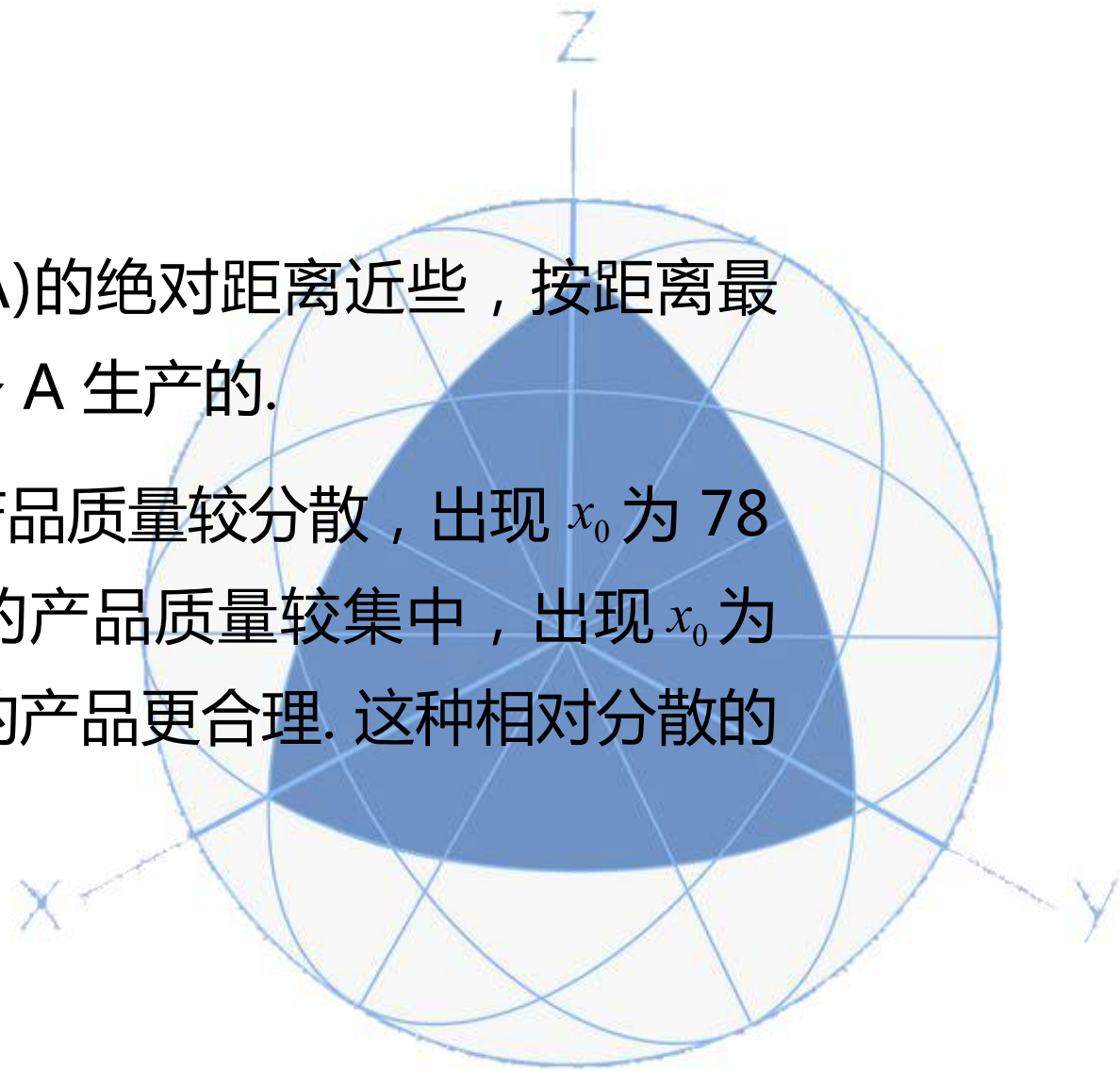
已知有两个类  $G_1$  和  $G_2$ ，比如  $G_1$  是设备 A 生产的产品， $G_2$  是设备 B 生产的同类产品. 设备 A 的产品质量高 (如考察指标为耐磨度  $X$ )，其平均耐磨度  $\mu_1 = 80$ ，反映设备精度的方差  $\sigma_1^2 = 0.25$ ；设备 B 的产品质量稍差，其平均耐磨度  $\mu_2 = 75$ ，反映设备精度的方差  $\sigma_2^2 = 4$  . 今有一产品  $x_0$ ，测得耐磨度  $x_0 = 78$ ，试判断该产品是哪一台设备生产的？





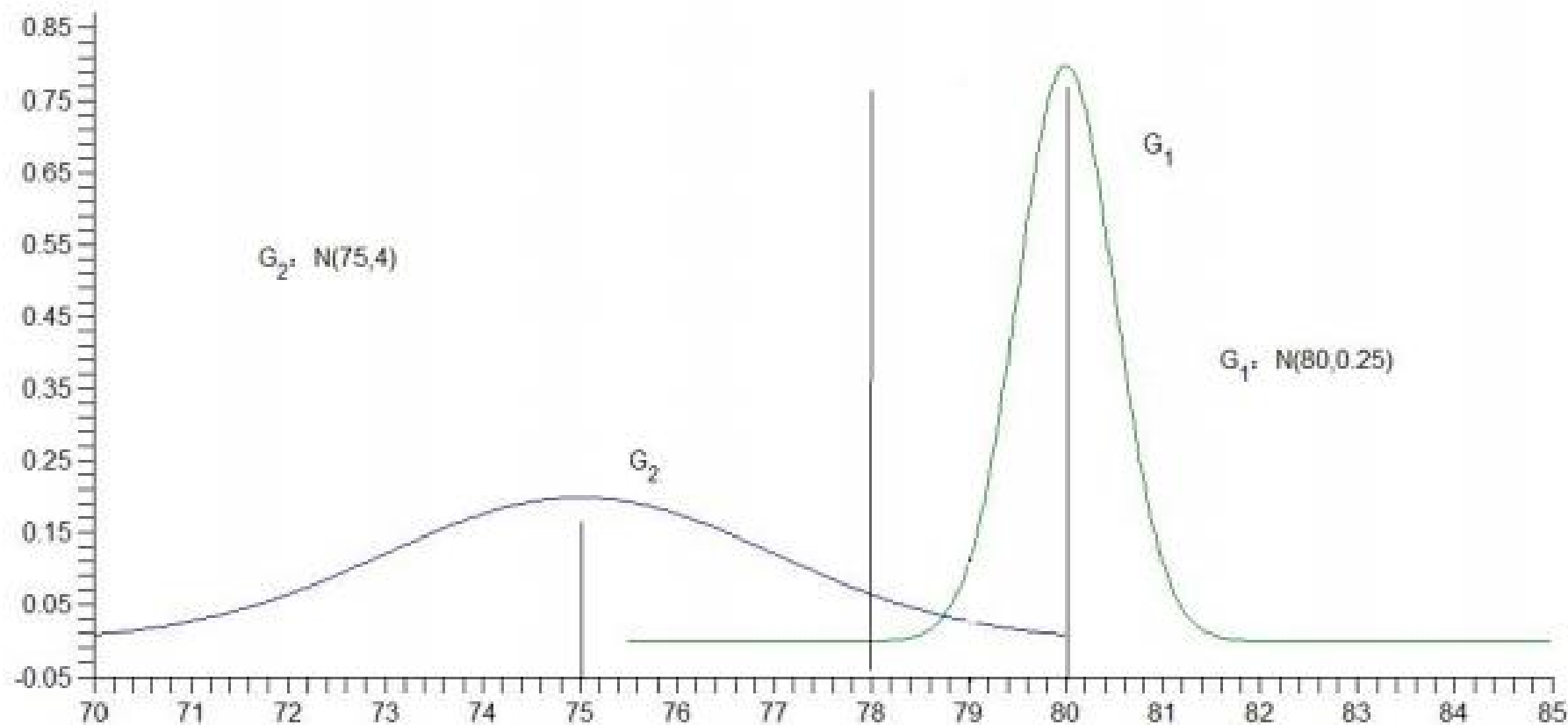
从绝对长度来看， $x_0$  与  $\mu_1$  (设备 A) 的绝对距离近些，按距离最近的原则应把该产品  $x_0$  判断为设备 A 生产的。

从概率观点来看，设备 B 生产的产品质量较分散，出现  $x_0$  为 78 的可能性仍较大；而设备 A 生产的产品质量较集中，出现  $x_0$  为 78 的可能性较小。判  $x_0$  为设备 B 的产品更合理。这种相对分散的距离我们称之为马氏距离。





z







## 1、马氏距离

**定义** 设总体  $G$  为  $m$  维总体 (考察  $m$  个指标), 均值向量为

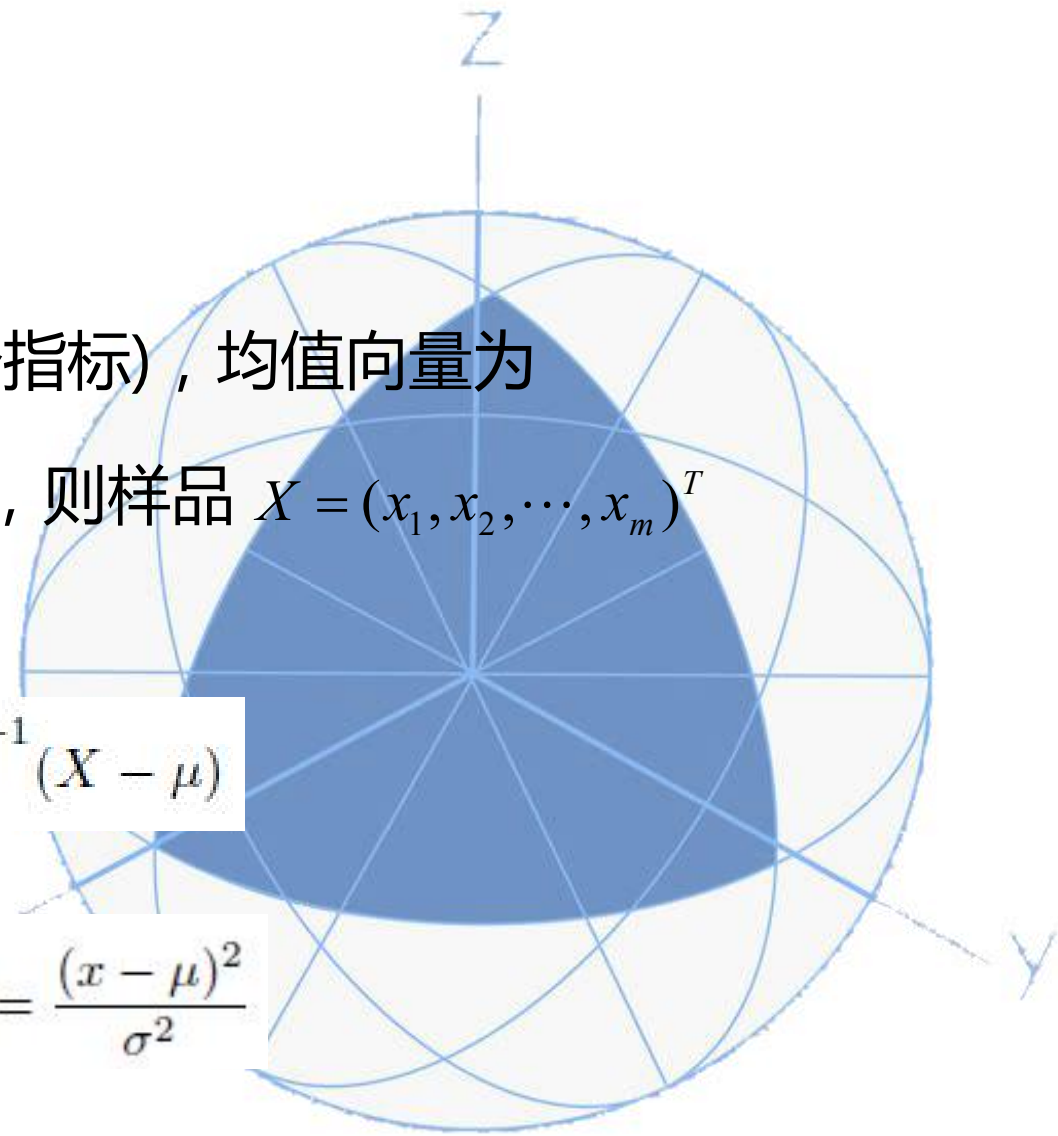
$\mu = (\mu_1, \mu_2, \dots, \mu_m)^T$ , 协方差阵为  $\Sigma = (\sigma_{ij})$ , 则样品  $X = (x_1, x_2, \dots, x_m)^T$

与总体  $G$  的马氏距离定义为:

$$d^2(X, G) = (X - \mu)^T \Sigma^{-1} (X - \mu)$$

当  $m = 1$  时,

$$d^2(x, G) = \frac{(x - \mu)^T (x - \mu)}{\sigma^2} = \frac{(x - \mu)^2}{\sigma^2}$$





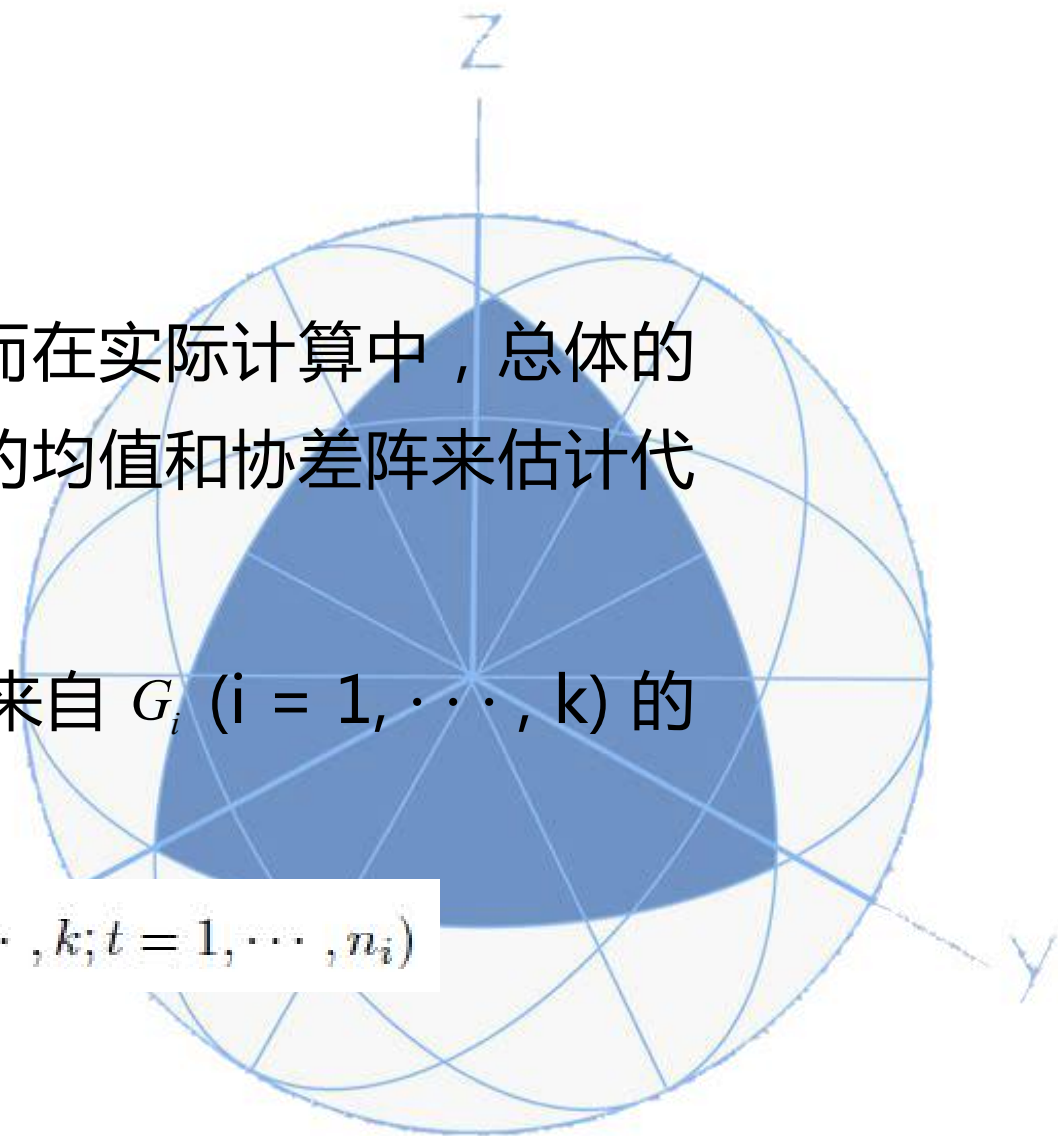
## 2、样本的特征量

马氏距离会用到总体的均值和方差，而在实际计算中，总体的均值和方差是未知的，因此常用样本的均值和协差阵来估计代替总体的协差阵.

设有  $k$  个总体  $G_i$  ( $i = 1, \dots, k$ )，已知来自  $G_i$  ( $i = 1, \dots, k$ ) 的训练样本为:

$$X_t^{(i)} = (x_{t1}^{(i)}, x_{t2}^{(i)}, \dots, x_{tm}^{(i)})' \quad (i = 1, \dots, k; t = 1, \dots, n_i)$$

其中  $n_i$  是取自  $G_i$  的样品个数，





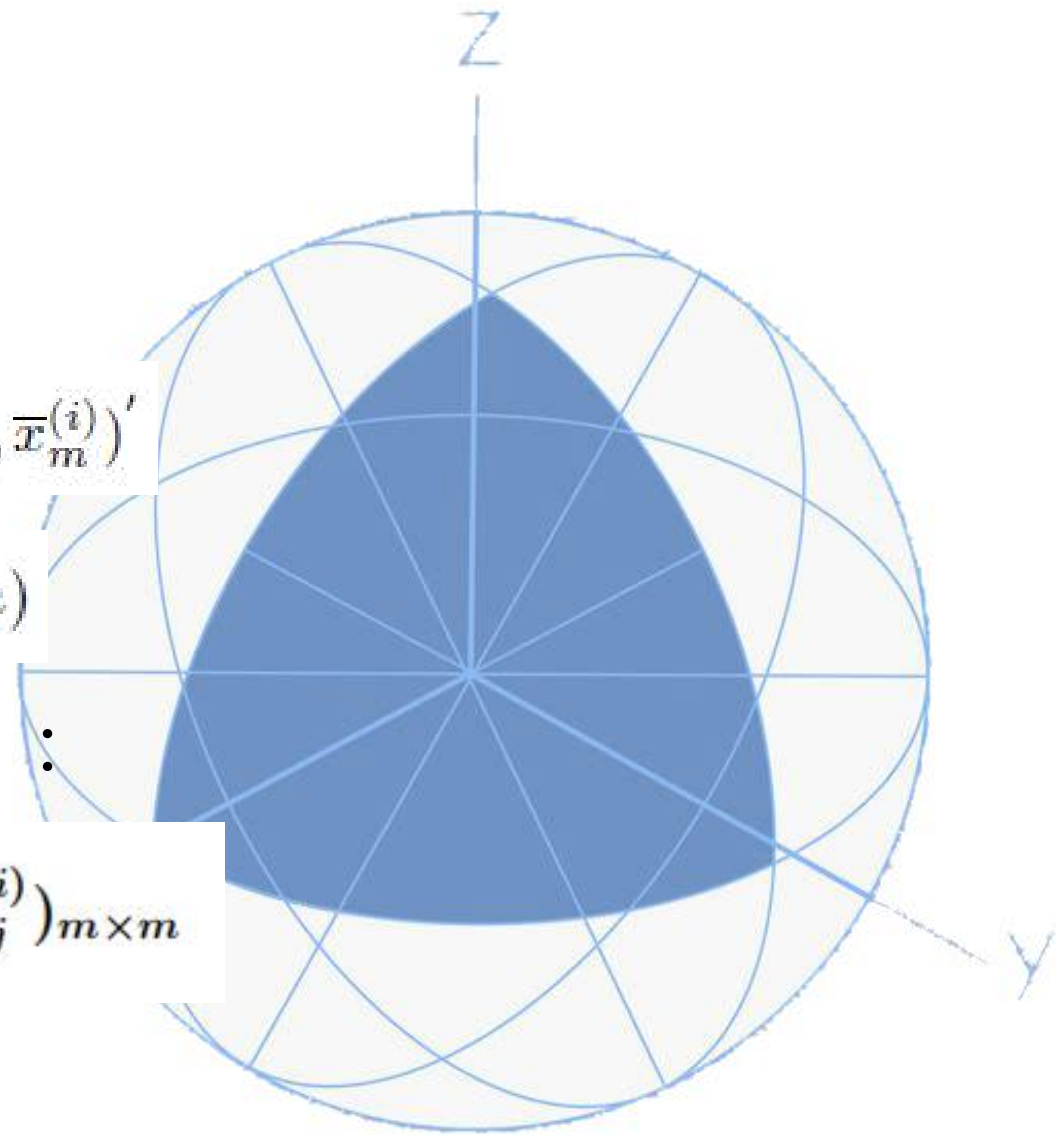
则均值向量  $\mu_i$  的估计量为：

$$\bar{X}^{(i)} = (\bar{x}_1^{(i)}, \bar{x}_2^{(i)}, \dots, \bar{x}_m^{(i)})'$$

其中  $\bar{x}_j^{(i)} = \frac{1}{n_i} \sum_{t=1}^{n_i} x_{tj}^{(i)} (j = 1, 2, \dots, m)$

总体  $G_i$  的协方差阵  $\Sigma_i$  的估计值  $S_i$ ：

$$S_i = \frac{1}{n_i - 1} A_i = (s_{lj}^{(i)})_{m \times m}$$



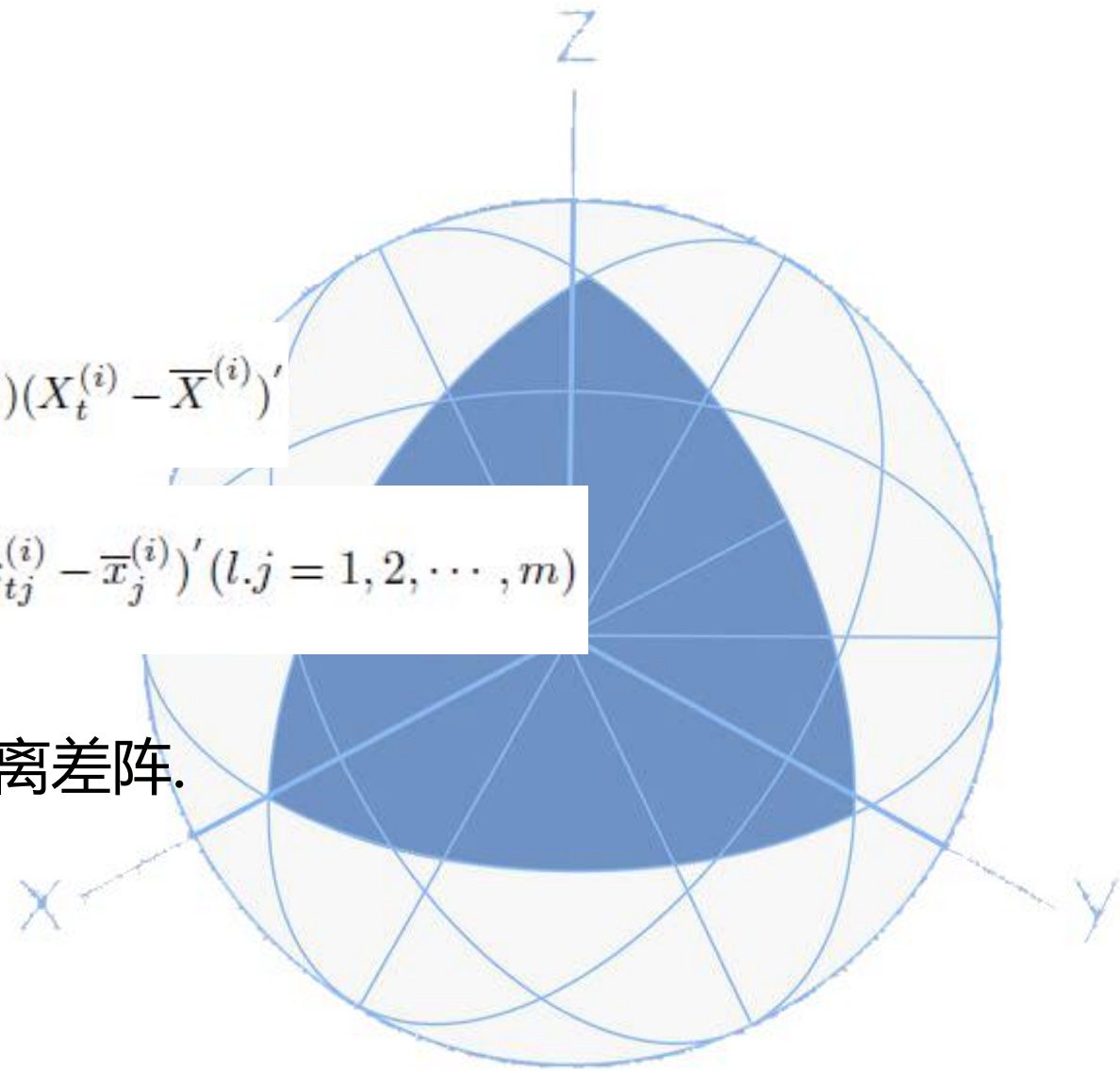


其中

$$A_i = \sum_{t=1}^{n_i} (X_t^{(i)} - \bar{X}^{(i)})(X_t^{(i)} - \bar{X}^{(i)})'$$

$$s_{lj}^{(i)} = \frac{1}{n_i - 1} \sum_{t=1}^{n_i} (x_{tl}^{(i)} - \bar{x}_l^{(i)})(x_{tj}^{(i)} - \bar{x}_j^{(i)})' (l, j = 1, 2, \dots, m)$$

$S_i$  称为组内协差阵， $A_i$  称为组内离差阵。





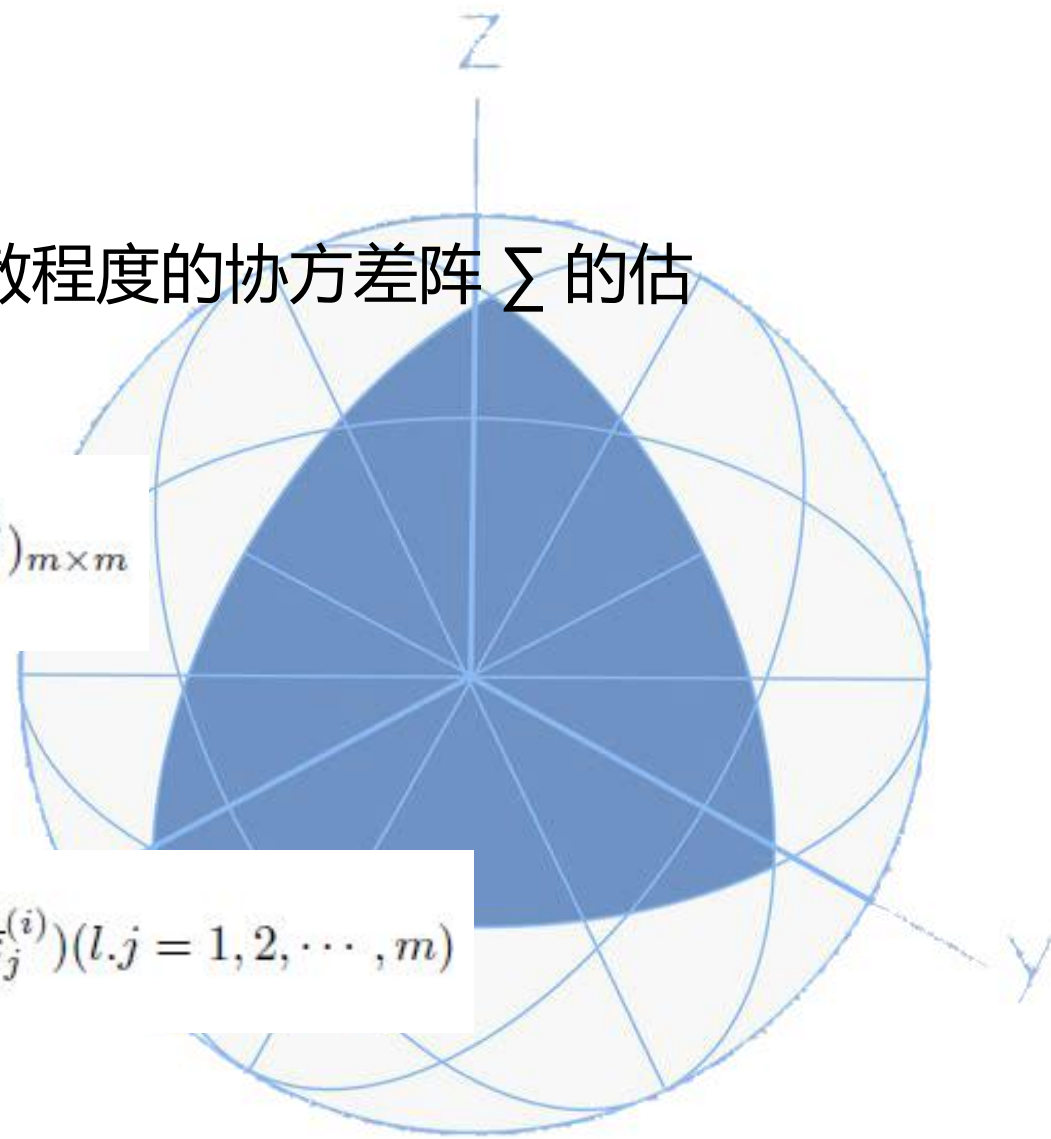


当假定  $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k = \Sigma$  时反应分散程度的协方差阵  $\Sigma$  的估计  $S$  为

$$S = \frac{1}{n-k} \sum_{i=1}^k A_i = (s_{lj}^{(i)})_{m \times m}$$

其中

$$s_{lj}^{(i)} = \frac{1}{n-k} \sum_{i=1}^k \sum_{t=1}^{n_i} (x_{tl}^{(i)} - \bar{x}_l^{(i)})(x_{tj}^{(i)} - \bar{x}_j^{(i)}) (l, j = 1, 2, \cdots, m)$$



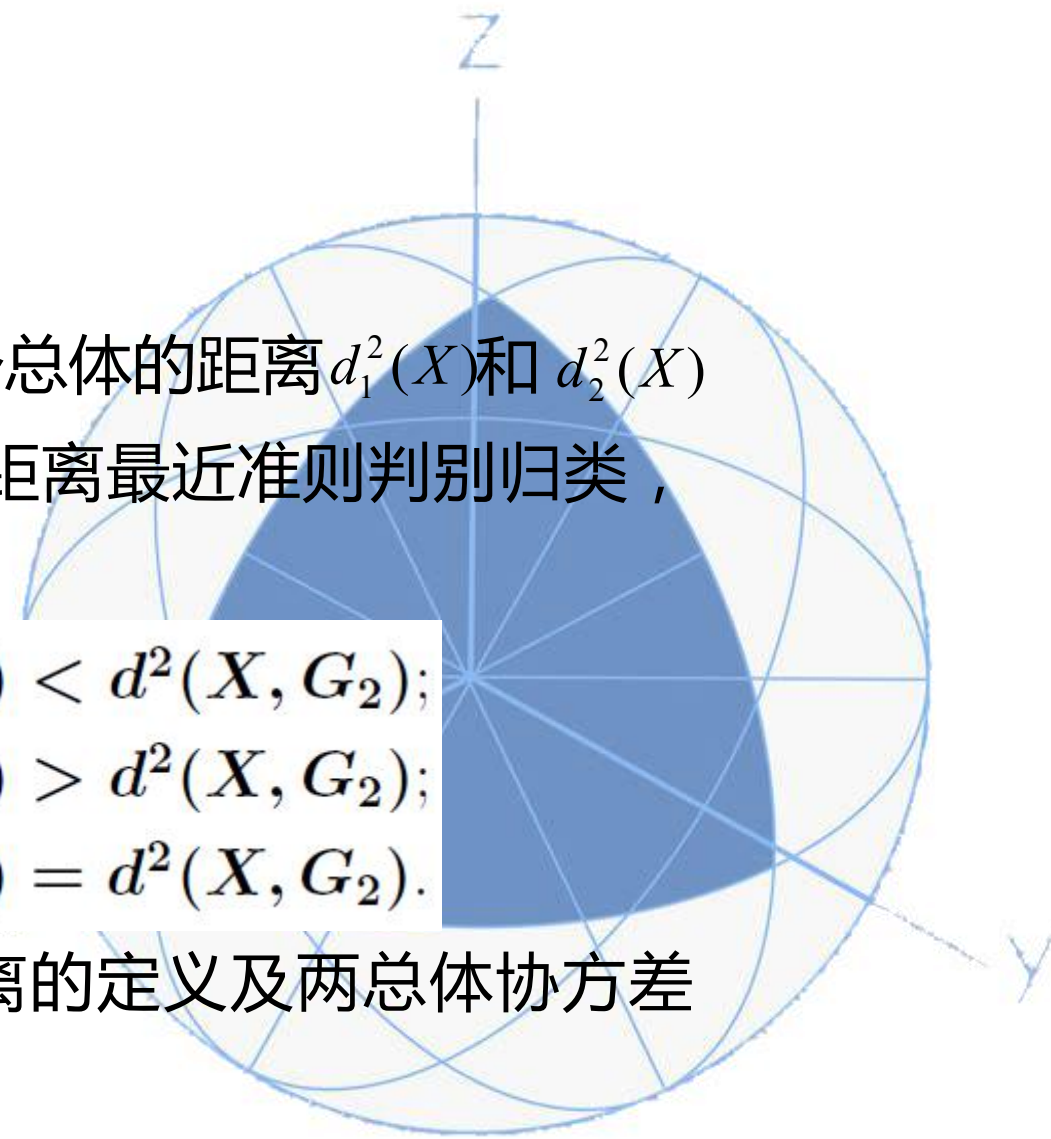


### 3、两总体判别(1)当 $\Sigma_1 = \Sigma_2$

最直观的想法是分别计算样品  $X$  到两个总体的距离  $d_1^2(X)$  和  $d_2^2(X)$  (或者记为  $d^2(X, G_1)$  和  $d^2(X, G_2)$ ), 并按距离最近准则判别归类, 即判别准则为:

$$\begin{cases} \text{判 } X \in G_1, & \text{当 } d^2(X, G_1) < d^2(X, G_2); \\ \text{判 } X \in G_2, & \text{当 } d^2(X, G_1) > d^2(X, G_2); \\ \text{待判,} & \text{当 } d^2(X, G_1) = d^2(X, G_2). \end{cases}$$

这里的距离指马氏距离, 利用马氏距离的定义及两总体协方差阵相等的假设化简





z  
|

$$\begin{aligned}d^2(X, G_i) &= (X - \bar{X}^{(i)})' S^{-1} (X - \bar{X}^{(i)}) \\&= X' S^{-1} X - 2[X' (S^{-1} \bar{X}^{(i)}) - \frac{1}{2}(\bar{X}^{(i)})' S^{-1} \bar{X}^{(i)}] \\&= X' S^{-1} X - 2Y_i(X)\end{aligned}$$

其中

$$Y_i(X) = X' (S^{-1} \bar{X}^{(i)}) - \frac{1}{2}(\bar{X}^{(i)})' S^{-1} \bar{X}^{(i)}$$

因为函数  $Y_i(X)$  是  $X$  的线性函数 ( $i = 1, 2$ )，故称为**线性判别函数**.



考察这两个马氏距离之差，经计算可得：

$$d_2^2(X) - d_1^2(X) = (X - \bar{X}^{(2)})' S^{-1} (X - \bar{X}^{(2)}) - (X - \bar{X}^{(1)})' S^{-1} (X - \bar{X}^{(1)})$$

$$W(X) = Y_1(X) - Y_2(X)$$

令  $X^* = \frac{1}{2}(\bar{X}^{(1)} + \bar{X}^{(2)})$ ,  $a = S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)})$ ，则

$W(X) = a'(X - X^*)$ ， $W(X)$  为  $X$  的线性判别函数， $a$  为判别系数。



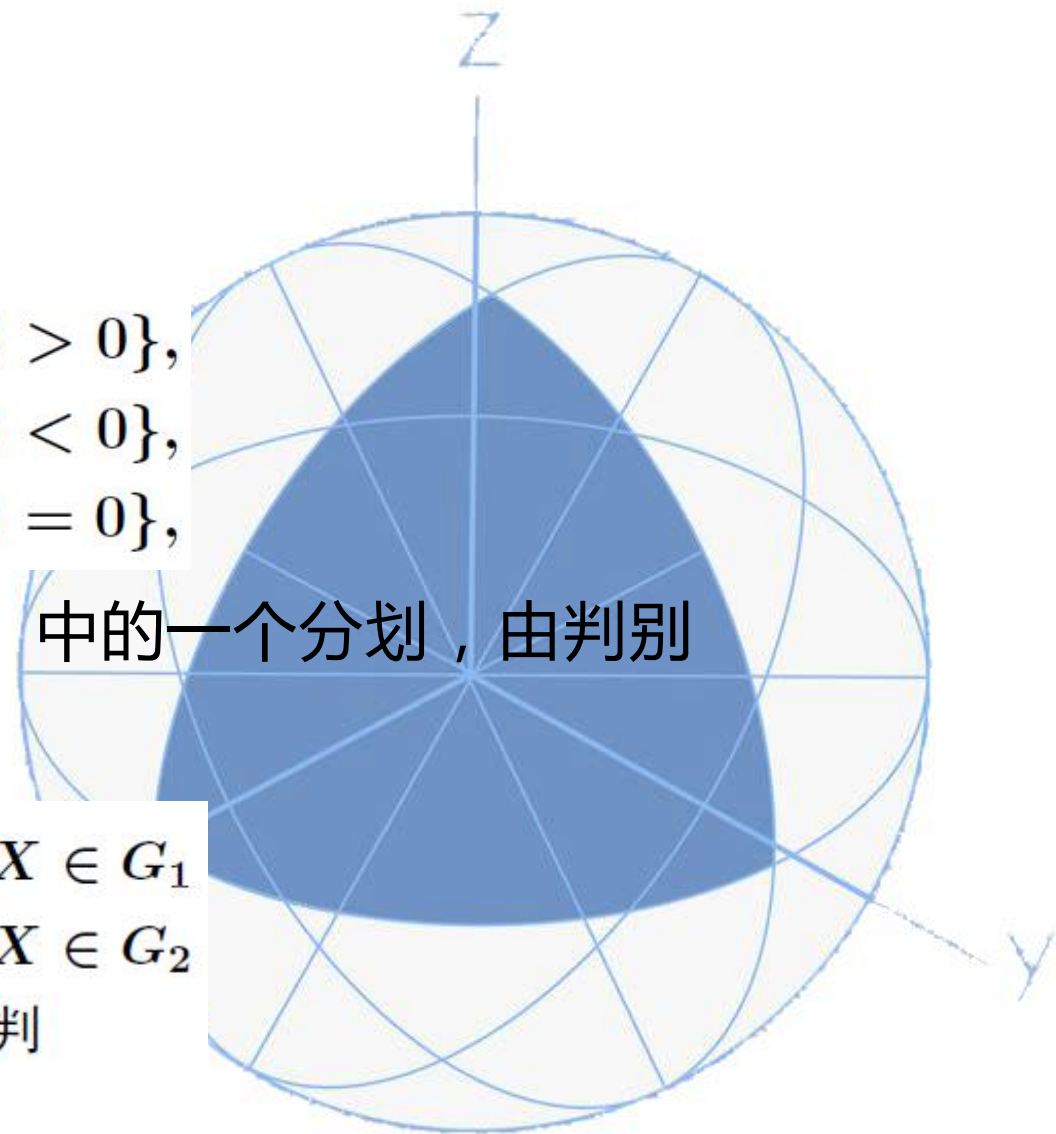


$W(X)$  把  $m$  维空间  $R^m$  划分为 3 部分：

$$\begin{cases} D_1 = \{X : W(X) > 0\}, \\ D_2 = \{X : W(X) < 0\}, \\ D_0 = \{X : W(X) = 0\}, \end{cases}$$

显然，判别方法的最终结果是得到  $R^m$  中的一个分划，由判别函数  $W(X)$  得到分划  $D_1, D_2, D_0$

$$\begin{cases} \text{当样品 } X \text{ 落入 } D_1 & \text{判 } X \in G_1 \\ \text{当样品 } X \text{ 落入 } D_2 & \text{判 } X \in G_2 \\ \text{当样品 } X \text{ 落入 } D_0 & \text{待判} \end{cases}$$



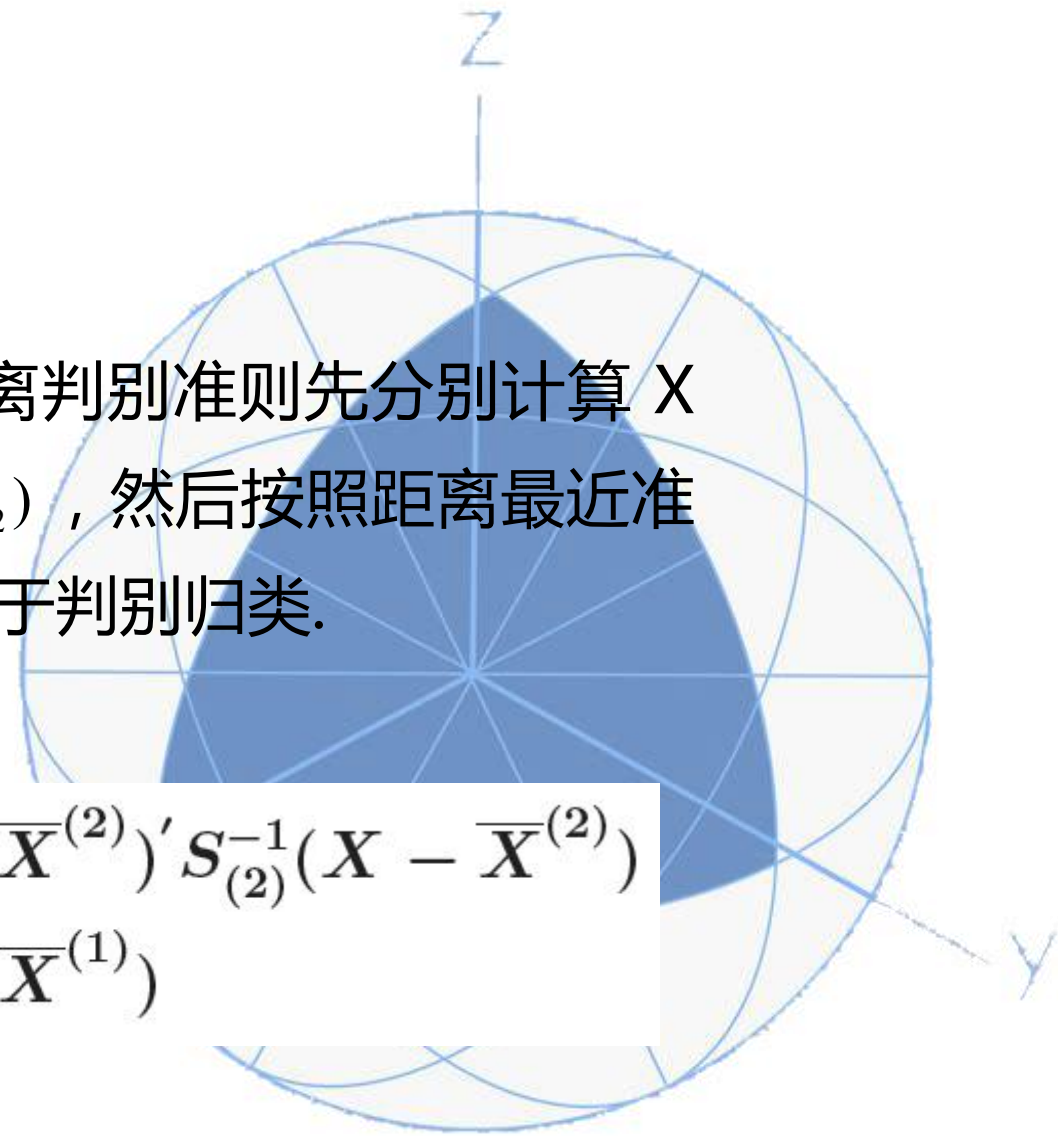


(2) 当  $\Sigma_1 \neq \Sigma_2$

当两总体协方差矩阵不相等时，按距离判别准则先分别计算  $X$  到两个总体的距离  $d^2(X, G_1)$  和  $d^2(X, G_2)$ ，然后按照距离最近准则判别归类，或者计算判别函数，并用于判别归类。

令

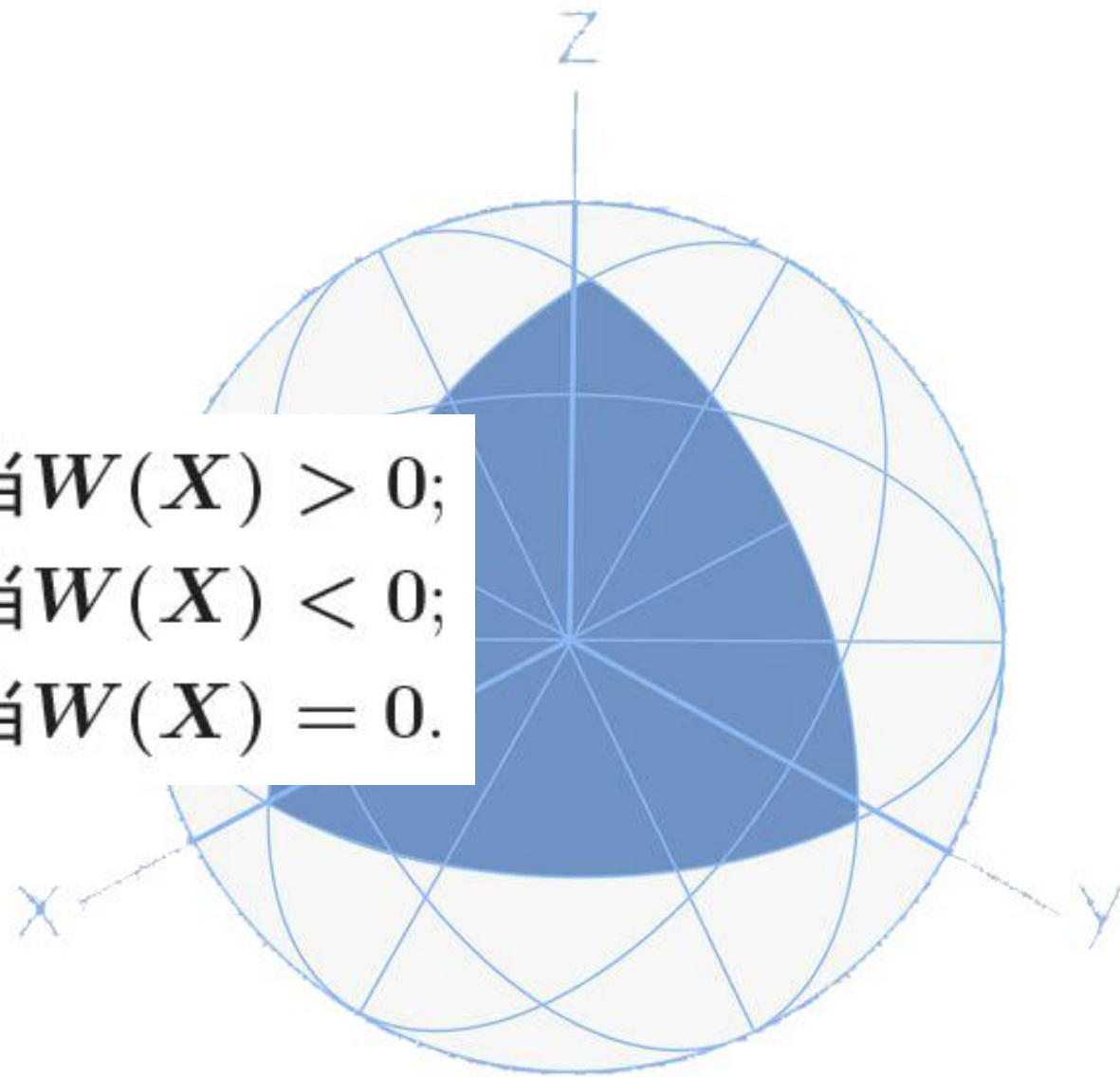
$$W(X) = d_2^2(X) - d_1^2(X) = (X - \bar{X}^{(2)})' S_{(2)}^{-1} (X - \bar{X}^{(2)}) - (X - \bar{X}^{(1)})' S_{(1)}^{-1} (X - \bar{X}^{(1)})$$





判别准则为：

$$\left\{ \begin{array}{ll} \text{判 } X \in G_1, & \text{当 } W(X) > 0; \\ \text{判 } X \in G_2, & \text{当 } W(X) < 0; \\ \text{待判,} & \text{当 } W(X) = 0. \end{array} \right.$$





## 4、多总体的距离判别

设有  $k(k > 2)$  个  $m$  维总体： $G_1, G_2, \dots, G_k$ . 它们的均值，协方差阵分别为： $\mu_i, \Sigma_i$  ( $i = 1, 2, \dots, k$ ). 对任给定的  $m$  维样品

$X = (x_1, x_2, \dots, x_m)'$ , 要判断它来自哪个总体.

多总体的情况，按距离最近的准则对  $X$  进行判别归类时，首先计算样品  $X$  到  $k$  个总体的马氏距离  $d^2(X, G_i)$  ( $i = 1, 2, \dots, k$ ). 然后进行比较，把  $X$  判归距离最小的那个总体.

设  $i = l$  时，若  $d_l^2(X) = \min_{i=1,2,\dots,k} \{d_i^2(X)\}$ ，则  $X \in G_l$

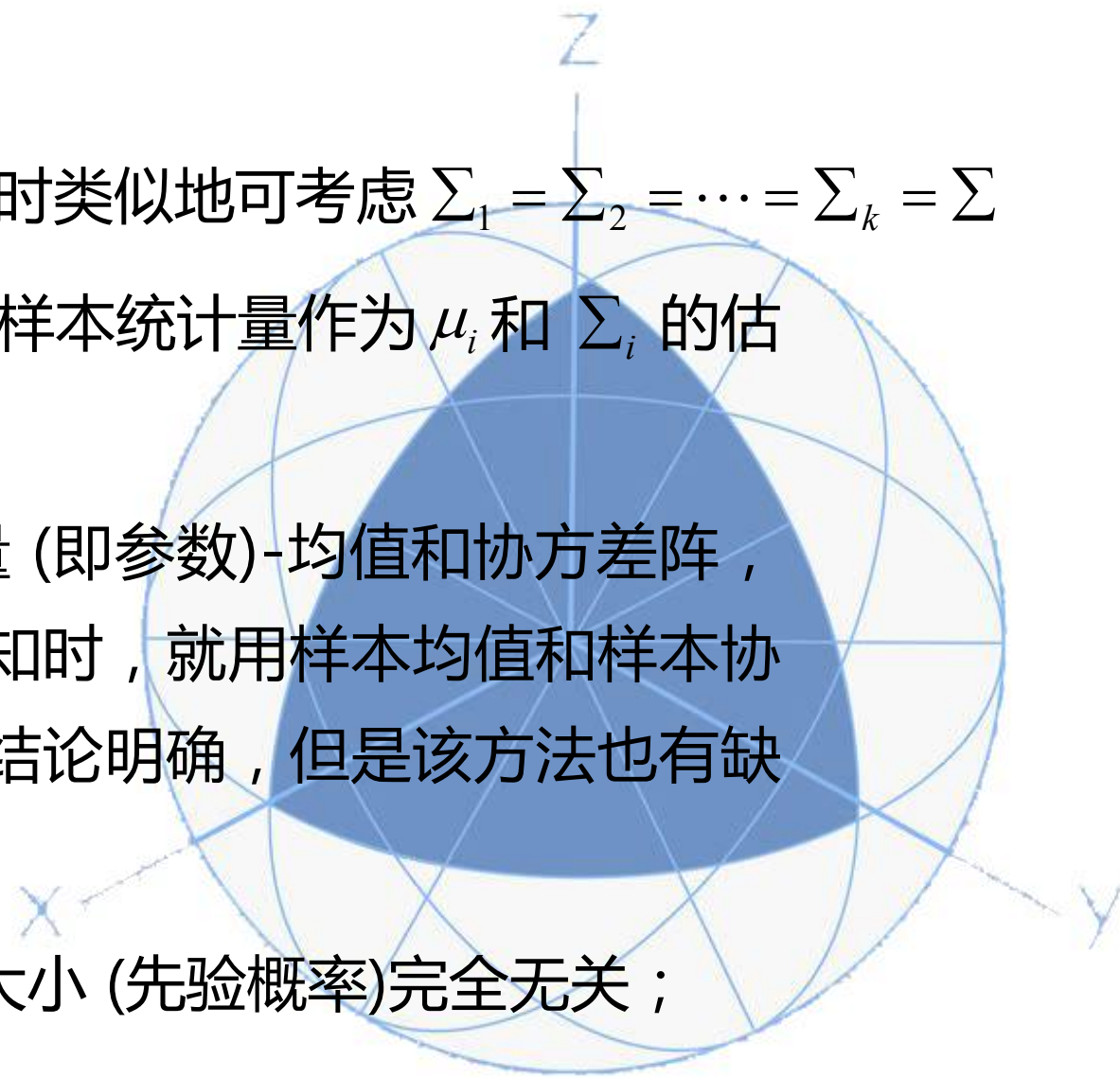




计算马氏距离  $d_i^2$  ( $i = 1, 2, \dots, k$ ) 时类似地可考虑  $\Sigma_1 = \Sigma_2 = \dots = \Sigma_k = \Sigma$  或者  $\Sigma_i$  不全相等的两种情况. 并用样本统计量作为  $\mu_i$  和  $\Sigma_i$  的估计.

距离判别只要求知道总体的特征量 (即参数)-均值和协方差阵, 不涉及总体的分布类型. 当参数未知时, 就用样本均值和样本协方差阵来估计. 距离判别方法简单结论明确, 但是该方法也有缺点:

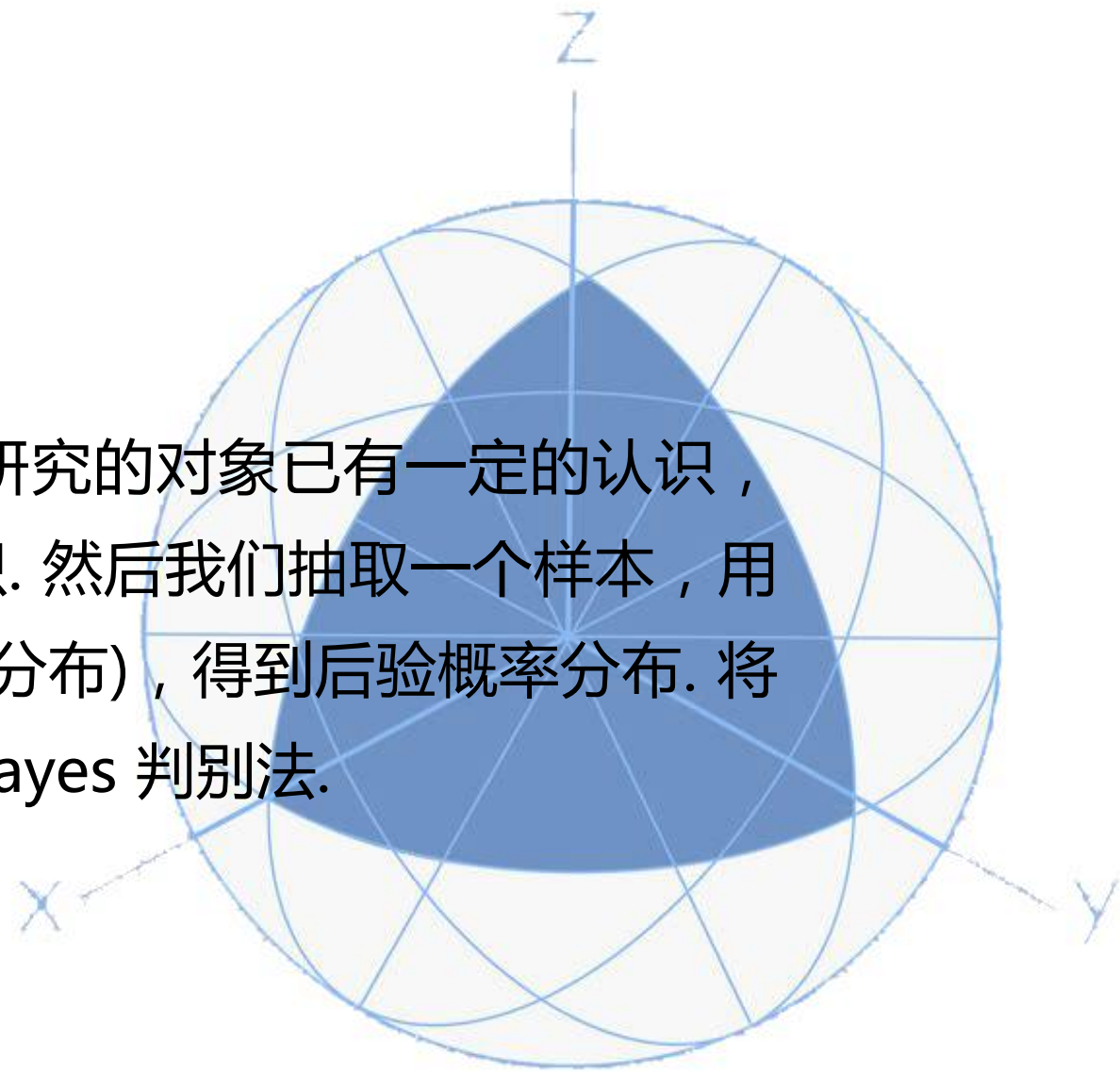
- (1) 该判别方法与各总体出现机会大小 (先验概率) 完全无关;
- (2) 判别方法没有考虑错判造成的损失.





### 三、Bayes 判别法

Bayes 的统计思想总是假定对所研究的对象已有一定的认识，常用先验概率分布来描述这种认识. 然后我们抽取一个样本，用样本来修正已有的认识（先验概率分布），得到后验概率分布. 将 Bayes 思想用于判别分析就得到 Bayes 判别法.

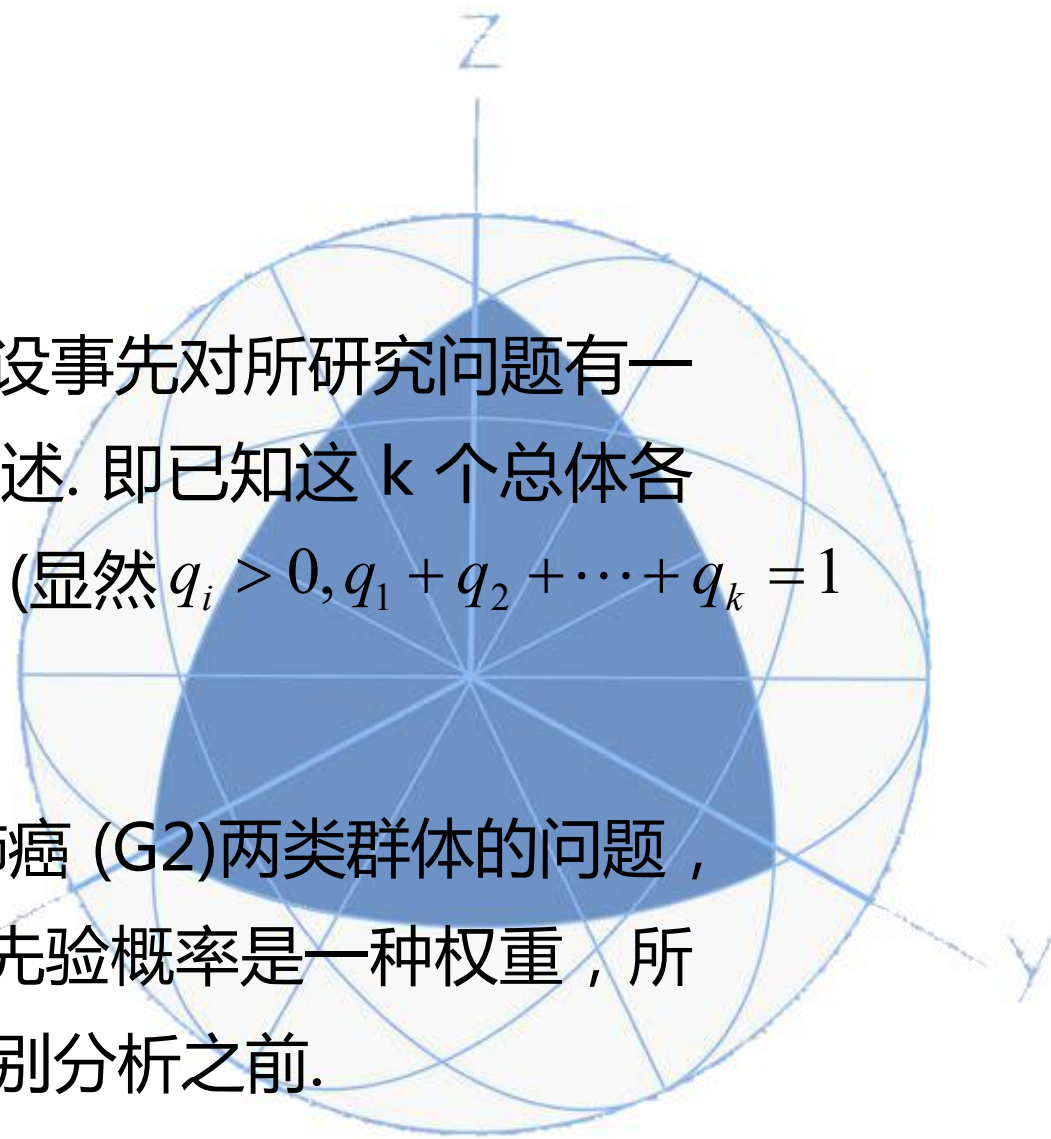




## 1、先验概率

设有  $k$  个  $m$  维总体： $G_1, G_2, \dots, G_k$  . 假设事先对所研究问题有一定的认识，这种认识常用先验概率来描述. 即已知这  $k$  个总体各自出现的概率 (先验概率) 为  $q_1, q_2, \dots, q_k$  (显然  $q_i > 0, q_1 + q_2 + \dots + q_k = 1$  ).

比如研究人群中得肺癌 ( $G_1$ ) 和没有得肺癌 ( $G_2$ ) 两类群体的问题，由长期经验知： $q_1 = 0.001, q_2 = 0.999$  . 先验概率是一种权重，所谓“先验”是指先于我们抽取样品做判别分析之前.



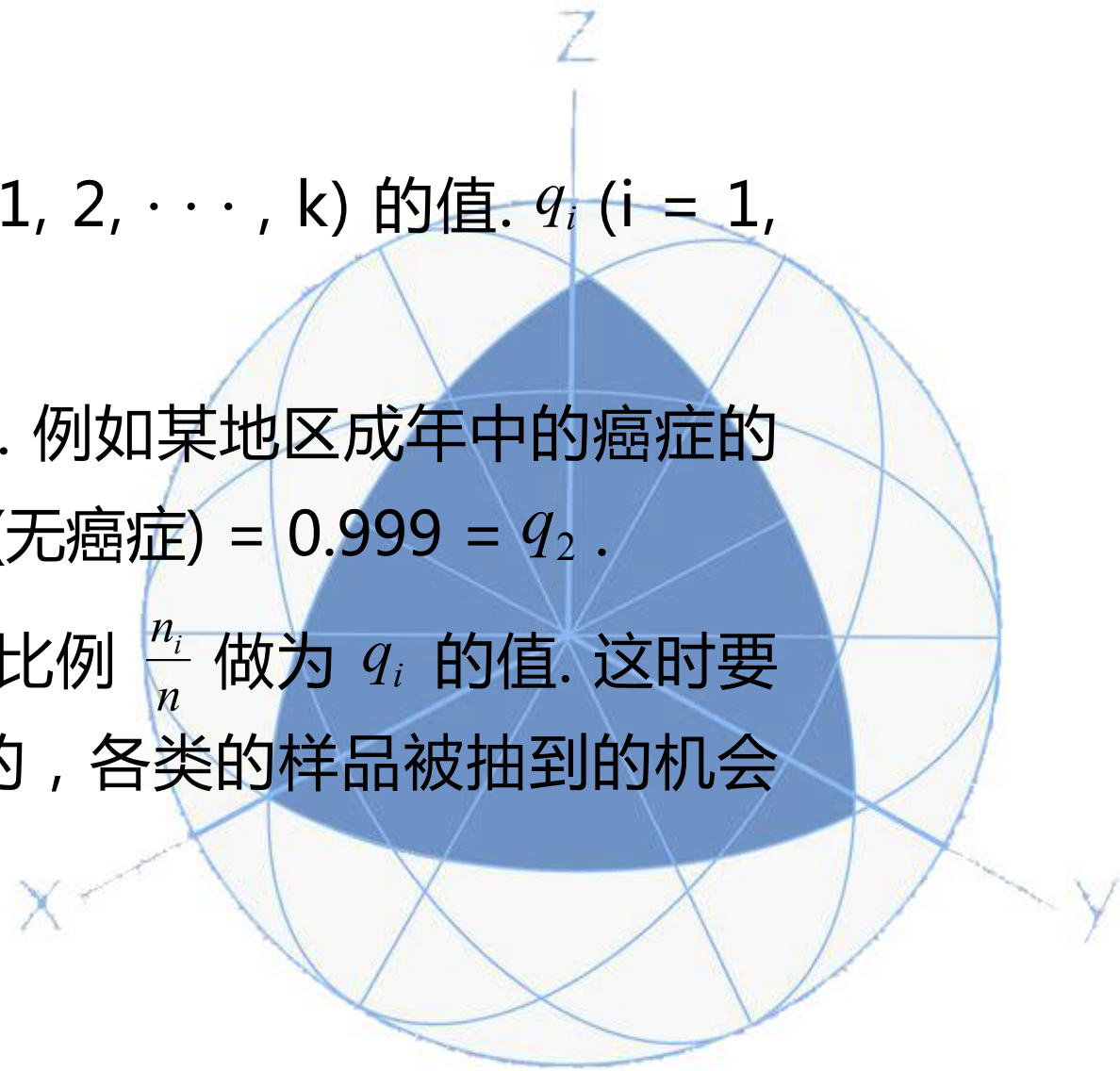


Bayes 判别准则要求给出  $q_i$  ( $i = 1, 2, \dots, k$ ) 的值.  $q_i$  ( $i = 1, 2, \dots, k$ ) 的赋值方法有以下几种:

(a) 利用历史资料及经验进行估计. 例如某地区成年中的癌症的概率为  $P(\text{癌症}) = 0.001 = q_1$ , 而  $P(\text{无癌症}) = 0.999 = q_2$ .

(b) 利用训练样本中各类样品占的比例  $\frac{n_i}{n}$  做为  $q_i$  的值. 这时要求训练样本是通过随机抽样得到的, 各类的样品被抽到的机会大小就是先验概率.

(c) 假定  $q_1 = q_2 = \dots = q_k = \frac{1}{k}$ .



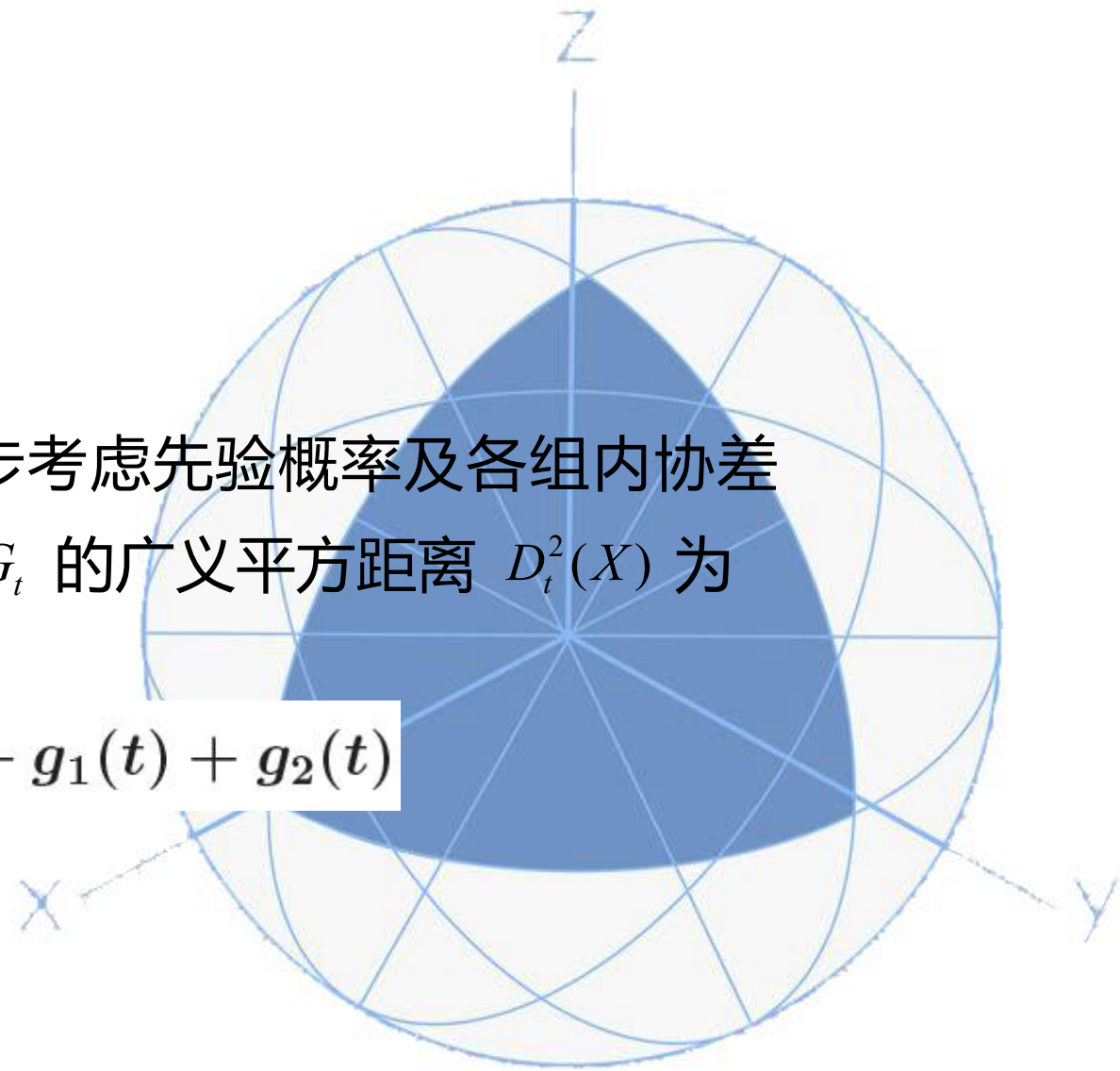




## 2、广义平方距离

在马氏距离判别的基础上，进一步考虑先验概率及各组内协差阵的不同，可定义样品  $X$  到总体  $G_t$  的广义平方距离  $D_t^2(X)$  为

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



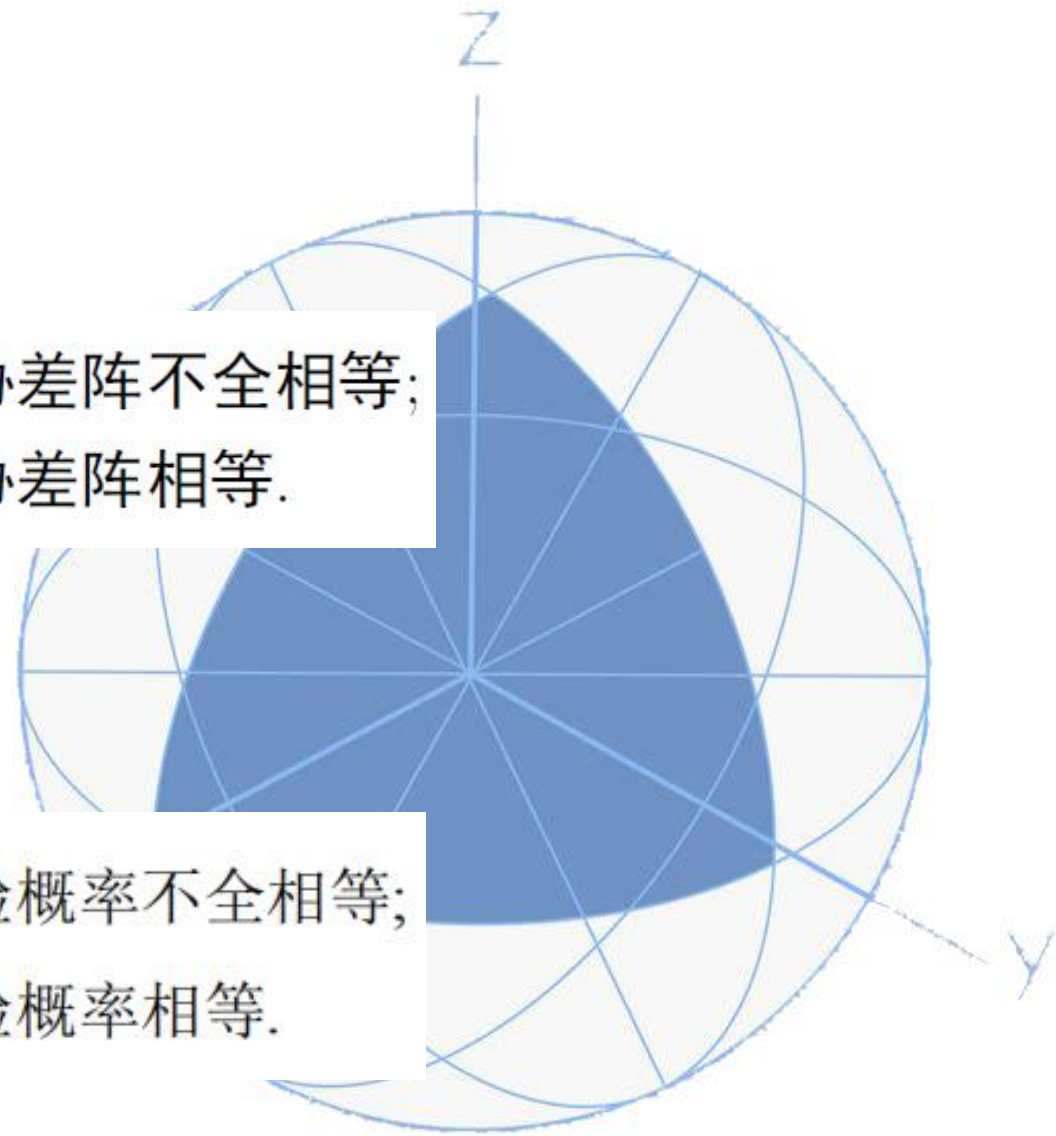


其中

$$g_1(t) = \begin{cases} \log |S_t|, & \text{若组内协差阵不全相等;} \\ 0, & \text{若组内协差阵相等.} \end{cases}$$

( $S_t$  为第  $t$  类的组内样本协差阵)

$$g_2(t) = \begin{cases} -2 \ln |q_t|, & \text{若先验概率不全相等;} \\ 0, & \text{若先验概率相等.} \end{cases}$$



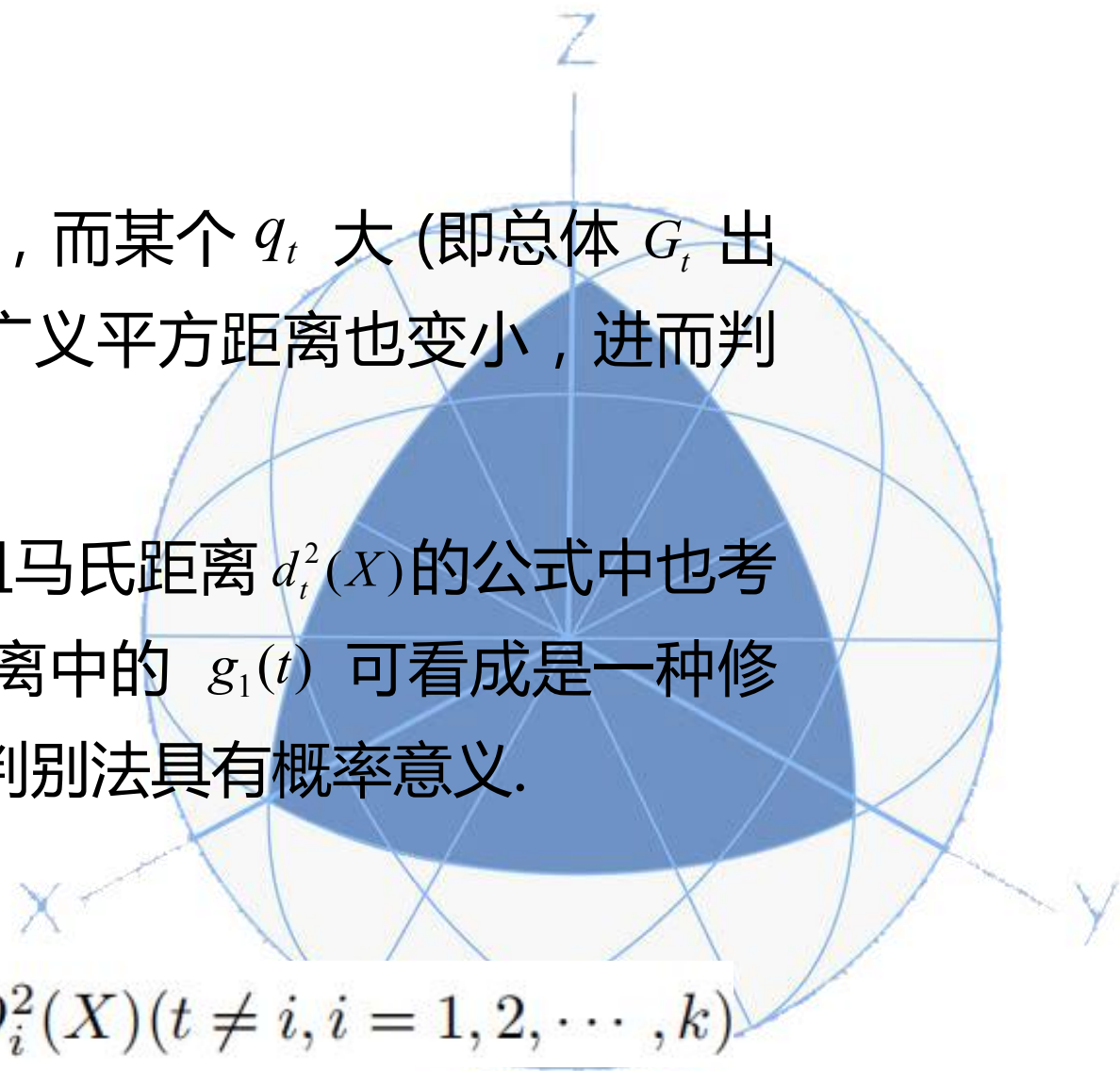


由  $D_t^2(X)$  的公式可见，当  $d_t^2(X)$  不变，而某个  $q_t$  大 (即总体  $G_t$  出现的机会大) 时，则  $g_2(t)$  变小，故广义平方距离也变小，进而判  $X$  为  $G_t$  的可能性大。

当  $\Sigma_i$  不全相等时， $g_1(t) = \log|S_t|$ ，且马氏距离  $d_t^2(X)$  的公式中也考虑了  $\Sigma_i$  的不等，这时广义平方距离中的  $g_1(t)$  可看成是一种修正。引入  $g_1(t)$  后，使广义平方距离判别法具有概率意义。

利用广义平方距离的判别法为：

判  $X \in G_t$ , 当  $D_t^2(X) < D_i^2(X) (t \neq i, i = 1, 2, \dots, k)$



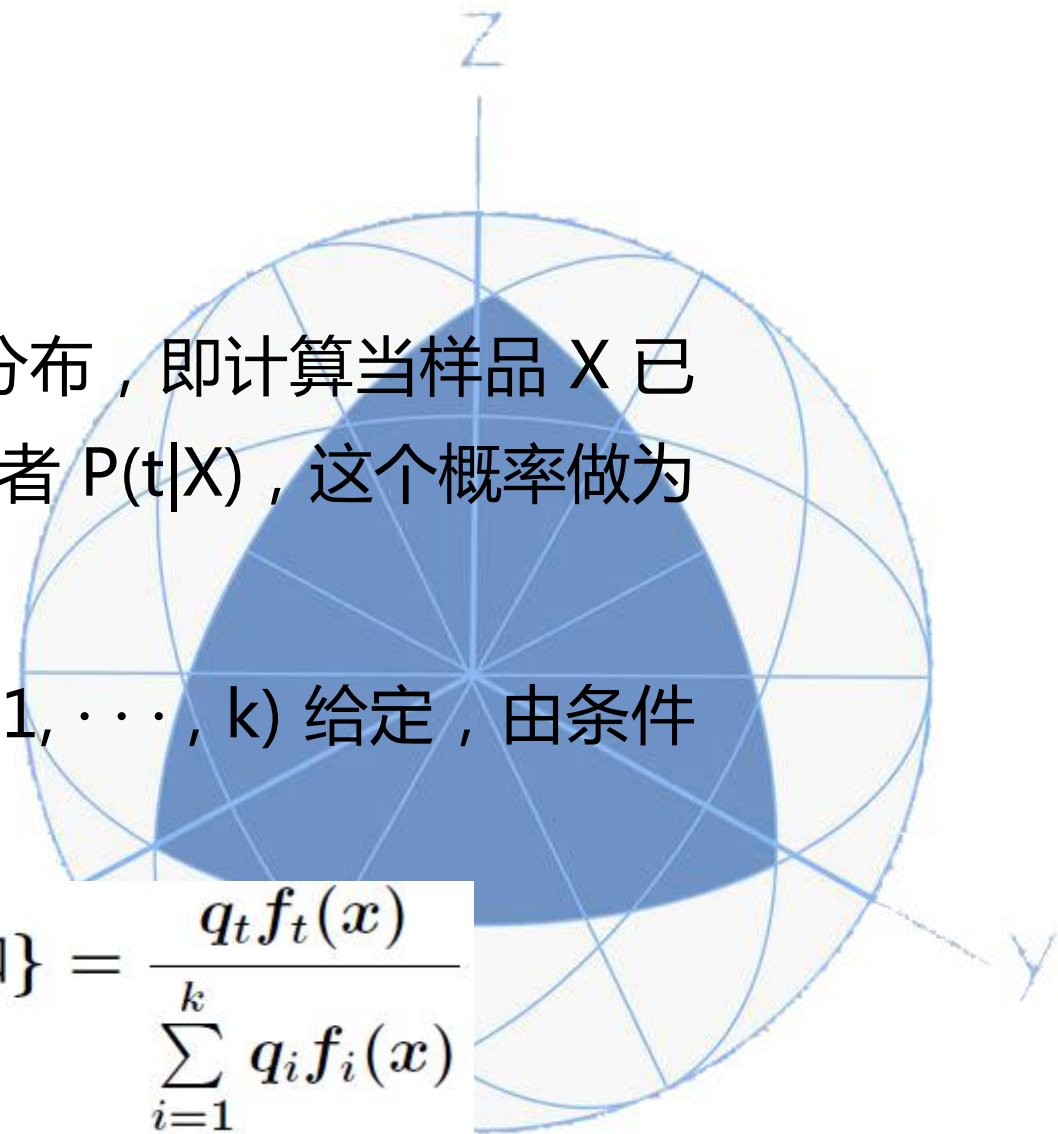


### 3、后验概率

标准的 Bayes 方法应该计算后验概率分布，即计算当样品  $X$  已知时，它属于  $G_t$  的概率记为  $P(G_t | t)$  或者  $P(t|X)$ ，这个概率做为判别归类的准则.

假定总体  $G_t$  的概率密度函数  $f_t(x)$  ( $t = 1, \dots, k$ ) 给定，由条件概率的定义可以导出：

$$P(t|X) = P\{X \in G_t | X \text{ 已知}\} = \frac{q_t f_t(x)}{\sum_{i=1}^k q_i f_i(x)}$$



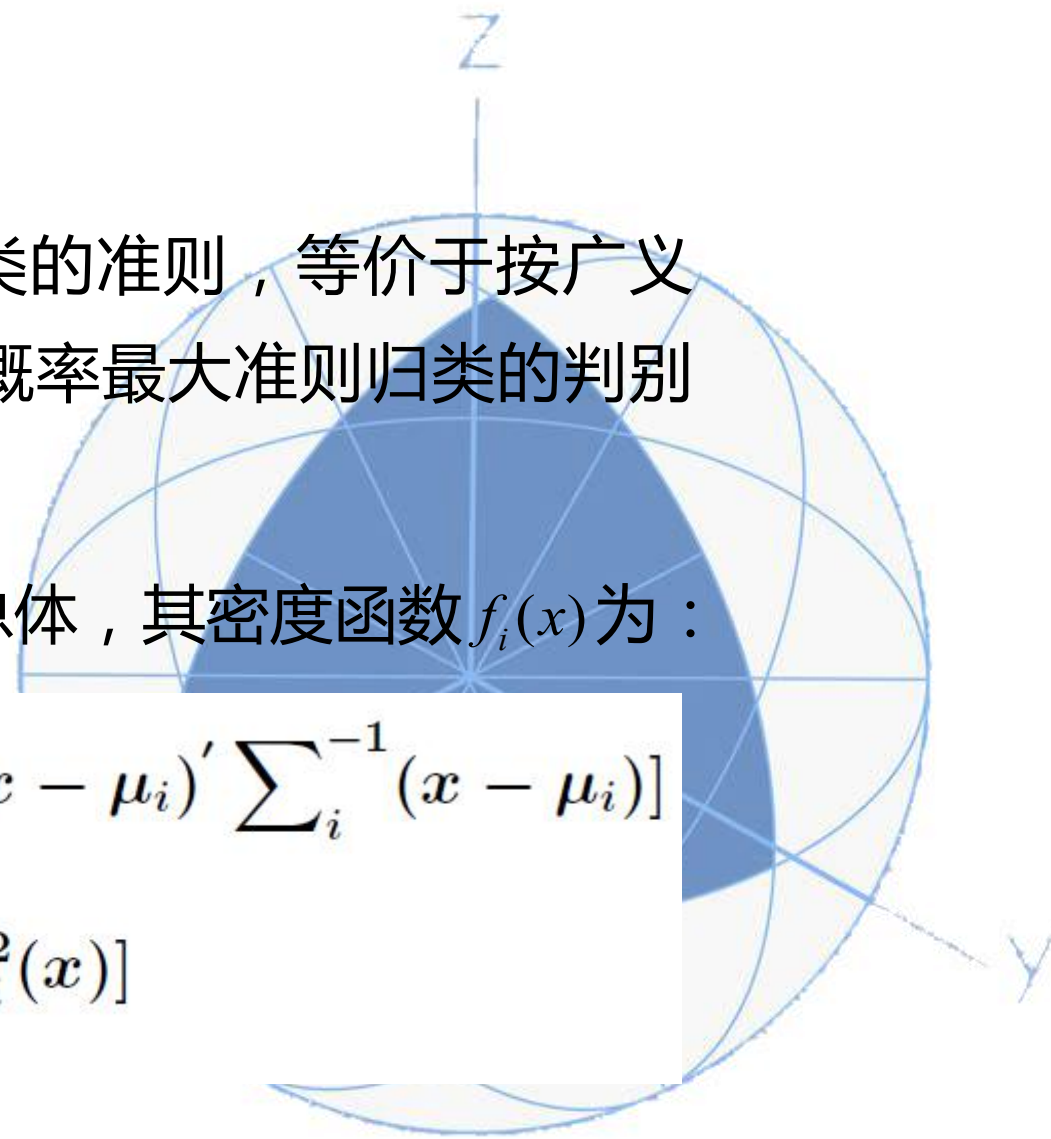




在正态假设下按后验概率最大进行归类的准则，等价于按广义平方距离最小准则进行归类.而按后验概率最大准则归类的判别法就是 Bayes 判别的一种情况.

若假设  $G_i$  ( $i = 1, \dots, k$ ) 为  $m$  维正态总体，其密度函数  $f_i(x)$  为：

$$\begin{aligned} f_i(x) &= \frac{1}{(2\pi)^{\frac{m}{2}} (|\sum_i|)^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(x - \mu_i)' \sum_i^{-1} (x - \mu_i)\right] \\ &= \frac{1}{(2\pi)^{\frac{m}{2}} (|\sum_i|)^{\frac{1}{2}}} \exp\left[-\frac{1}{2}d_i^2(x)\right] \end{aligned}$$



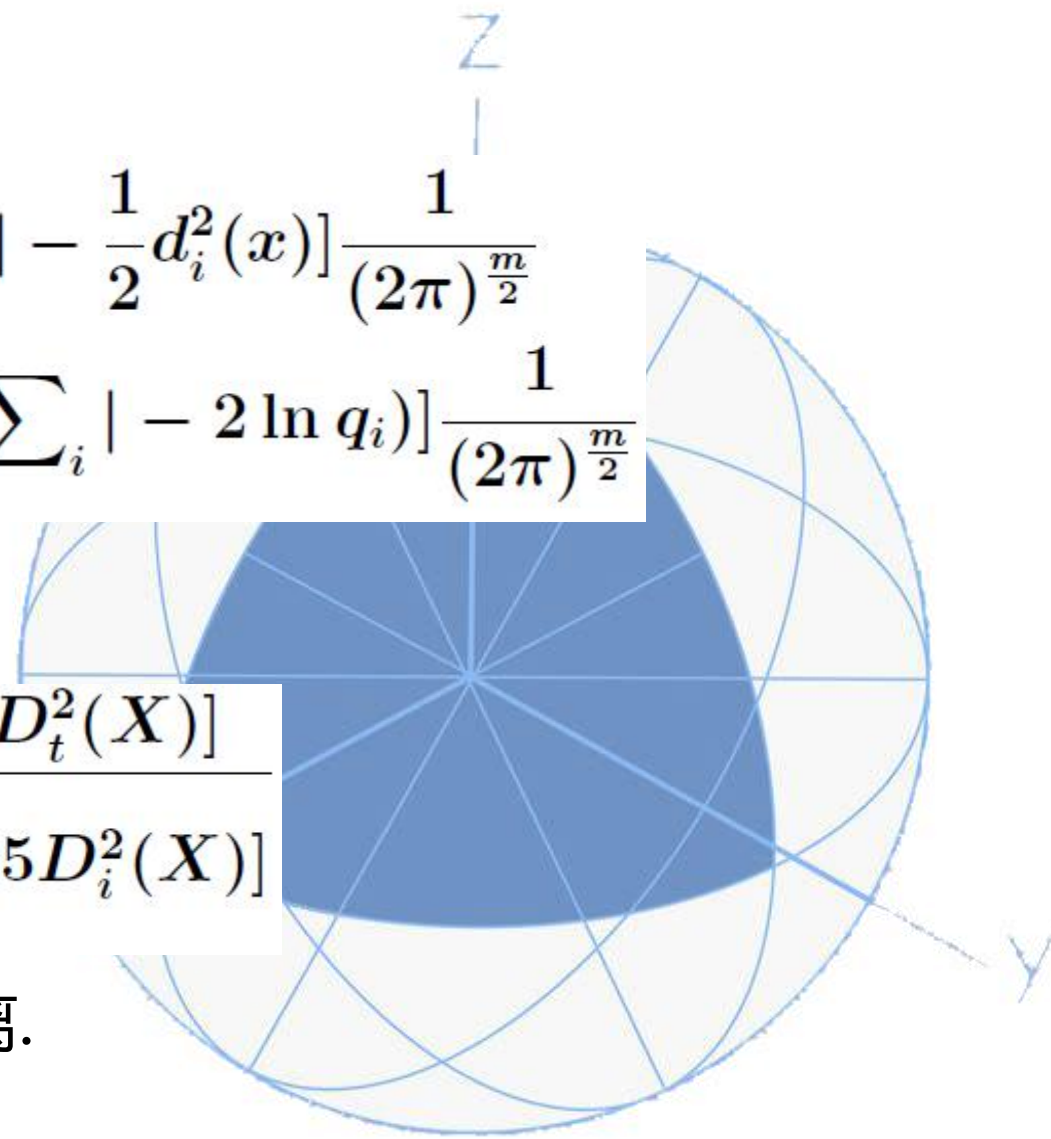


$$\begin{aligned} q_i f_i(x) &= \exp[\ln q_i - \frac{1}{2} \ln |\sum_i| - \frac{1}{2} d_i^2(x)] \frac{1}{(2\pi)^{\frac{m}{2}}} \\ &= \exp[-\frac{1}{2} (d_i^2(x) + \ln |\sum_i| - 2 \ln q_i)] \frac{1}{(2\pi)^{\frac{m}{2}}} \end{aligned}$$

则  $X$  属于第  $t$  组的后验概率为:

$$P(t|X) = \frac{\exp[-0.5 D_t^2(X)]}{\sum_{i=1}^k \exp[-0.5 D_i^2(X)]}$$

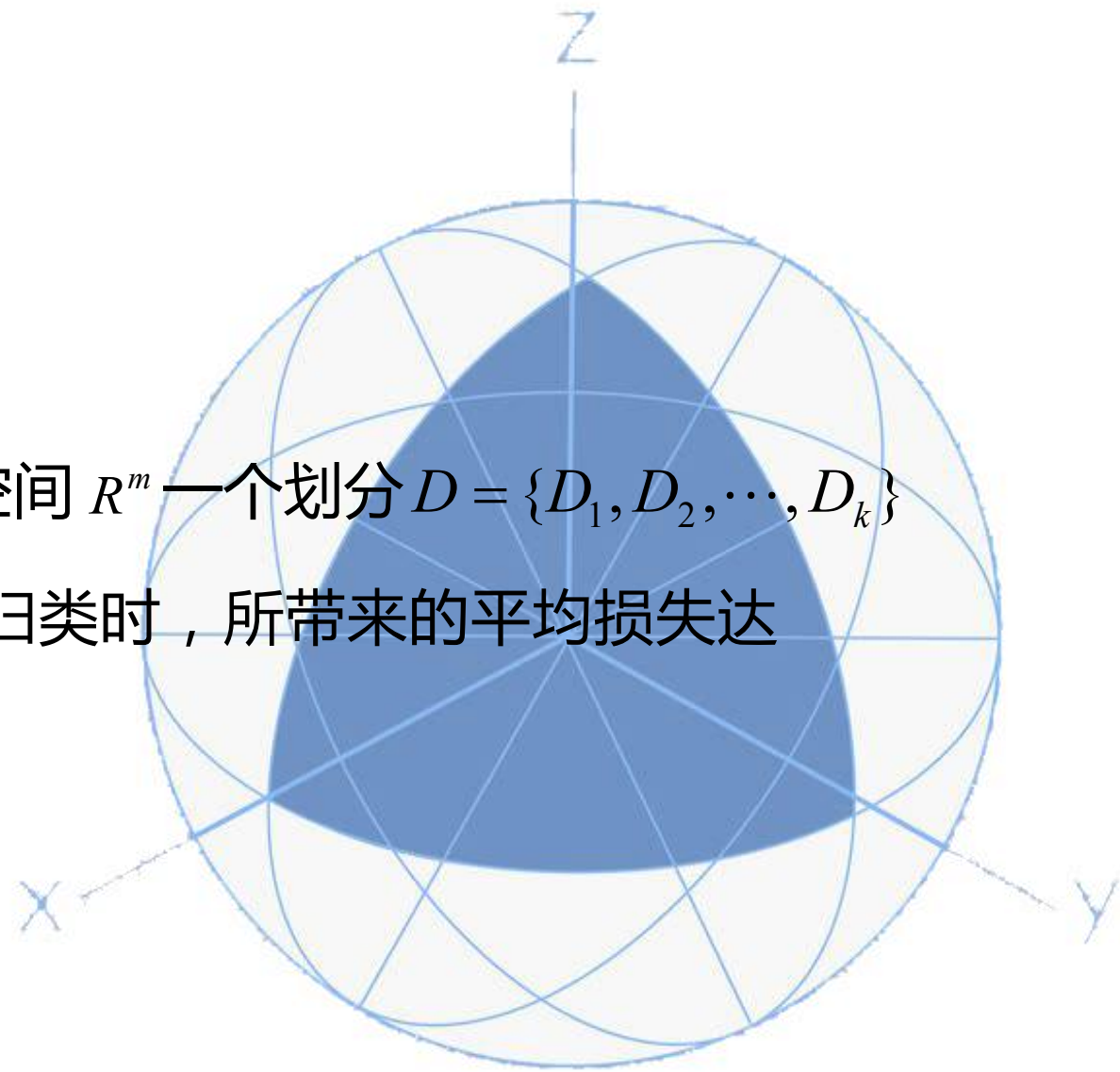
其中  $D_i^2(X)$  是  $X$  到第  $i$  组的广义平方距离.





## 4、Bayes 判别准则

所谓 Bayes 判别准则，就是给出空间  $R^m$  一个划分  $D = \{D_1, D_2, \dots, D_k\}$ ，使得当通过这个划分  $D$  来判别归类时，所带来的平均损失达到最小。

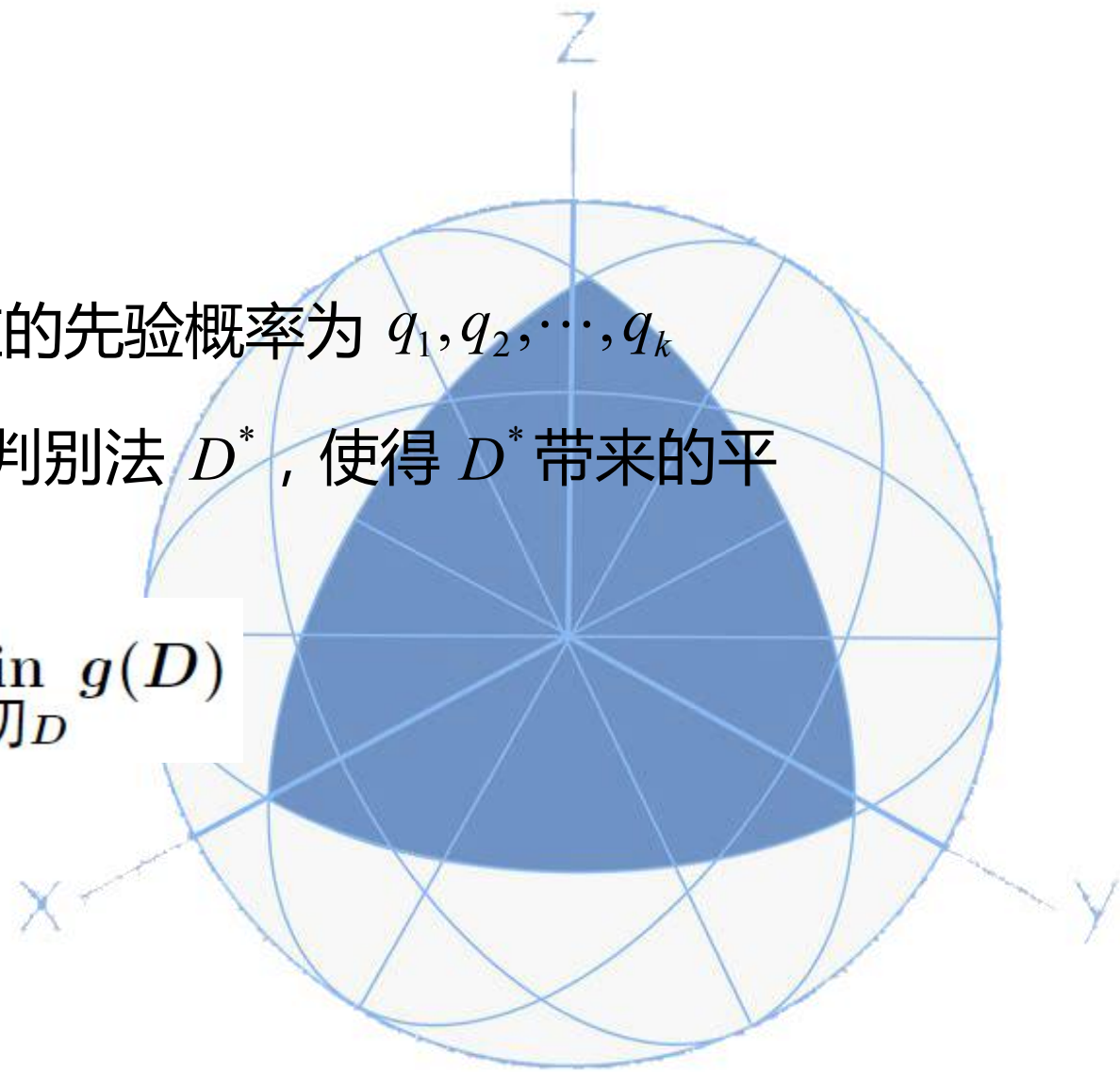




设有  $k$  个总体： $G_1, G_2, \dots, G_k$ ，相应的先验概率为  $q_1, q_2, \dots, q_k$   
( $q_i > 0, q_1 + q_2 + \dots + q_k = 1$ )。如果有判别法  $D^*$ ，使得  $D^*$  带来的平均损失  $g(D^*)$  达最小，即

$$g(D^*) = \min_{\text{一切 } D} g(D)$$

则称判别法  $D^*$  符合 Bayes 准则。





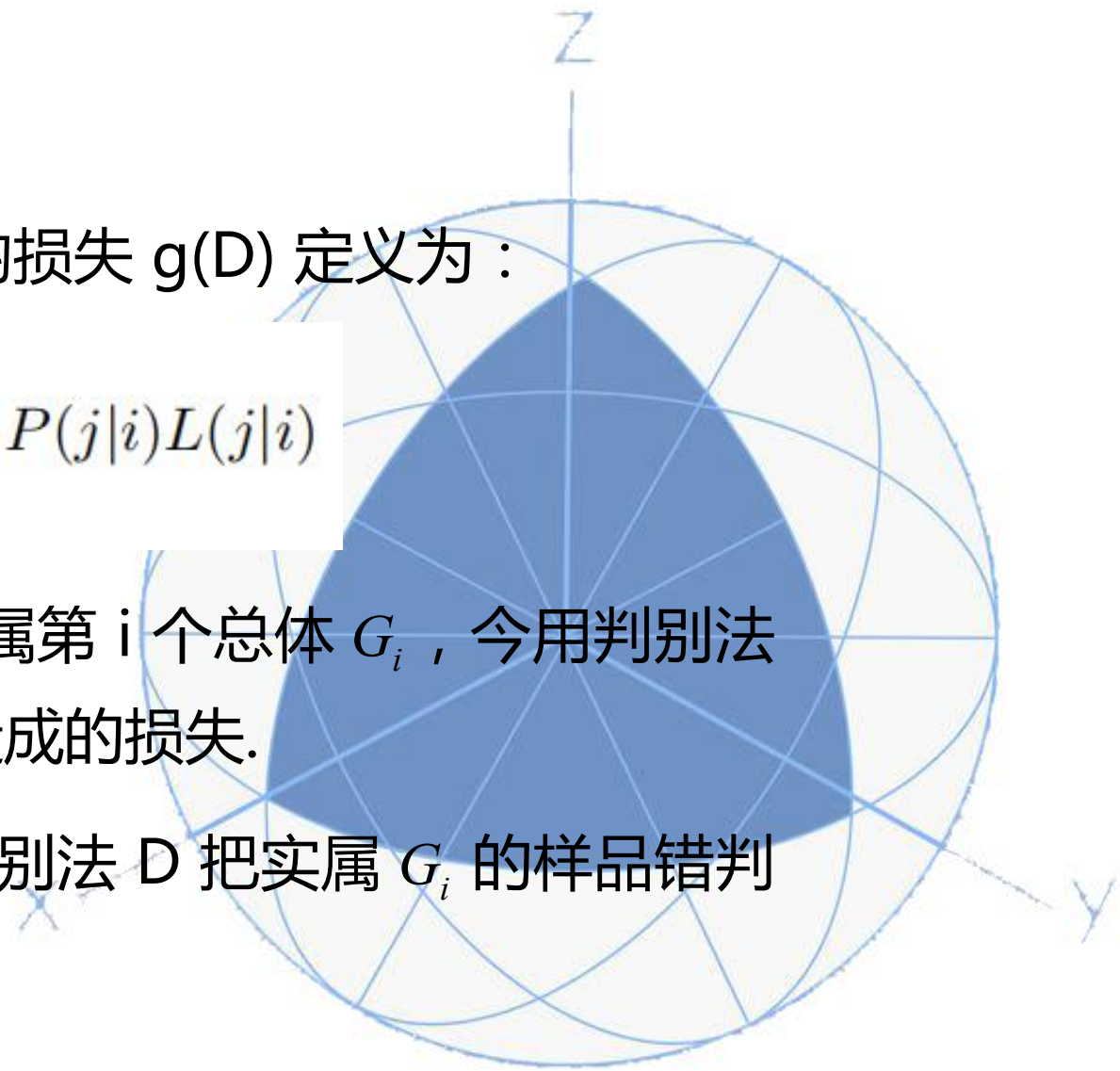


判别法  $D$  关于先验概率的错判平均损失  $g(D)$  定义为：

$$g(D) = \sum_{i=1}^k q_i \sum_{j=1}^k P(j|i) L(j|i)$$

$L(j|i; D)$  (简记为  $L(j|i)$ ) 表示样品实属第  $i$  个总体  $G_i$ ，今用判别法  $D$  判别时被错判为  $G_j (j \neq i)$  时所造成的损失。

$P(j|i; D)$  (或简记为  $P(j|i)$ ) 表示用判别法  $D$  把实属  $G_i$  的样品错判为  $G_j$  的概率。

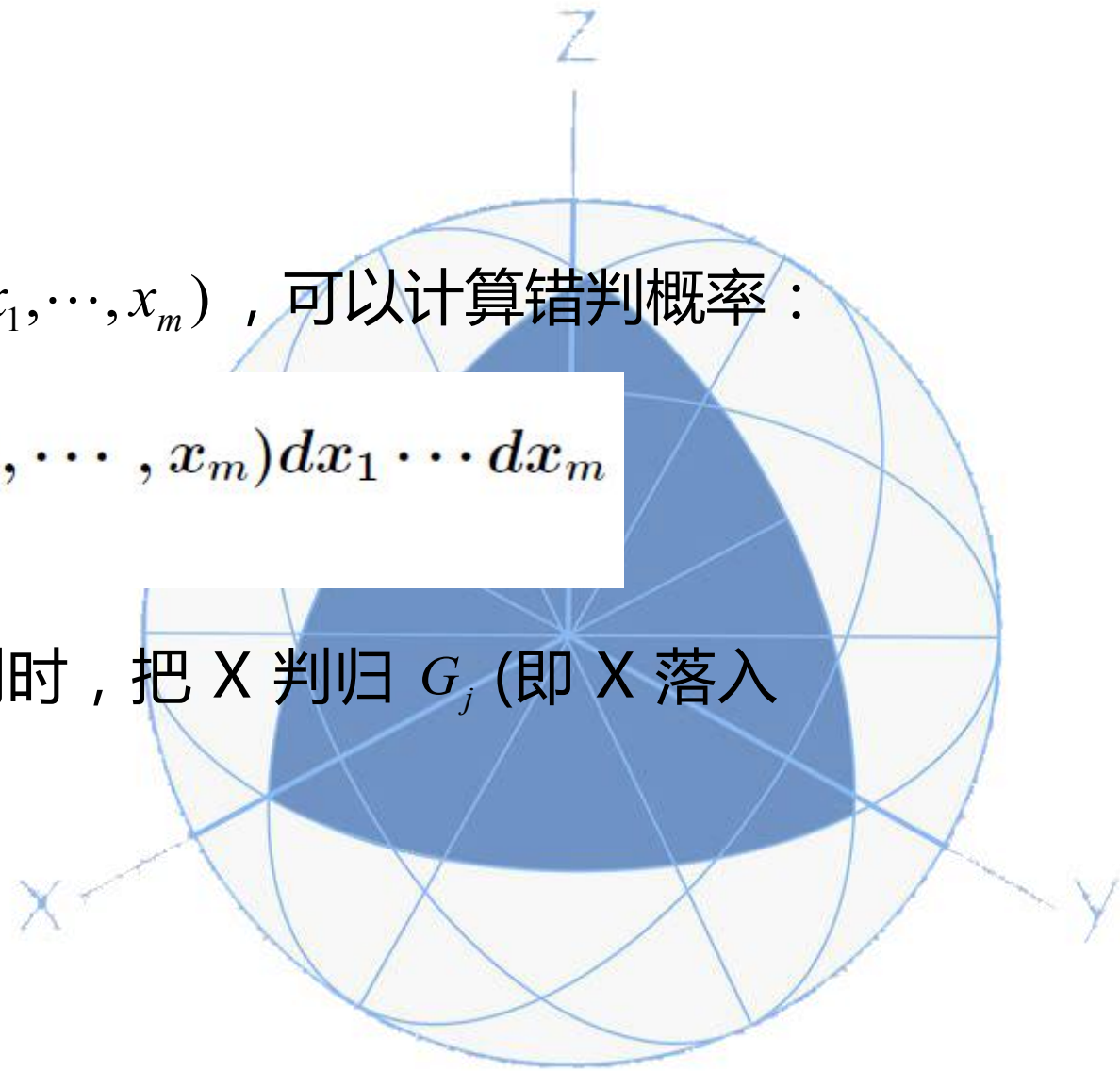




当总体  $G_i$  的分布密度已知记为  $f_i(x_1, \dots, x_m)$  , 可以计算错判概率 :

$$P(j|i; D) = \int \cdots \int_{D_j} f_i(x_1, \dots, x_m) dx_1 \cdots dx_m$$

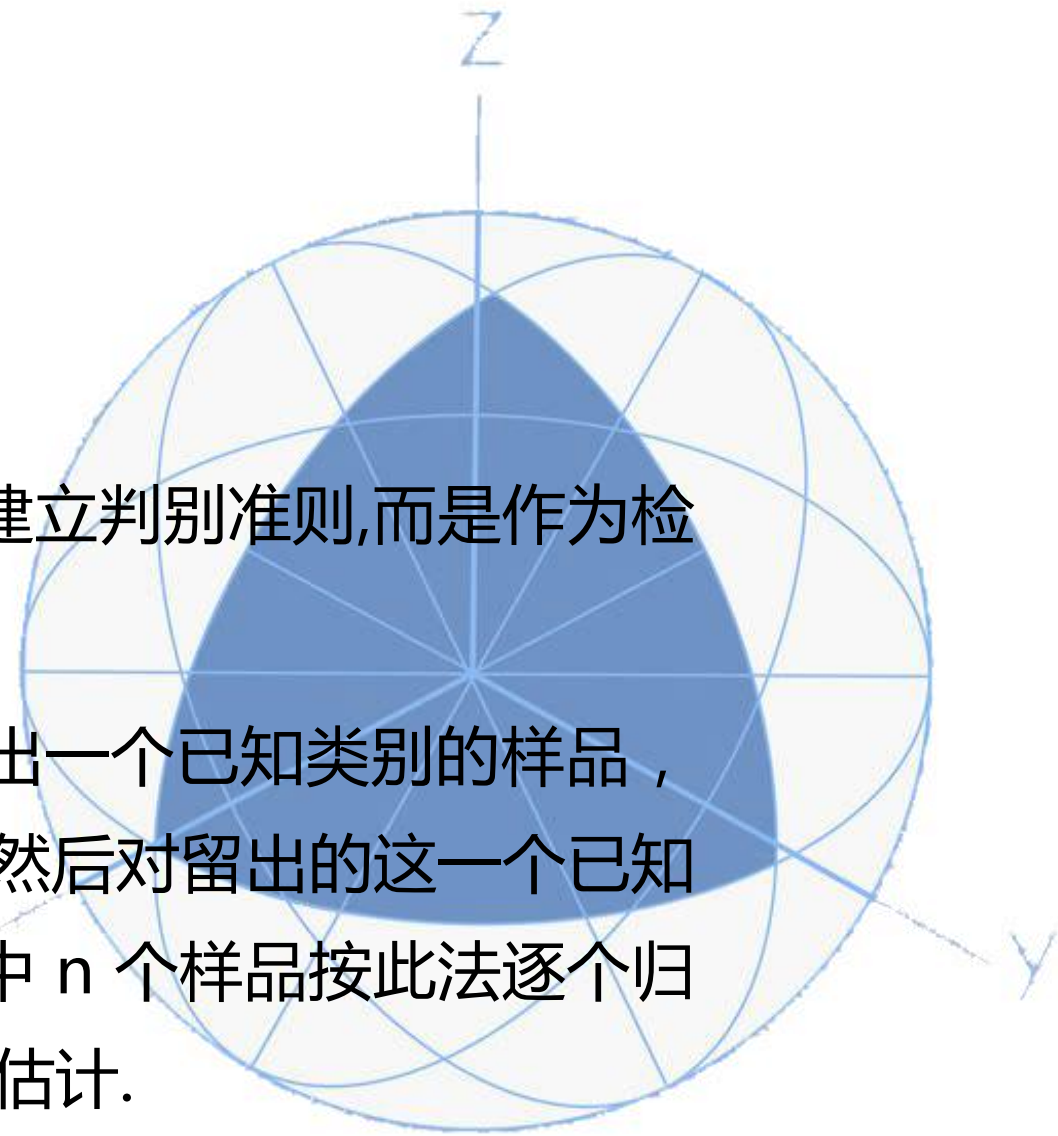
当样品  $X \in G_i$  , 但用判别法  $D$  判别时 , 把  $X$  判归  $G_j$  (即  $X$  落入区域  $D_j, j \neq i$  ) , 这时判错了.





错判概率  $P(j|i)$  估计方法有以下几种：

- (1) 利用训练样本作为检验集；
- (2) 可留出一些已知类别的样品不参加建立判别准则,而是作为检验集；
- (3) 舍一法 (或称交叉确认法)，每次留出一个已知类别的样品，而用其余  $n - 1$  个样品建立判别准则，然后对留出的这一个已知类别的样品进行判别归类. 对训练样本中  $n$  个样品按此法逐个归类后，最后把错判的比率作为错判率的估计.

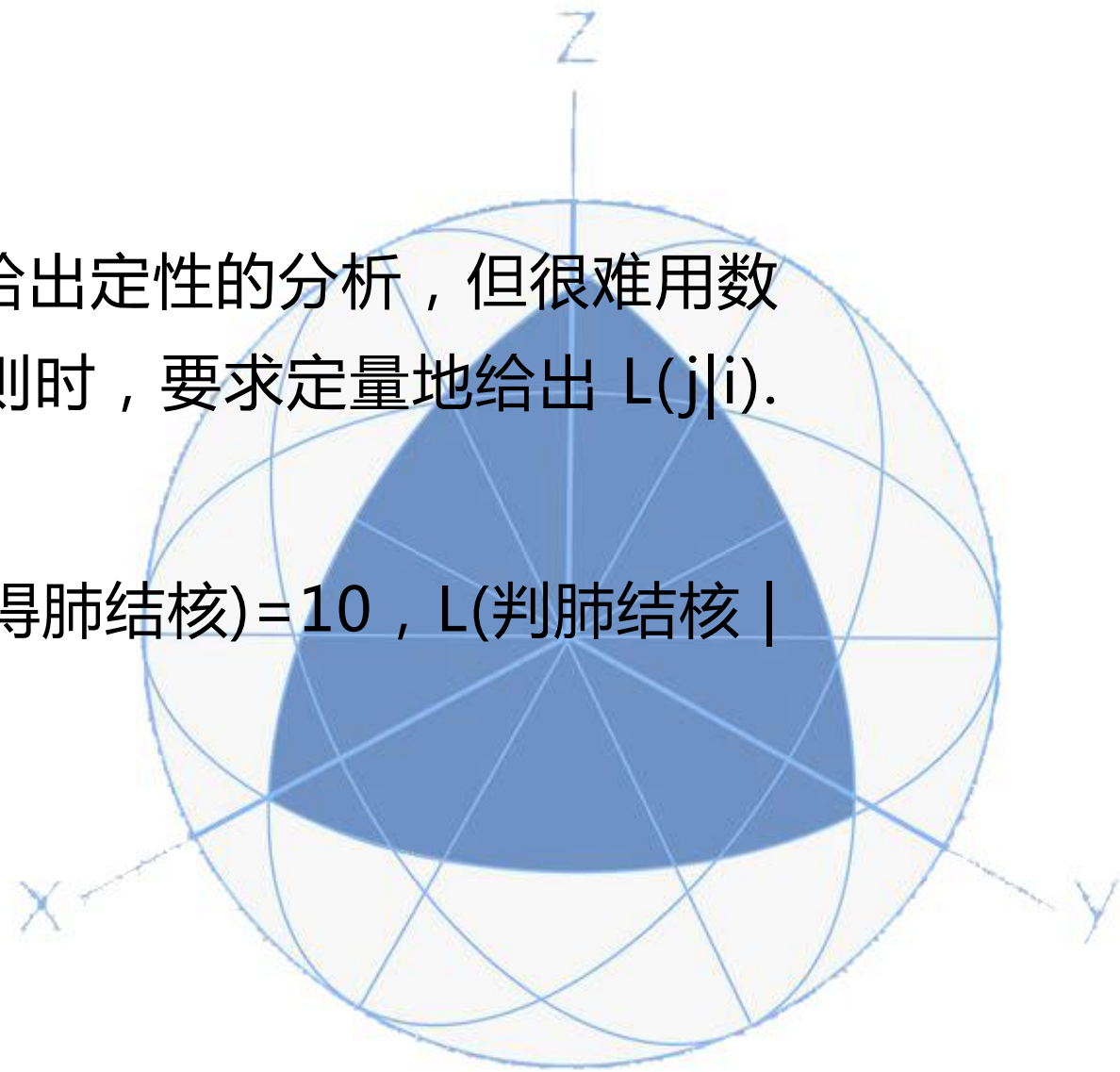




在实际问题中，错判的损失可以给出定性的分析，但很难用数值来表示. 但应用 Bayes 判别准则时，要求定量地给出  $L(j|i)$ .  $L(j|i)$  的赋值法常用的有以下两种：

a 由经验人为赋值. 例如  $L(\text{判癌} | \text{得肺结核}) = 10$ ， $L(\text{判肺结核} | \text{得癌症}) = 1000$

b 假定各种错判损失都相等.



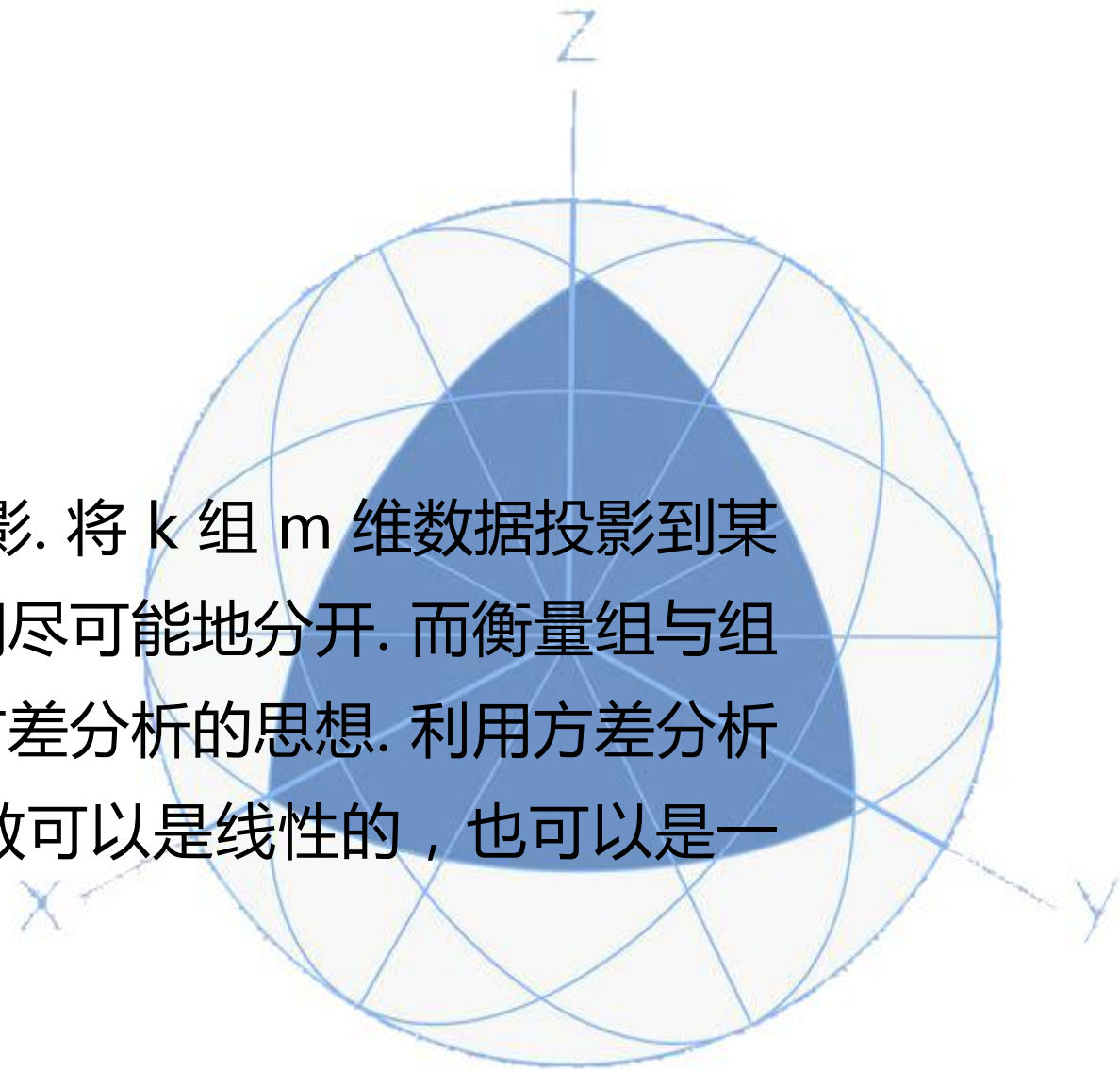


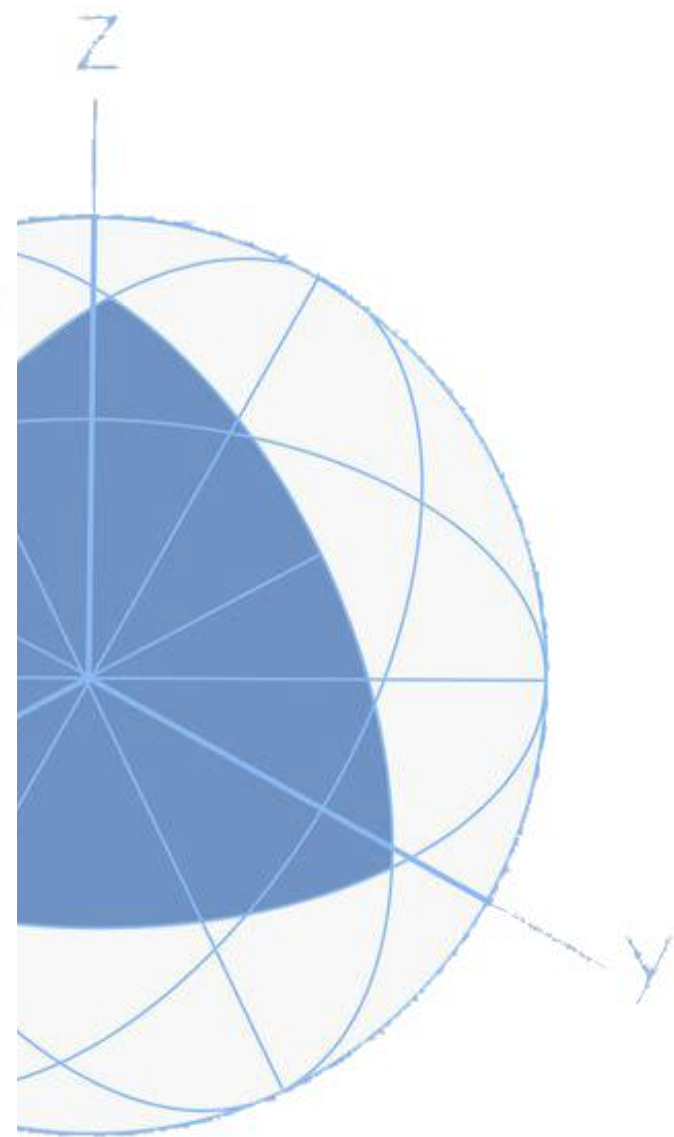
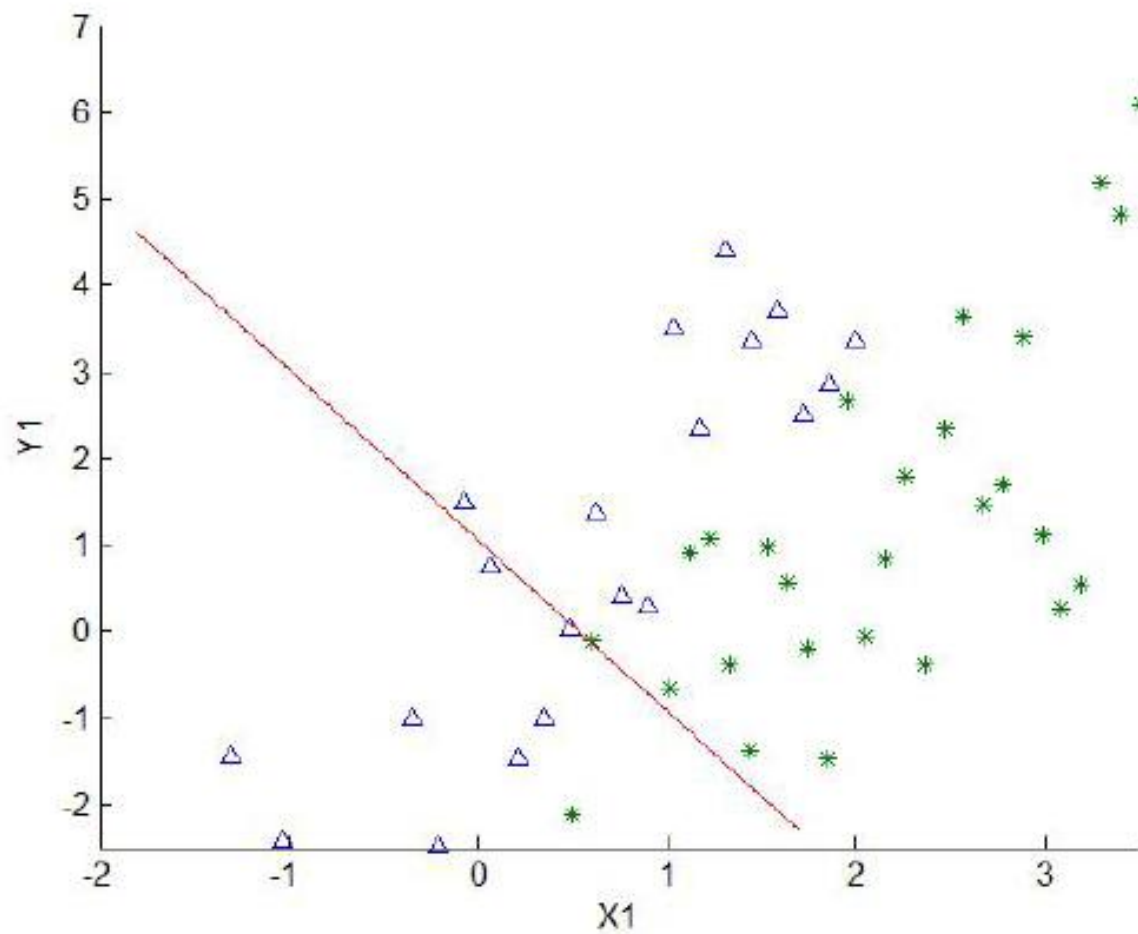


## 四、Fisher 判别法

### 1、Fisher 判别的基本思想

Fisher(费歇)判别的基本思想是投影. 将  $k$  组  $m$  维数据投影到某一个方向, 使得投影后组与组之间尽可能地分开. 而衡量组与组之间是否分开的方法借助于一元方差分析的思想. 利用方差分析的思想来导出判别函数, 这个函数可以是线性的, 也可以是一般的函数.





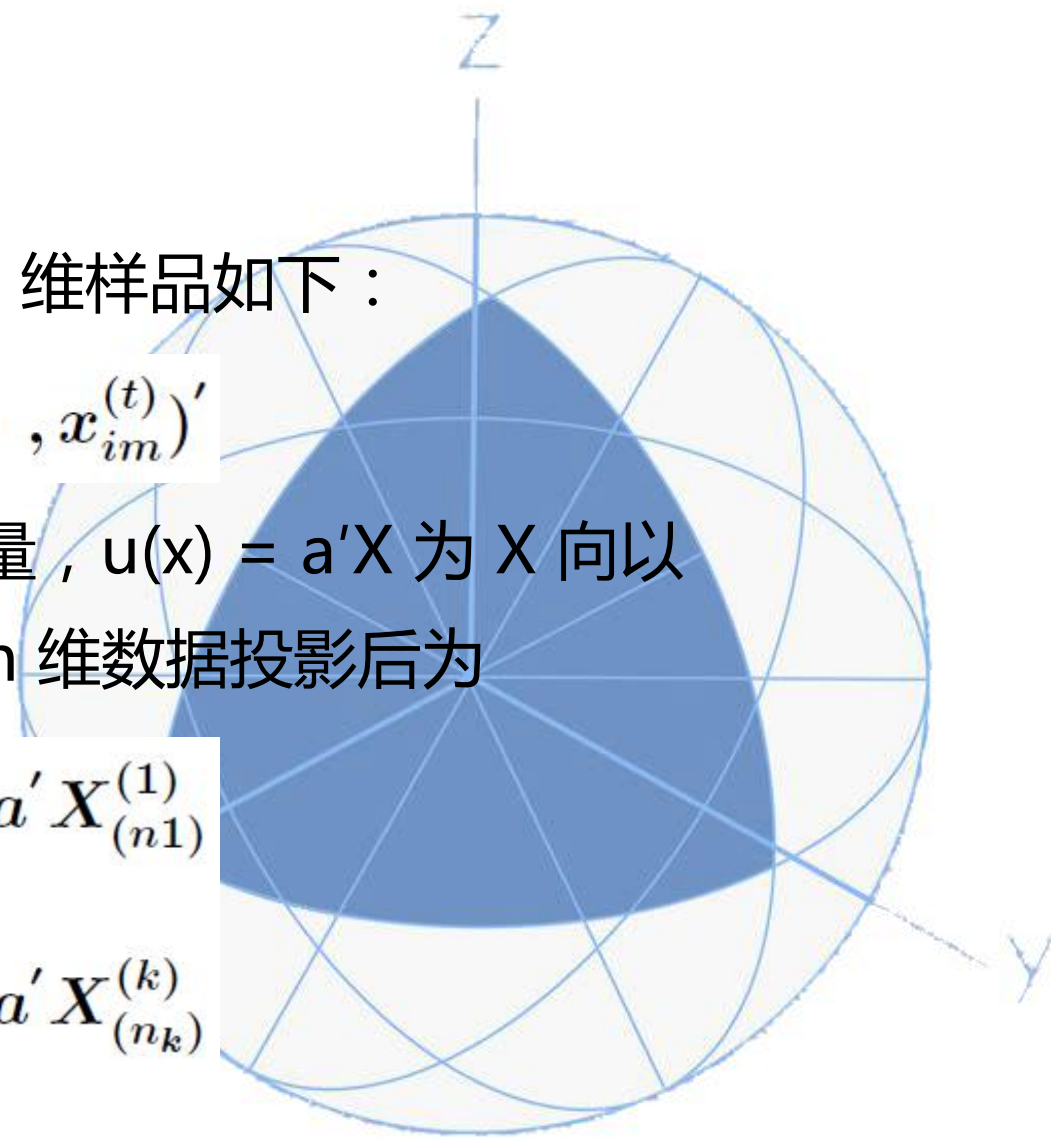


设从总体  $G_t$  ( $t = 1, \cdots, k$ ) 分别抽取  $m$  维样品如下：

$$X_{(i)}^{(t)} = (x_{i1}^{(t)}, x_{i2}^{(t)}, \cdots, x_{im}^{(t)})'$$

令  $a = (a_1, \cdots, a_m)'$  为  $m$  维空间的任一向量， $u(x) = a'X$  为  $X$  向以  $a$  为法线的方向投影。上述  $k$  个组中的  $m$  维数据投影后为

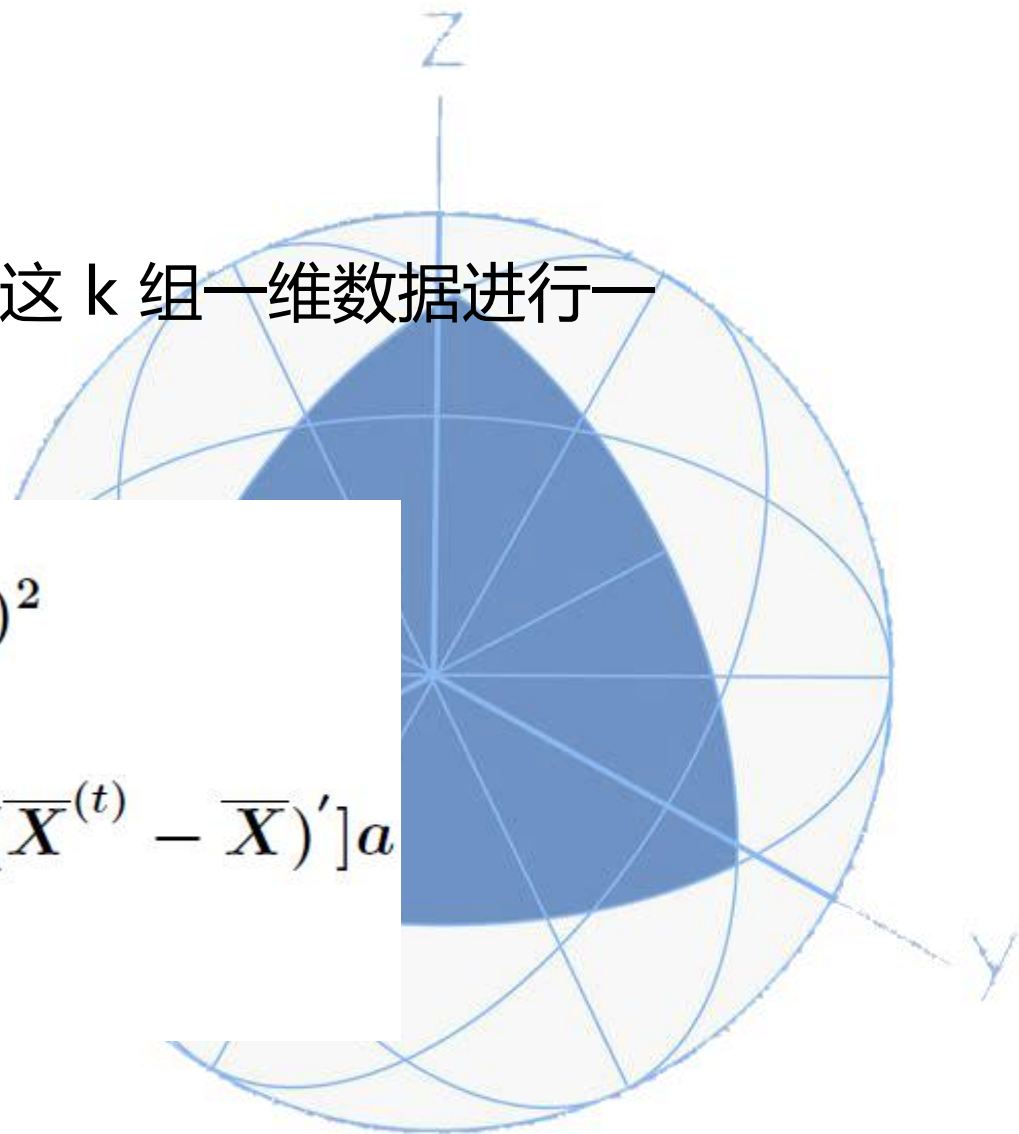
$$\begin{aligned} G_1 : & a' X_{(1)}^{(1)}, \cdots, a' X_{(n_1)}^{(1)} \\ & \cdots \qquad \qquad \cdots \\ G_k : & a' X_{(1)}^{(k)}, \cdots, a' X_{(n_k)}^{(k)} \end{aligned}$$





每个总体的数据投影后均为一维数据. 对这  $k$  组一维数据进行一元方差分析, 其组间平方和为

$$\begin{aligned} B_0 &= \sum_{t=1}^k n_t (a' \bar{X}^{(t)} - a' \bar{X})^2 \\ &= a' \left[ \sum_{t=1}^k n_t (\bar{X}^{(t)} - \bar{X})(\bar{X}^{(t)} - \bar{X})' \right] a \\ &= a' B a \end{aligned}$$





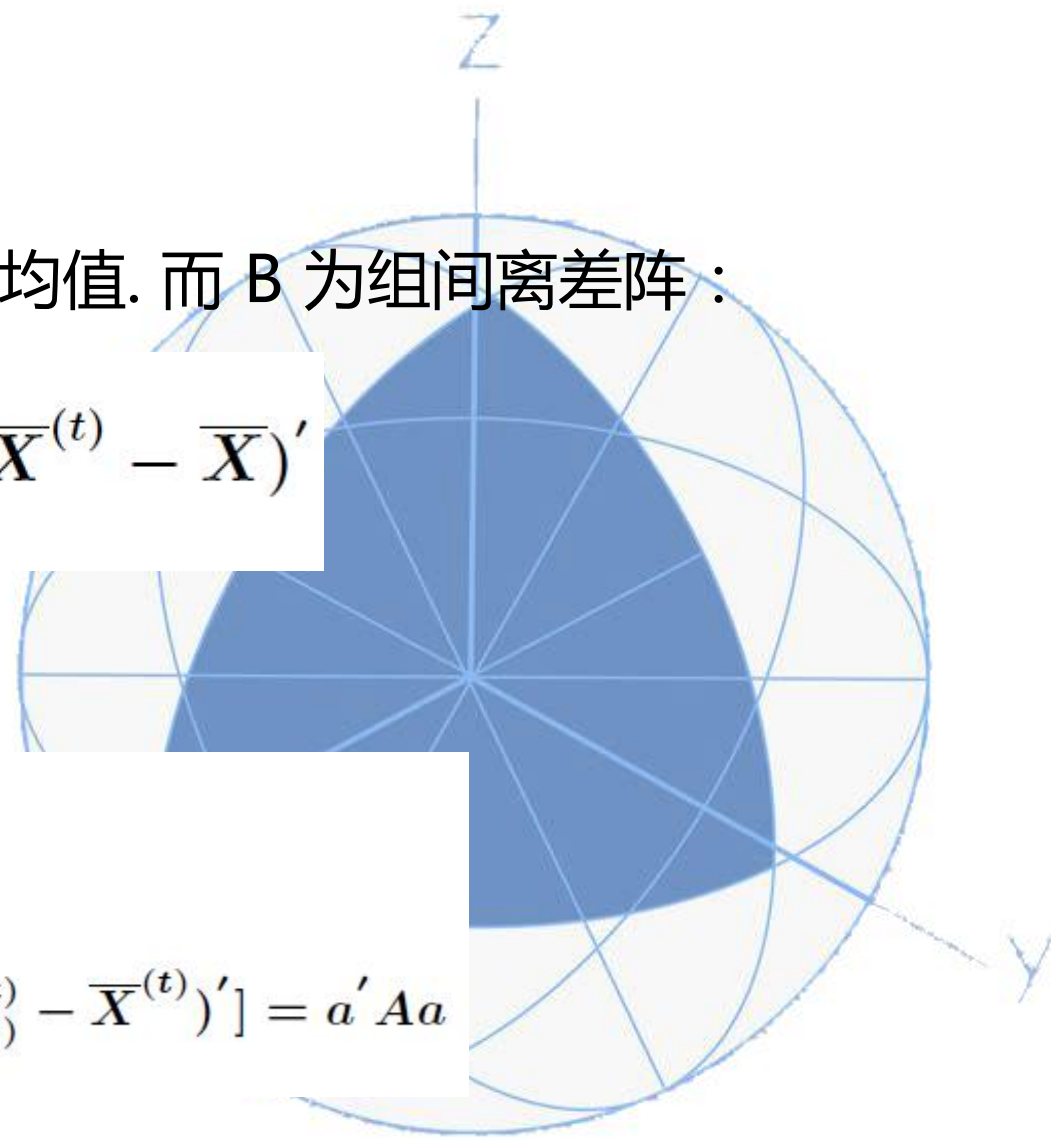


其中  $\bar{X}^{(t)}$  和  $\bar{X}$  分别为  $G_t$  的样本均值和总均值. 而  $B$  为组间离差阵:

$$B = \sum_{t=1}^k n_t (\bar{X}^{(t)} - \bar{X})(\bar{X}^{(t)} - \bar{X})'$$

合并的组内平方和为

$$\begin{aligned} A_0 &= \sum_{t=1}^k \sum_{j=1}^{n_t} (a' X_{(j)}^{(t)} - a' \bar{X}^{(t)})^2 \\ &= a' \left[ \sum_{t=1}^k \sum_{j=1}^{n_t} (X_{(j)}^{(t)} - \bar{X}^{(t)})(X_{(j)}^{(t)} - \bar{X}^{(t)})' \right] = a' A a \end{aligned}$$





其中合并的组内离差阵  $A$  为

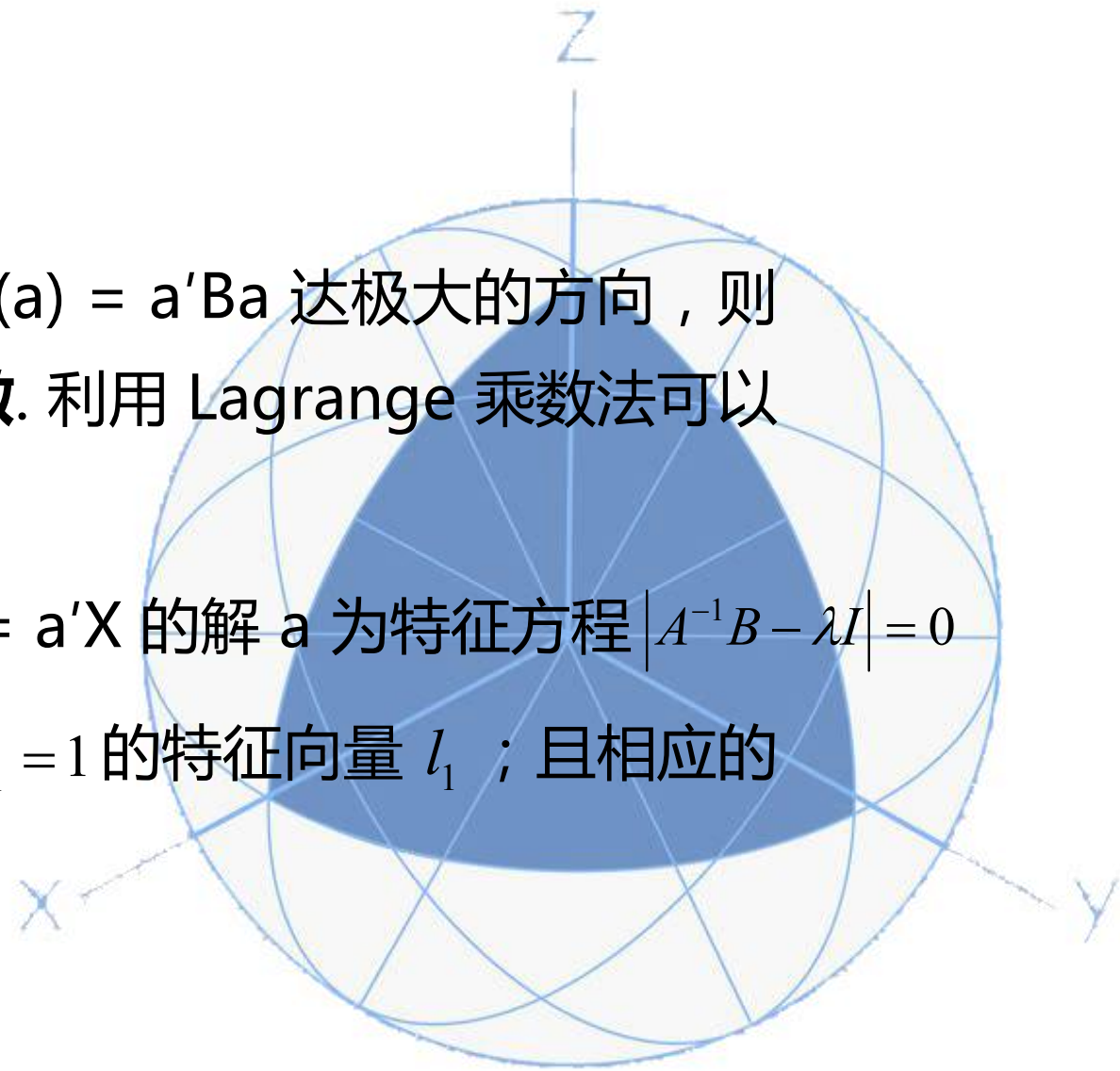
$$A = \sum_{t=1}^k \sum_{j=1}^{n_t} (X_{(j)}^{(t)} - \bar{X}^{(t)})(X_{(j)}^{(t)} - \bar{X}^{(t)})'$$

若  $k$  个类的均值有显著差异，则比值  $\frac{a'Ba}{a'Aa} = \Delta(a)$  应充分大. 利用方差分析的思想，问题化为求投影方向  $a$ ，使  $\Delta(a)$  达极大值，显然使  $\Delta(a)$  达极大的解  $a$  不唯一. 若  $a$  使  $\Delta(a)$  达极大，则  $Ca$  ( $C$  是任意不为零常数) 也使  $\Delta(\cdot)$  达极大，故对  $a$  加一约束条件，即选取  $a$  使  $a'Aa = 1$ . 问题化为求  $a$ ，使  $\Delta(a) = a'Ba$  在  $a'Aa = 1$  条件下达极大.



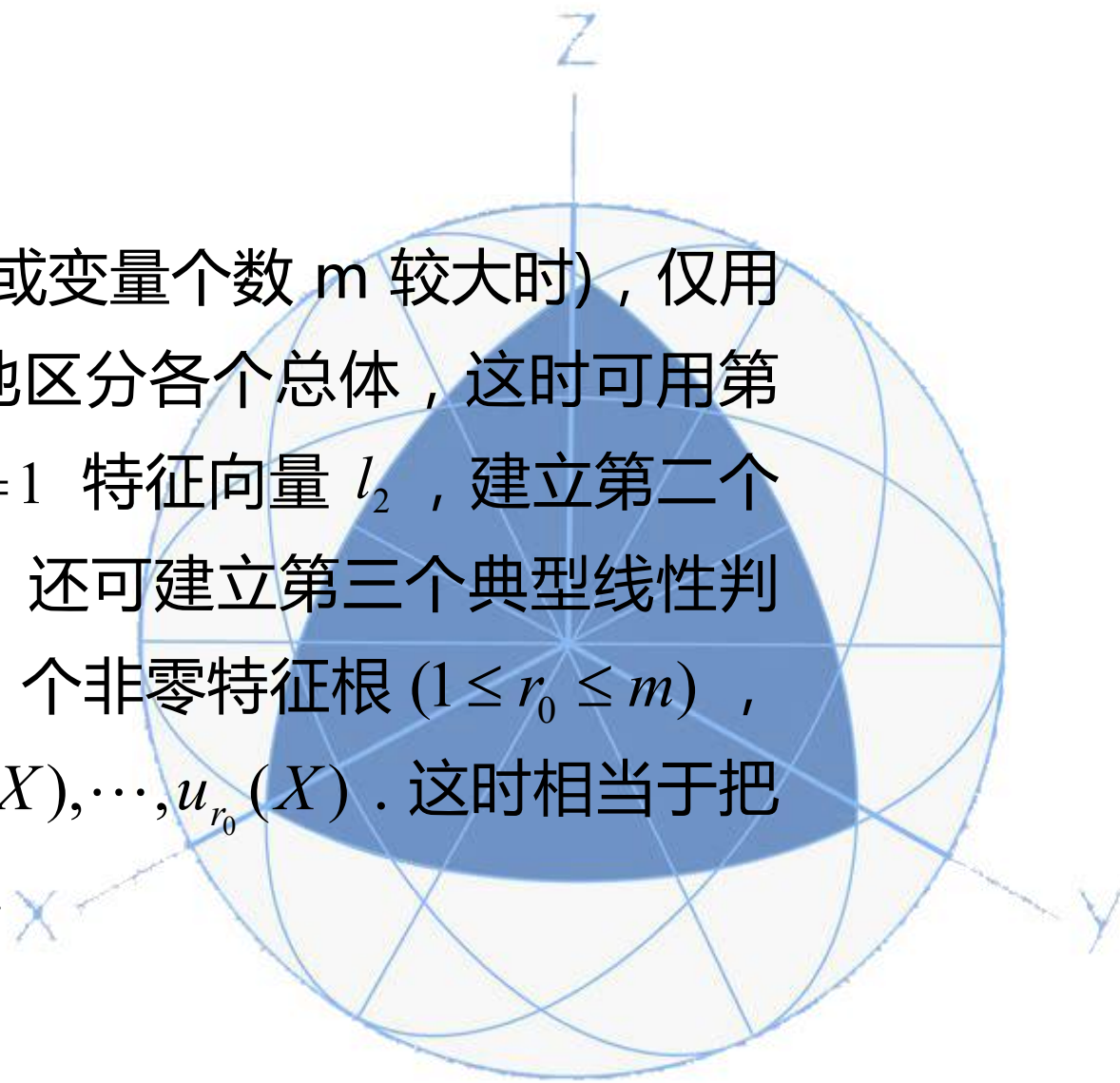
如果  $a$  是在  $a'Aa = 1$  条件下使  $\Delta(a) = a'Ba$  达极大的方向, 则称  $u(X) = a'X$  为**典型线性判别函数**. 利用 Lagrange 乘数法可以求条件极值问题的解.

Fisher 准则下线性判别函数  $u(X) = a'X$  的解  $a$  为特征方程  $|A^{-1}B - \lambda I| = 0$  的最大特征根  $\lambda_1$  所对应的满足  $l_1'Al_1 = 1$  的特征向量  $l_1$ ; 且相应的判别效率  $\Delta(l_1) = \lambda_1$ .





在有些问题中 (如分类个数  $k$  较大或变量个数  $m$  较大时), 仅用一个典型线性判别函数不能很好地区分各个总体, 这时可用第二大特征值  $\lambda_2$  对应的满足  $l_2' A l_2 = 1$  特征向量  $l_2$ , 建立第二个典型线性判别函数  $l_2' X$ ; 如还不够, 还可建立第三个典型线性判别函数  $l_3' X$ ; 依次类推. 如果有  $r_0$  个非零特征根 ( $1 \leq r_0 \leq m$ ), 相应地有  $r_0$  个典型线性判别函数  $u_1(X), \dots, u_{r_0}(X)$ . 这时相当于把原来  $m$  个变量综合成  $r_0$  个新变量.



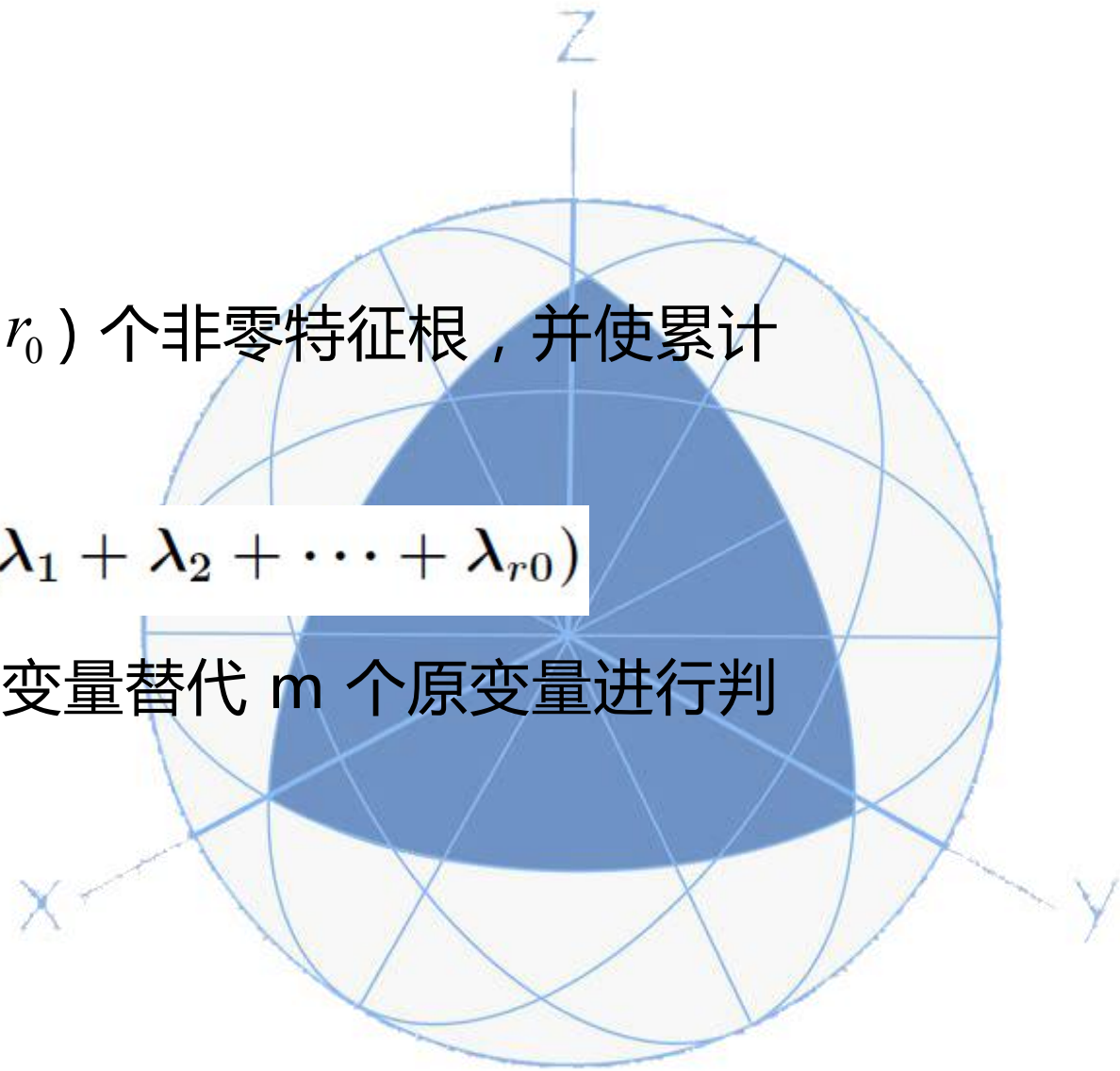




在实际应用中，常取前  $r(1 \leq r \leq r_0)$  个非零特征根，并使累计判别能力 (记为  $p_{(r)}$ )

$$p_{(r)} = (\lambda_1 + \cdots + \lambda_r) / (\lambda_1 + \lambda_2 + \cdots + \lambda_{r_0})$$

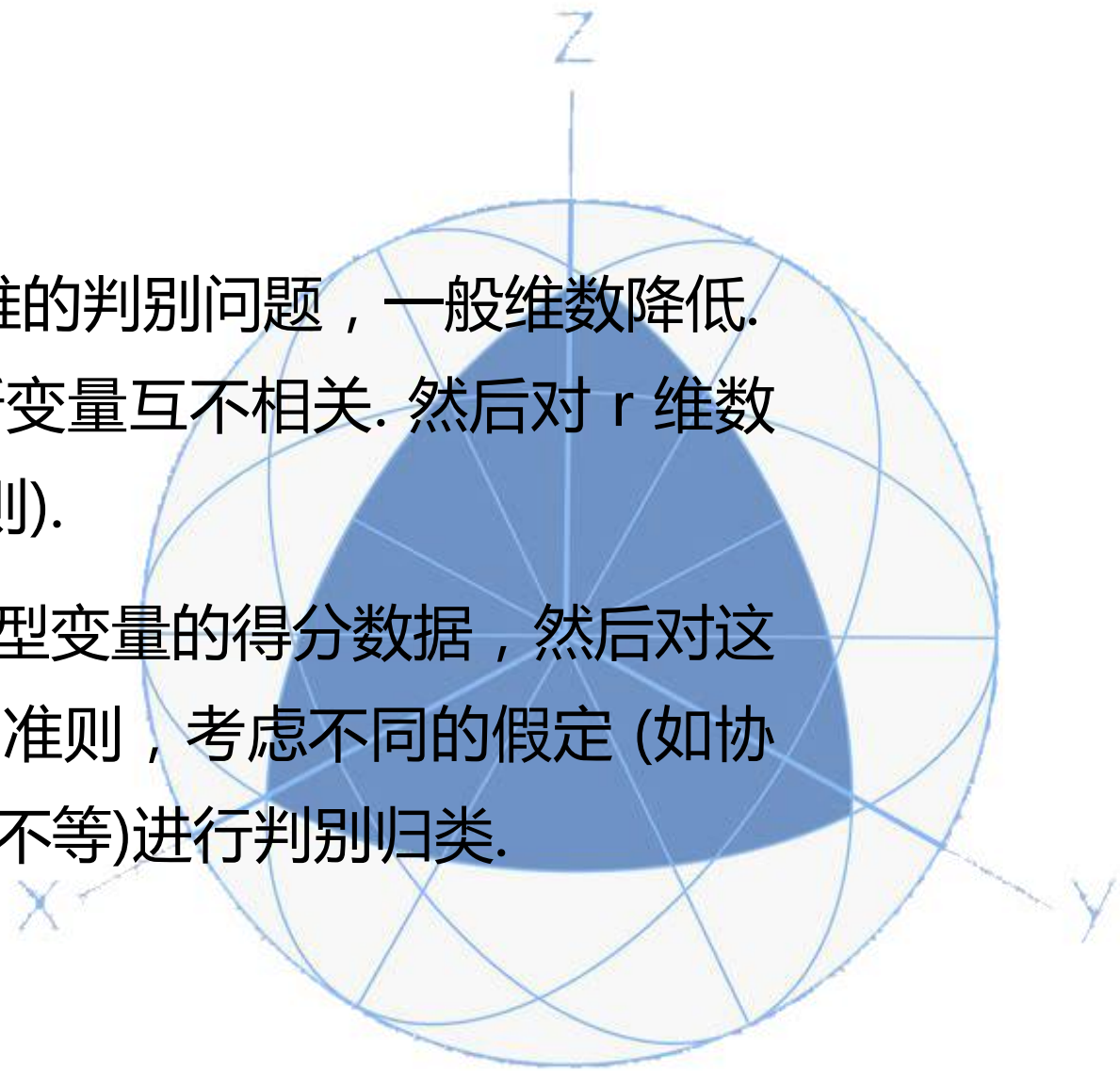
达到 0.8 以上 (这表示用这  $r$  个新变量替代  $m$  个原变量进行判别归类损失的信息不会超过 0.2).





这样  $m$  维总体的判别问题化为  $r$  维的判别问题，一般维数降低. 由于特征向量线性无关，故  $r$  个新变量互不相关. 然后对  $r$  维数据进行判别归类 (比如距离判别准则).

具体判别归类时，首先计算  $r$  个典型变量的得分数据，然后对这组  $r$  维的新数据，可以使用不同的准则，考虑不同的假定 (如协差阵相等或不等；先验概率相等或不等) 进行判别归类.



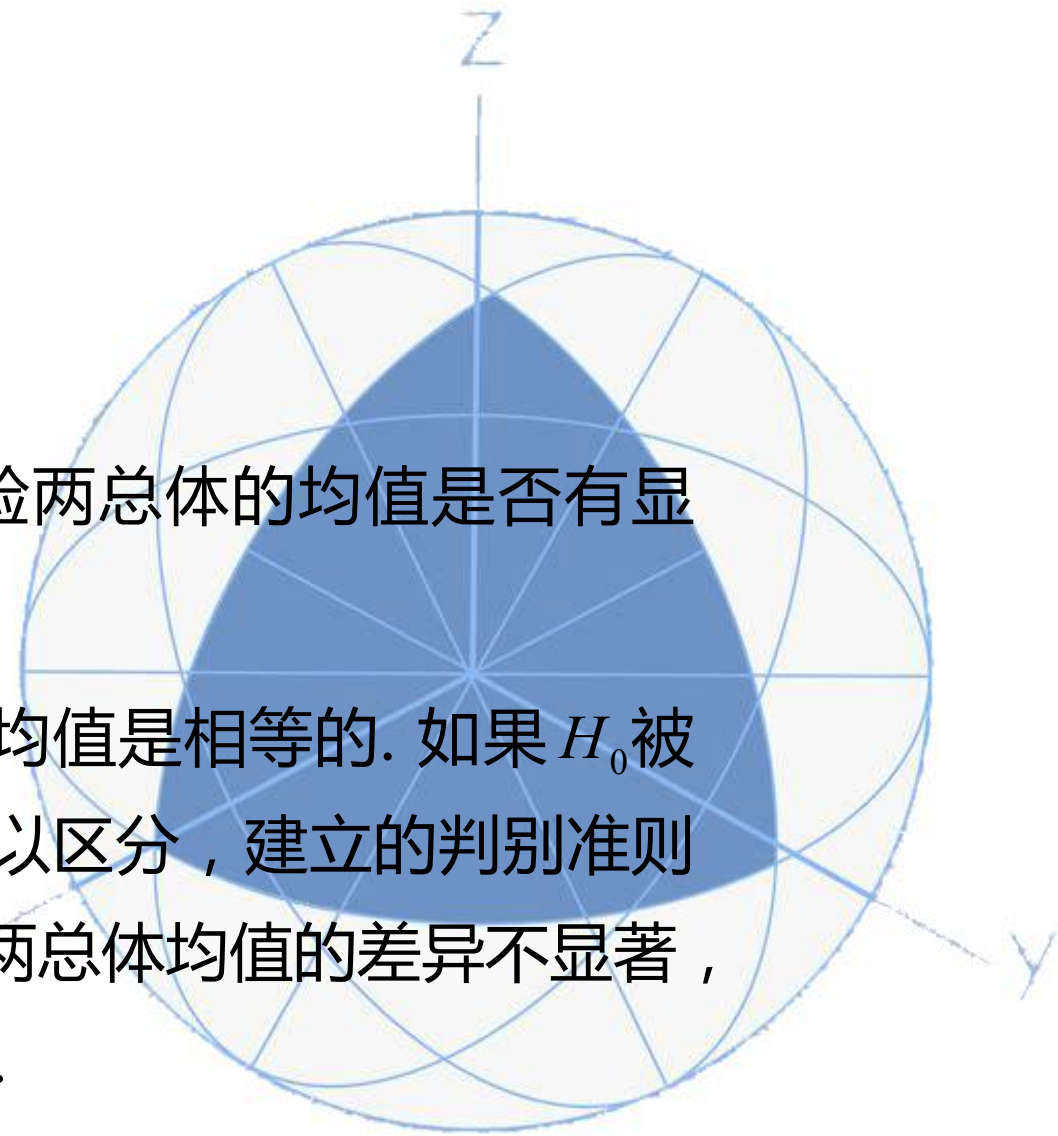


## 五、判别效果的检验

### 1、两总体判别效果的检验

所谓两总体判别效果的检验，就是检验两总体的均值是否有显著地差异.

一般我们提出的原假设  $H_0$  为两总体的均值是相等的. 如果  $H_0$  被否定，则说明两总体  $G_1$  和  $G_2$  确实可以区分，建立的判别准则是有意义的. 如果  $H_0$  不能被拒绝，说明两总体均值的差异不显著，毫无意义，除非考虑其他新的其他变量.





假设  $G_i$  为  $N(\mu^{(i)}, \Sigma_i) (i = 1, 2)$ . 检验两总体的均值是否有显著性差异 (即检验  $H_0 : \mu^{(1)} = \mu^{(2)}$ ). 由马氏距离可以构造检验统计量 -F 统计量

$$F = \frac{n_1 + n_2 - m - 1}{m(n_1 + n_2 - 2)} \frac{n_1 n_2}{n_1 + n_2} d^2(1, 2)$$

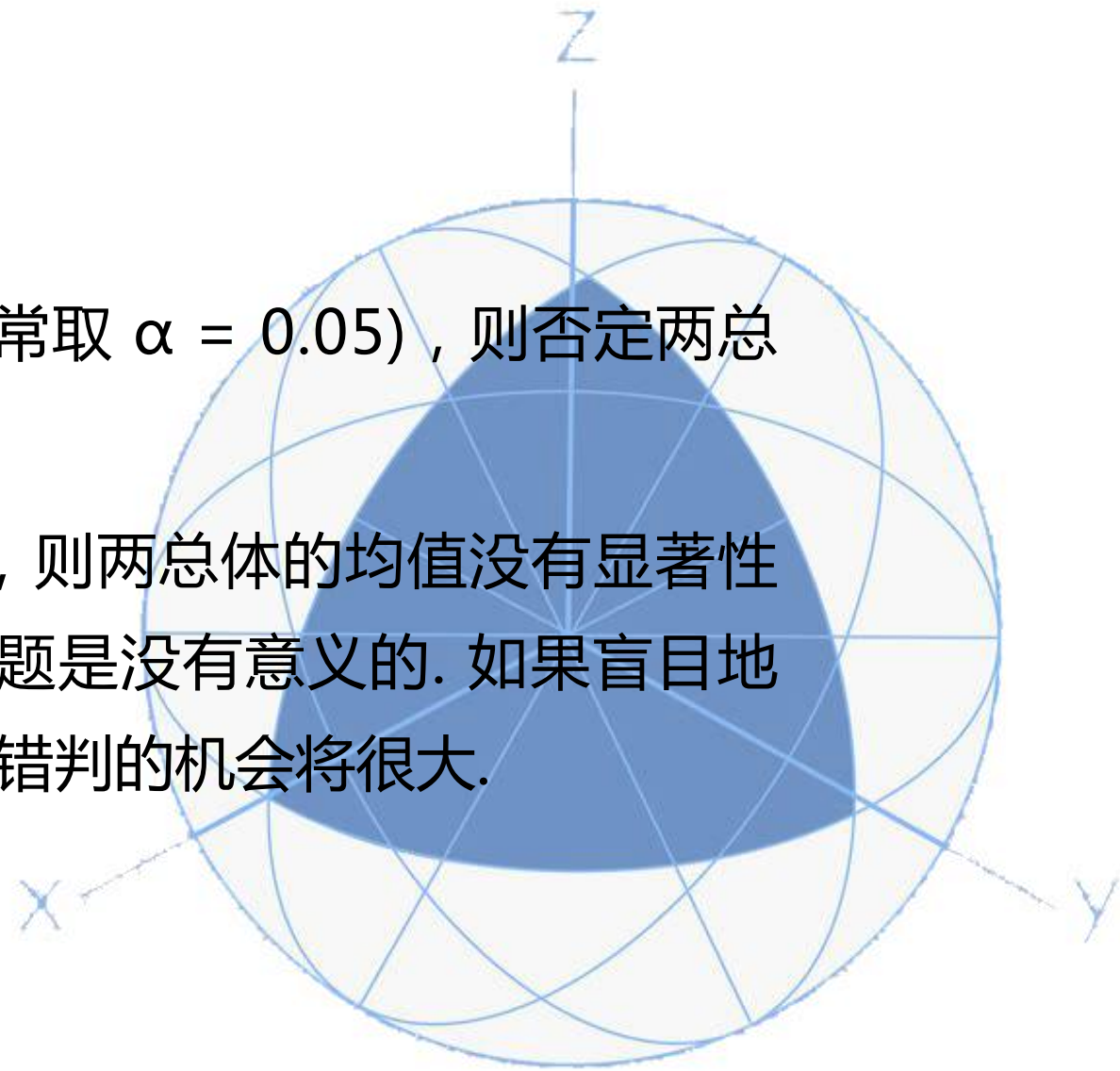
其中  $n_i$  是第  $i$  个总体的样品个数 ( $i = 1, 2$ ). 在两总体均值相等的假设成立下, F 统计量服从分子自由度为  $m$ , 而分母自由度为  $n_1 + n_2 - m - 1$  的 F 分布. 利用样本可计算 F 统计量的值, 由该值还可求出显著性概率值 (p 值).





若  $p$  值小于给定的显著性水平  $\alpha$  (常取  $\alpha = 0.05$ )，则否定两总体的均值是相等的假设.

若  $p$  值大于给定的显著性水平  $\alpha$ ，则两总体的均值没有显著性的差异. 这时讨论两总体的判别问题是没有意义的. 如果盲目地应用以上的方法进行判别归类，则错判的机会将很大.





## 2、多总体判别效果的检验

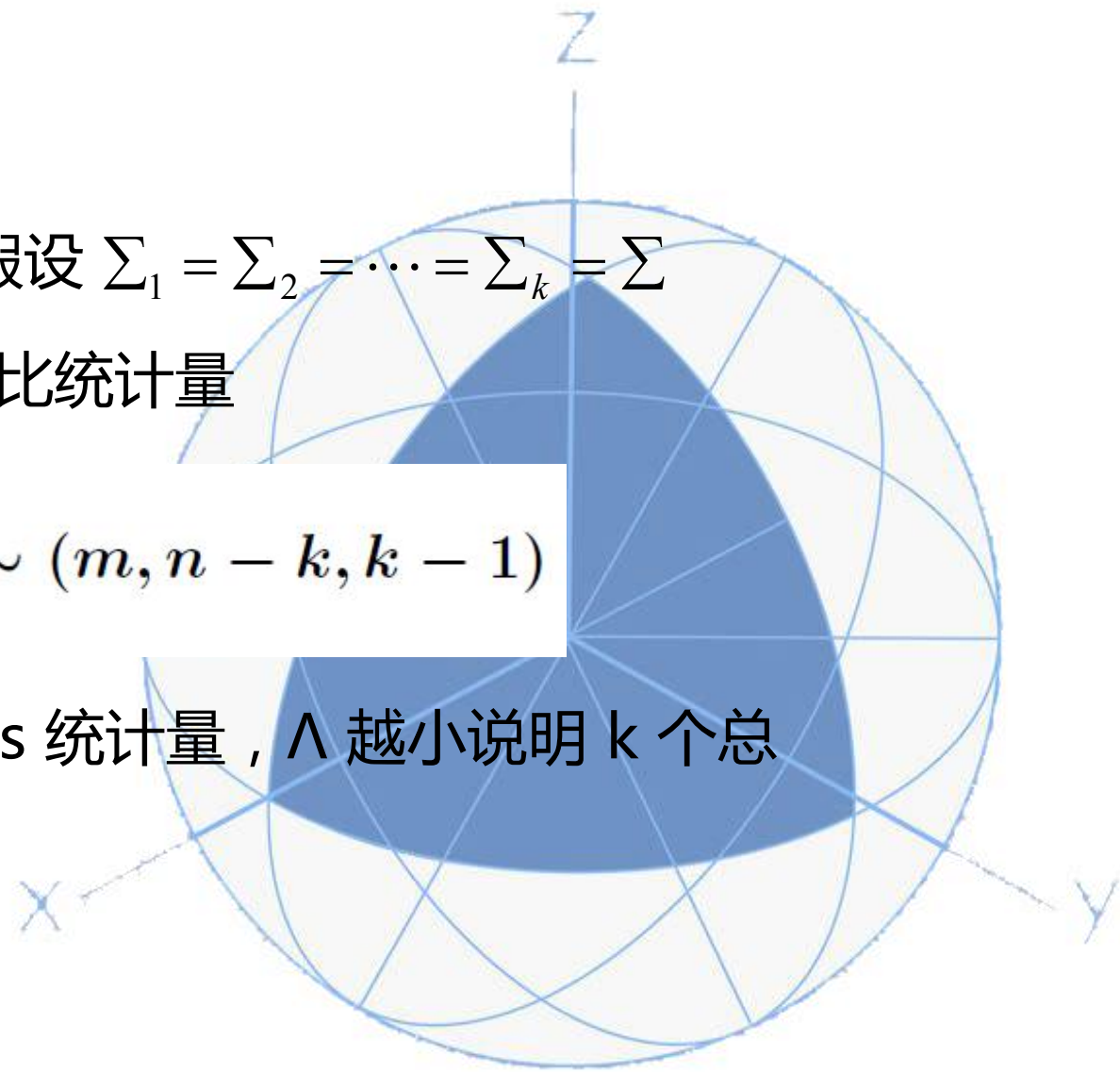
当  $k \geq 2$  时，判别效果的检验问题包括以下两方面：首先检验  $k$  个类的均值向量是否全都相等（即检验  $H_0 : \mu^{(1)} = \mu^{(2)} = \cdots = \mu^{(k)}$ ）；若不全相等，则进一步对  $k$  个总体两两配对，然后逐对检验这两个总体的均值是否有显著差异（检验  $H(ij)_0 : \mu^{(i)} = \mu^{(j)}, i \neq j$ ），也就是检验这两总体的判别效果是否显著. 具体方法仍是通过计算各总体间的马氏距离及  $F$  统计量，并利用  $p$  值的大小来判断其判别效果.



(1) 检验  $H_0 : \mu^{(1)} = \mu^{(2)} = \cdots = \mu^{(k)}$  假设  $\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k = \Sigma$   
利用似然比方法可推导出广义似然比统计量

$$\Lambda = \frac{|A|}{|A + B|} = \Lambda(m) \sim (m, n - k, k - 1)$$

其中  $n = n_1 + n_2 + \cdots + n_k$  ,  $\Lambda$  是 Wilks 统计量 ,  $\Lambda$  越小说明  $k$  个总体的差异就越显著.

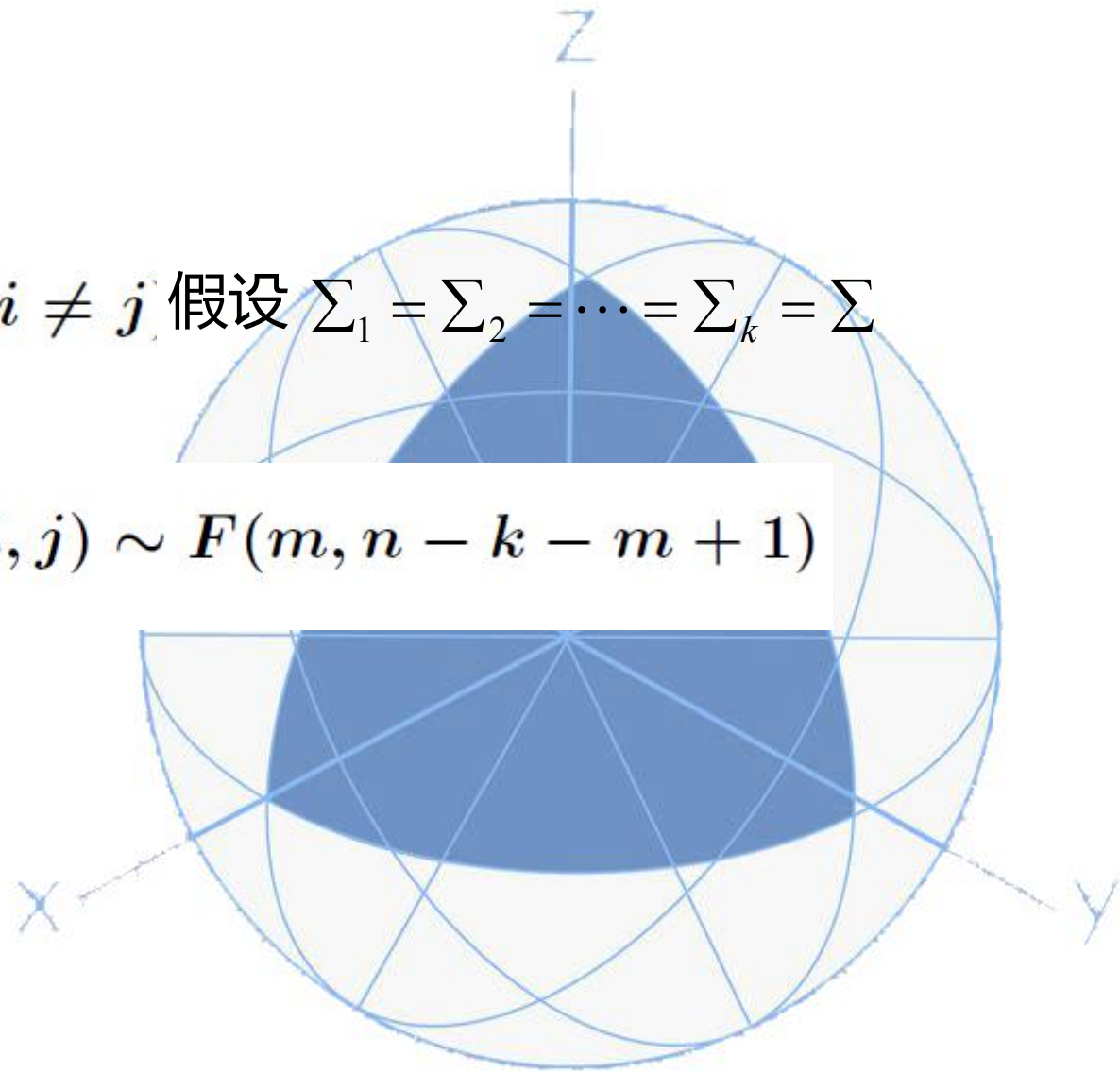




(2) 分别检验  $H(ij)_0 : \mu^{(i)} = \mu^{(j)}, i \neq j$  假设  $\Sigma_1 = \Sigma_2 = \dots = \Sigma_k = \Sigma$

$$F = \frac{n - k - m + 1}{m(n - k)} \frac{n_i n_j}{n_i + n_j} d^2(i, j) \sim F(m, n - k - m + 1)$$

利用 F 统计量对假设  $H_0$  做检验.

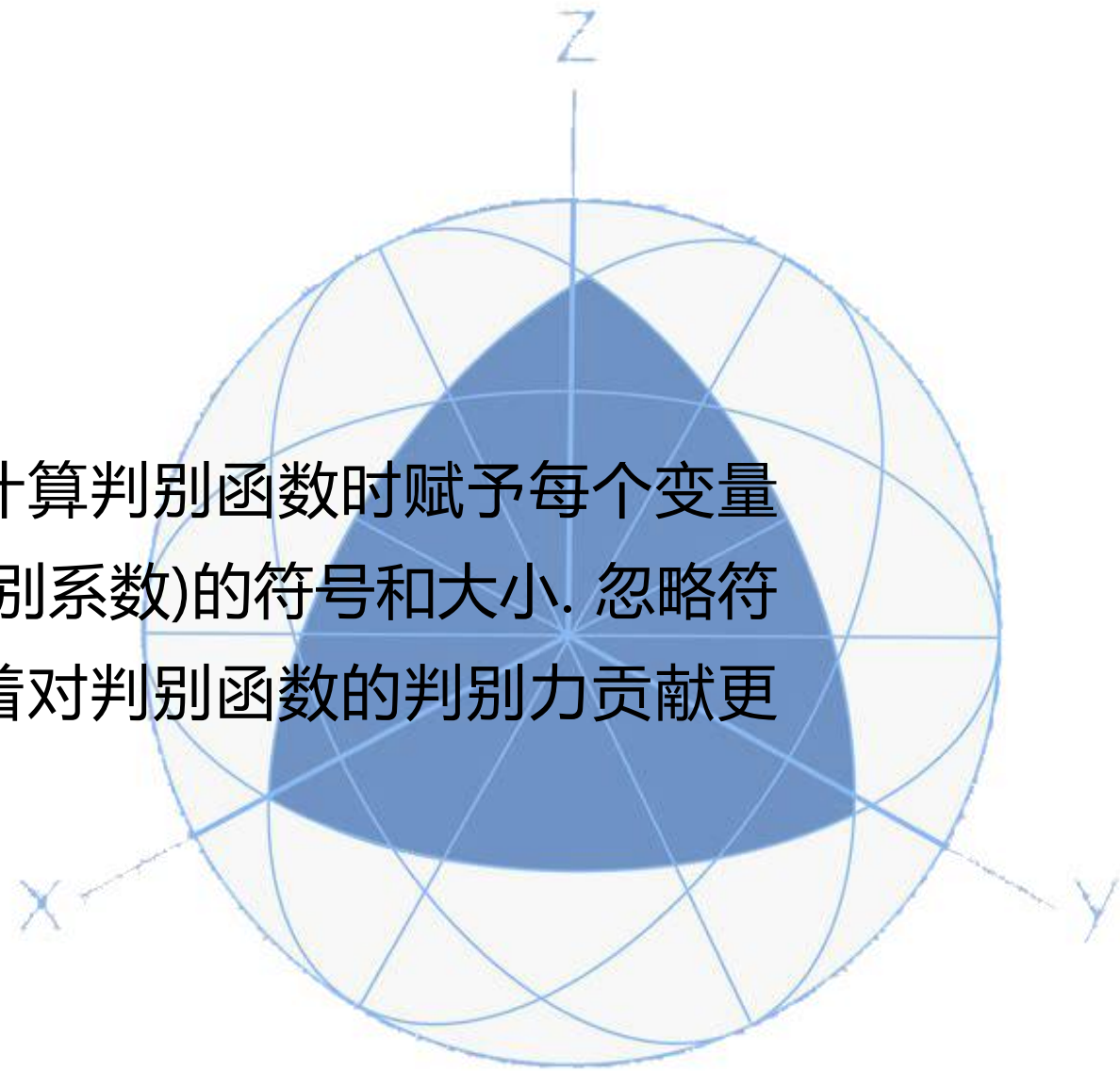






### 3、判别变量重要性的检验

解释判别函数传统的方法是观察计算判别函数时赋予每个变量的标准化判别权重 (有时也称为判别系数) 的符号和大小. 忽略符号时, 较大权重的解释变量意味着对判别函数的判别力贡献更多.





廈門大學  
XIAMEN UNIVERSITY

Part 3

# 案例分析

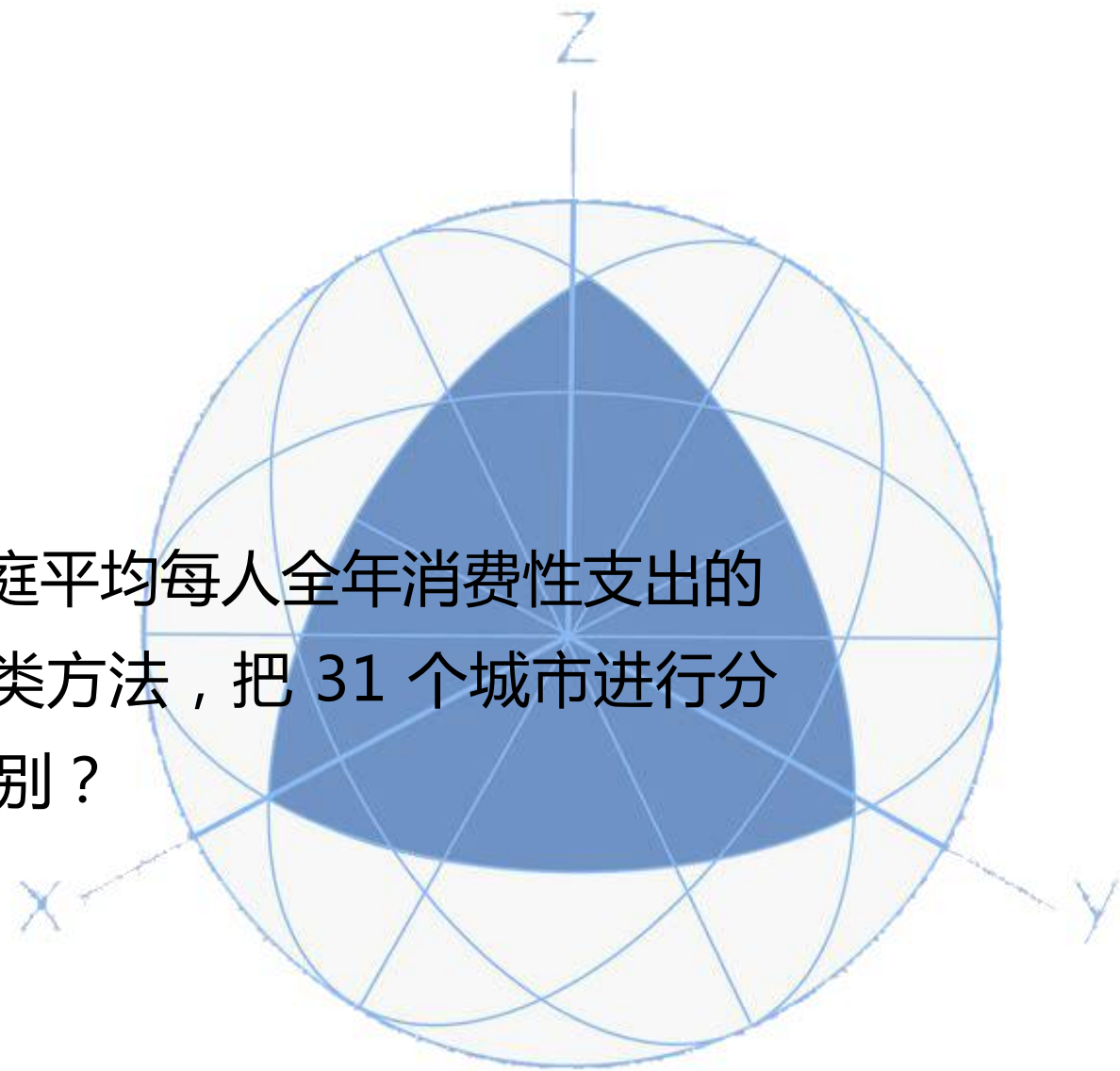




## 案例一 城市聚类

### (1) 问题描述

下表为某年我国 31 个城市居民家庭平均每人全年消费性支出的 8 个主要变量数据. 请用不同的聚类方法, 把 31 个城市进行分类, 并比较不同聚类方法之间的区别?







X1	X2	X3	X4	X5	X6	X7	X8
2959.19	730.79	749.41	513.34	467.87	1141.82	478.42	457.64
2459.77	495.47	697.33	302.87	284.19	735.97	570.84	305.08
1495.63	515.9	362.37	285.32	272.95	540.58	364.91	188.63
1046.33	477.77	290.15	208.57	201.5	414.72	281.84	212.1
1303.97	524.29	254.83	192.17	249.81	463.09	287.87	192.96
1730.84	553.9	246.91	279.81	239.18	445.2	330.24	163.86
1561.86	492.42	200.49	218.36	220.69	459.62	360.48	147.76
1410.11	510.71	211.88	277.11	224.65	376.82	317.61	152.85
3712.31	550.74	893.37	346.93	527	1034.98	720.33	462.03
2207.58	449.37	572.4	211.92	302.09	585.23	429.77	252.54
2629.16	557.32	689.73	435.69	514.66	795.87	575.76	323.36
1844.78	430.29	271.28	126.33	250.56	513.18	314	151.39
2709.46	428.11	334.12	160.77	405.14	461.67	535.13	232.29
1563.78	303.65	233.81	107.9	209.7	393.99	509.39	160.12

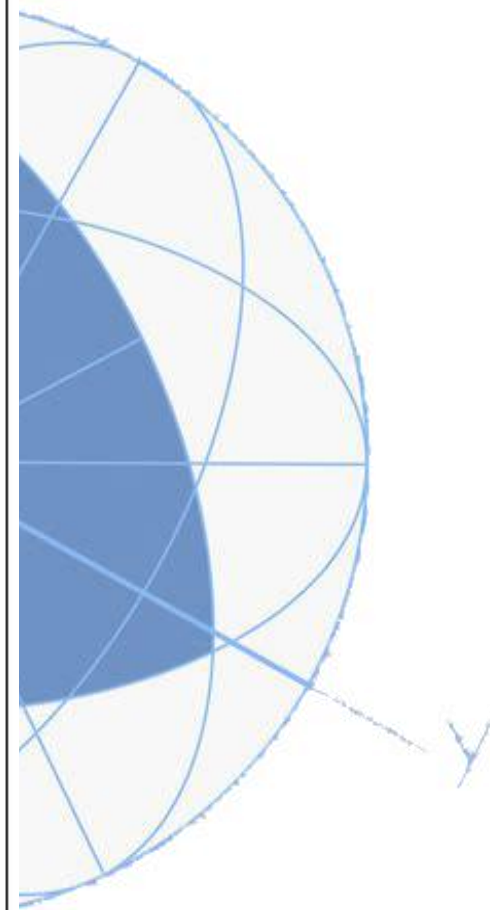






z  
|

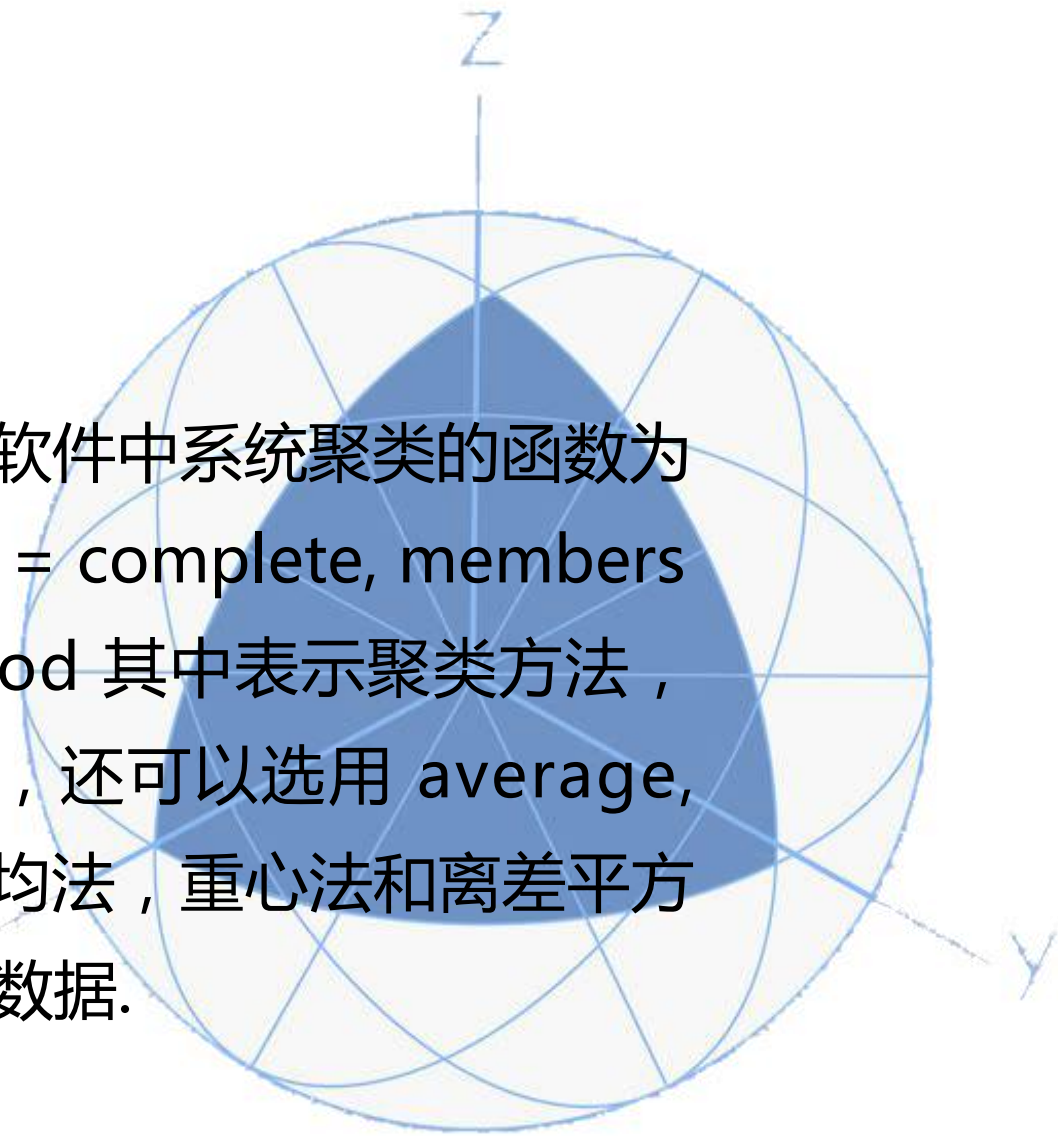
1427.65	431.79	288.55	208.14	217	337.76	421.31	165.32
1783.43	511.88	282.84	201.01	237.6	617.74	523.52	182.52
1942.23	512.27	401.39	206.06	321.29	697.22	492.6	226.45
3055.17	353.23	564.56	356.27	811.88	873.06	1082.82	420.81
2033.87	300.82	338.65	157.78	329.06	621.74	587.02	218.27
2057.86	186.44	202.72	171.79	329.65	477.17	312.93	279.19
2303.29	589.99	516.21	236.55	403.92	730.05	438.41	225.8
1974.28	507.76	344.79	203.21	240.24	575.1	430.36	223.46
1673.82	437.75	461.61	153.32	254.66	445.59	346.11	191.48
2194.25	537.01	369.07	249.54	290.84	561.91	407.7	330.95
2646.61	839.7	204.44	209.11	379.3	371.04	269.59	389.33
1472.95	390.89	447.95	259.51	230.61	490.9	469.1	191.34
1525.57	472.98	328.9	219.86	206.65	449.69	249.66	228.19
1654.69	437.77	258.78	303	244.93	479.53	288.56	236.51
1375.46	480.99	273.84	317.32	251.08	424.75	228.73	195.93





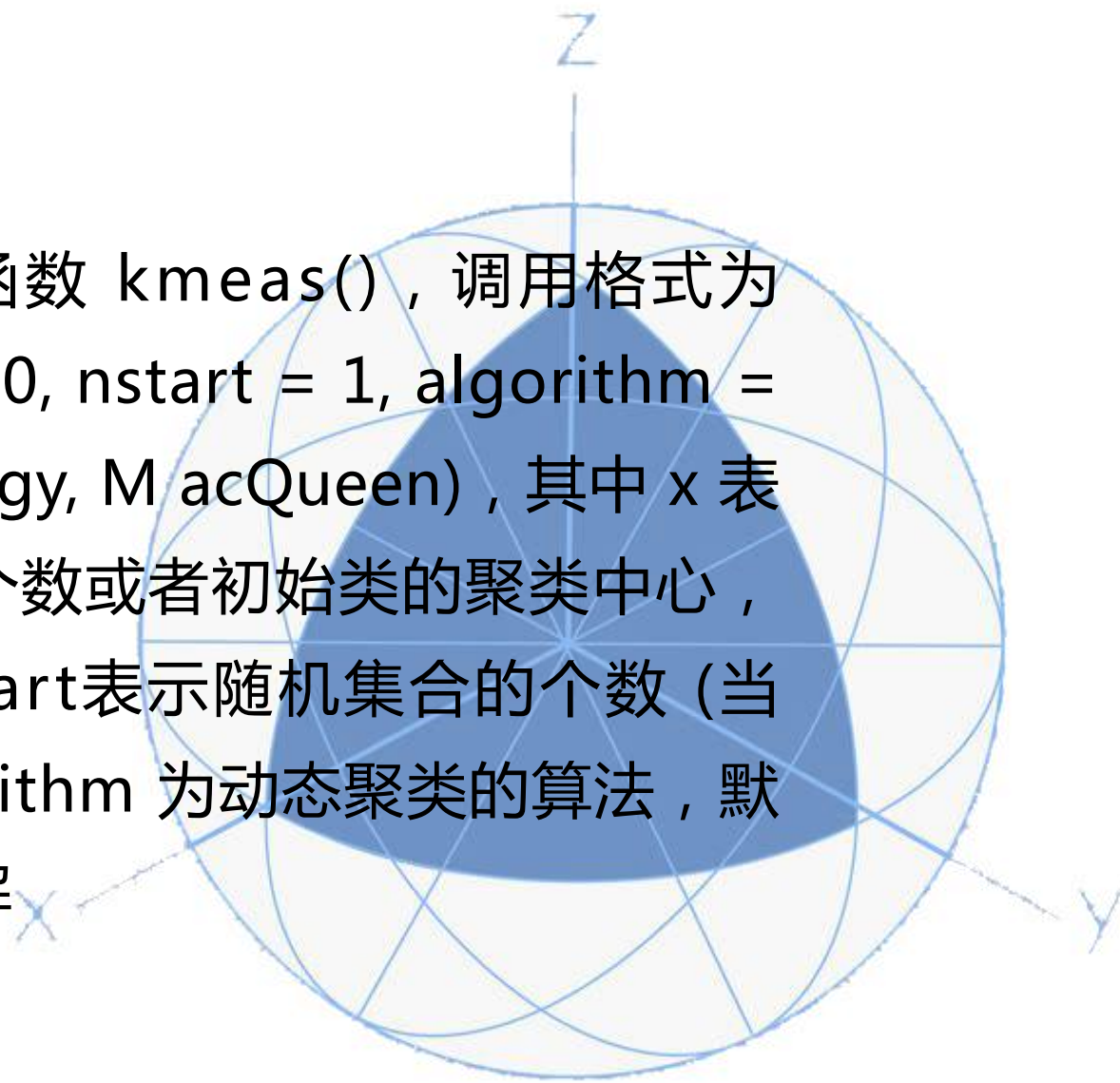
## (2) 问题分析

数据较大时，可借助软件进行分类，R 软件中系统聚类的函数为 `hclust()`，调用格式是 `clust(d, method = complete, members = NULL)`，表示聚类的数据集，`method` 其中表示聚类方法，默认值为 `complete` 表示最长距离法，还可以选用 `average`, `centroid`, `ward` 等方法（分别代表类平均法，重心法和离差平方和法），`number = NULL` 表示使用所有数据。





kmeans 动态聚类方法可调用函数 `kmeans()`，调用格式为 `kmeans(x, centers, iter.max = 10, nstart = 1, algorithm = c(Hartigan - Wong, Lloyd, Forgy, MacQueen))`，其中  $x$  表示数据集， $centers$  表示聚类的个数或者初始类的聚类中心， $iter.max$  表示迭代的次数， $nstart$  表示随机集合的个数（当  $centers$  为聚类的个数时）， $algorithm$  为动态聚类的算法，默认值为 Hartigan - Wong. (3)求解







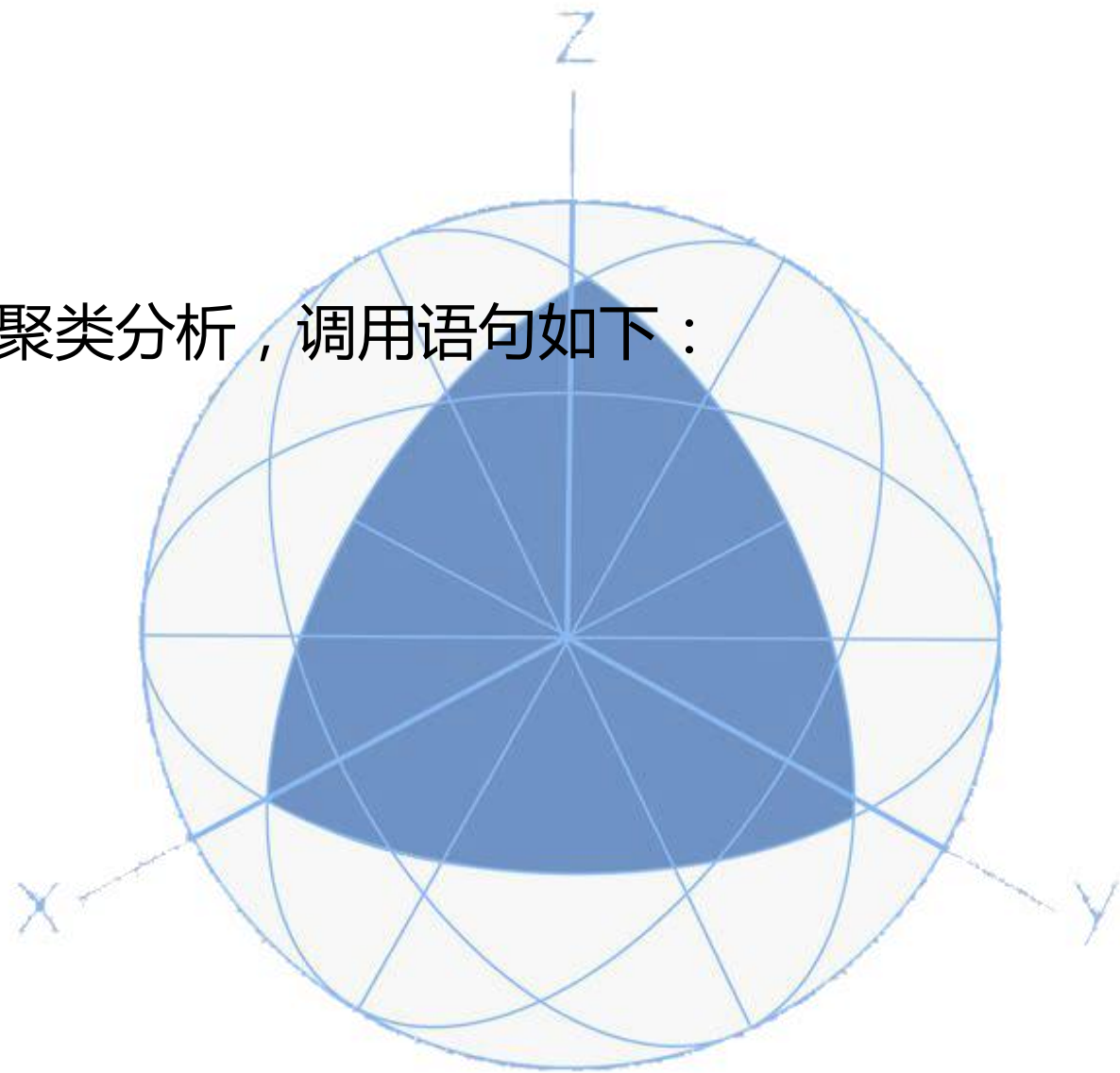
调用 R 程序中 `hclust()` 函数做系统聚类分析，调用语句如下：

```
d <- -dist(scale(X))scale(x)
```

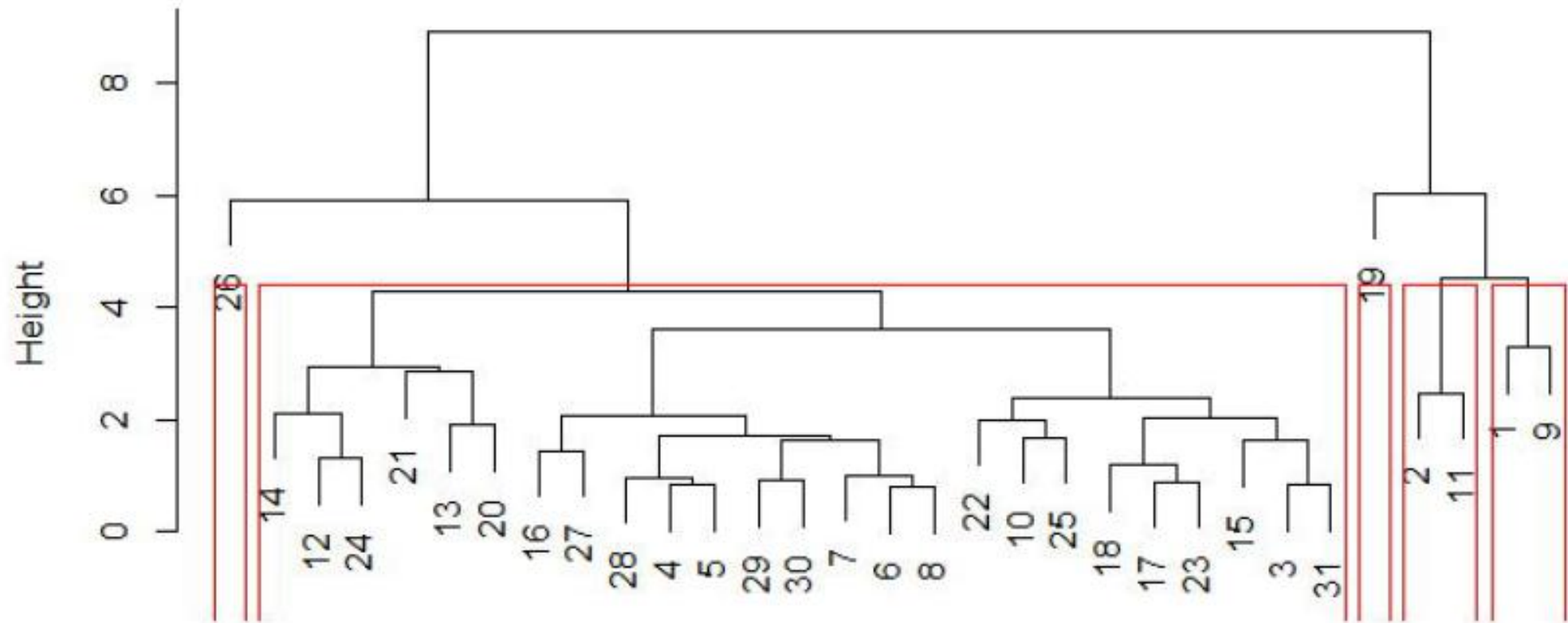
```
h <- -hclust(d)
```

```
plclust(h)
```

```
r <- -rect.hclust(h, 5)
```









由上图可以看出,最后聚为 5 类 :

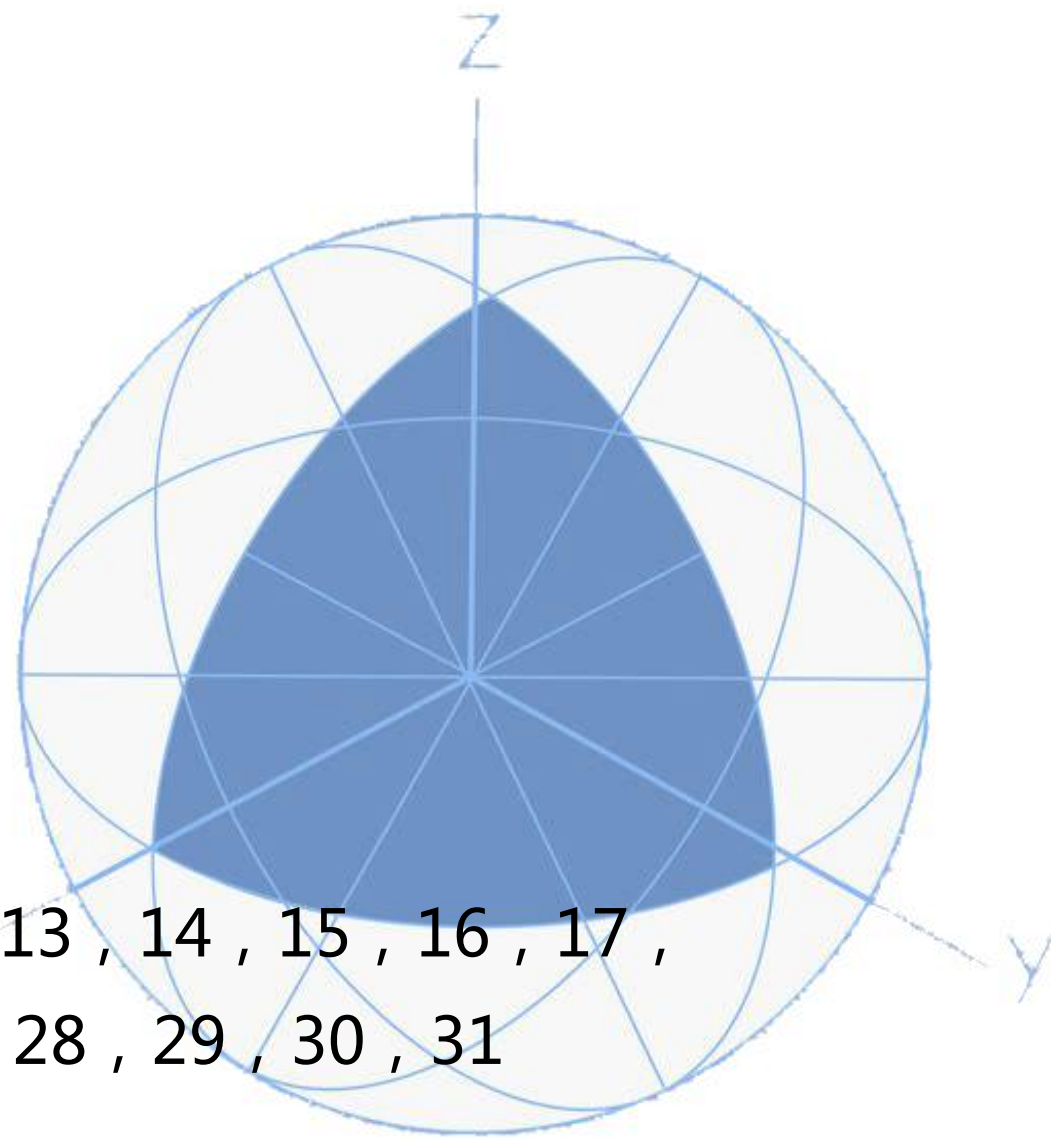
第一类 26

第二类 1, 9

第三类 2, 11

第四类 19

第五类 3, 4, 5, 6, 7, 8, 10, 12, 13, 14, 15, 16, 17, 18, 20, 21, 22, 23, 24, 25, 27, 28, 29, 30, 31





用 kmeans 动态聚类法去分析，调用 R 程序中 kmeans() 函数，

```
km <- kmeans(scale(X), 5, nstart = 20);
```

程序运行结果如下：

```
> km
K-means clustering with 5 clusters of sizes 3, 1, 7, 16, 4

cluster means:
      x1      x2      x3      x4      x5      x6      x7      x8
1  1.8790347  1.02836873  2.1203833  2.1727806  1.49972764  2.2232050  0.95830640  1.9453274
2  1.8042004 -1.12776493  0.9368961  1.2959544  3.90904835  1.6014419  3.88031413  2.0187653
3  0.3906401  0.72770263  0.4284646 -0.1235496  0.08595291  0.2215108 -0.02724055  0.3904549
4 -0.6867323 -0.05815552 -0.4787096 -0.1598851 -0.57749718 -0.5070907 -0.49317064 -0.6033238
5  0.2029830 -1.53019285 -0.6594861 -1.0978219  0.05751333 -0.4270452  0.33154520 -0.2336878

Clustering vector:
[1] 1 3 4 4 4 4 4 4 1 3 1 4 5 5 3 4 4 3 2 5 5 3 4 4 3 3 4 4 4 4 4

within cluster sum of squares by cluster:
[1] 10.191335  0.000000 23.259407 20.384632  9.035662
(between_ss / total_ss = 73.8 %)
```



由上图可以看出每个五个类的聚类中心为，并且这 31 个城市聚为 5 类结果如下：

第一类 2 , 10 , 15 , 18 , 22 , 25 , 26

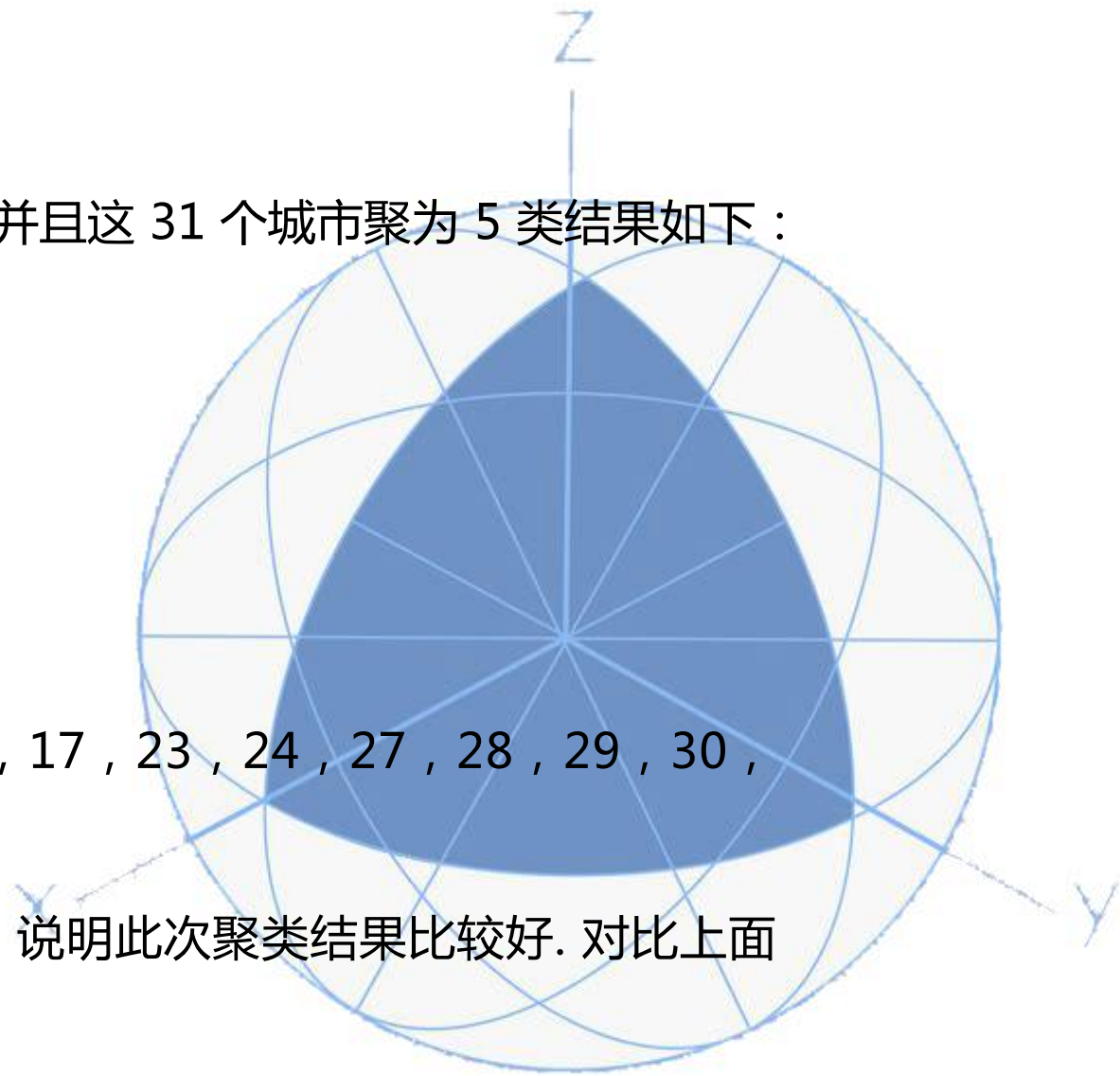
第二类 13 , 14 , 20 , 21

第三类 1 , 9 , 11

第四类 19

第五类 3 , 4 , 5 , 6 , 7 , 8 , 12 , 13 , 16 , 17 , 23 , 24 , 27 , 28 , 29 , 30 , 31

类内部的方差与类之间的方差比为 73.8% ，说明此次聚类结果比较好. 对比上面系统聚类与动态聚类结果相差较大.

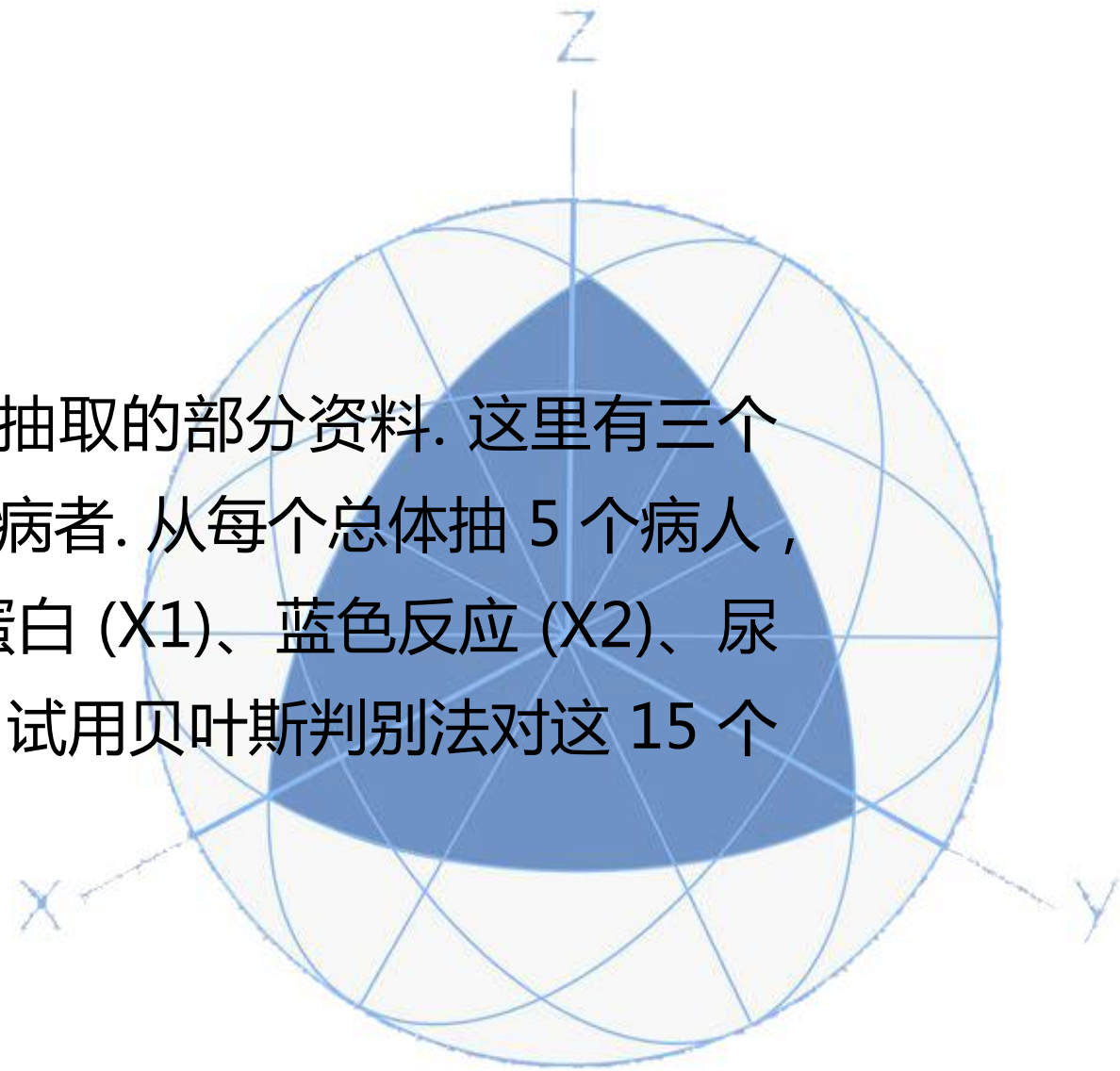






## 案例二 病人识别

(1) 问题背景下表是从病例中随机抽取的部分资料. 这里有三个总体：胃癌、萎缩性胃炎和非胃炎病者. 从每个总体抽 5 个病人，每人化验 4 项生化指标：血清铜蛋白 ( $X_1$ )、蓝色反应 ( $X_2$ )、尿吡啶乙酸 ( $X_3$ ) 和中性硫化物 ( $X_4$ ). 试用贝叶斯判别法对这 15 个样品进行判别归类.





Z  
|

类别	序号	血清铜蛋白	颜色反应	尿吡啶乙酸	中性硫化物
胃癌患者	1	228	134	20	11
	2	245	134	10	40
	3	200	167	12	27
	4	170	150	7	8
	5	100	167	20	14
萎缩性胃炎	1	225	125	7	14
	2	130	100	6	12
	3	150	117	7	6
	4	120	133	10	26
	5	160	100	5	10
非胃炎患者	1	185	115	5	19
	2	170	125	6	4
	3	165	142	5	3
	4	135	108	2	12
	5	100	117	7	2



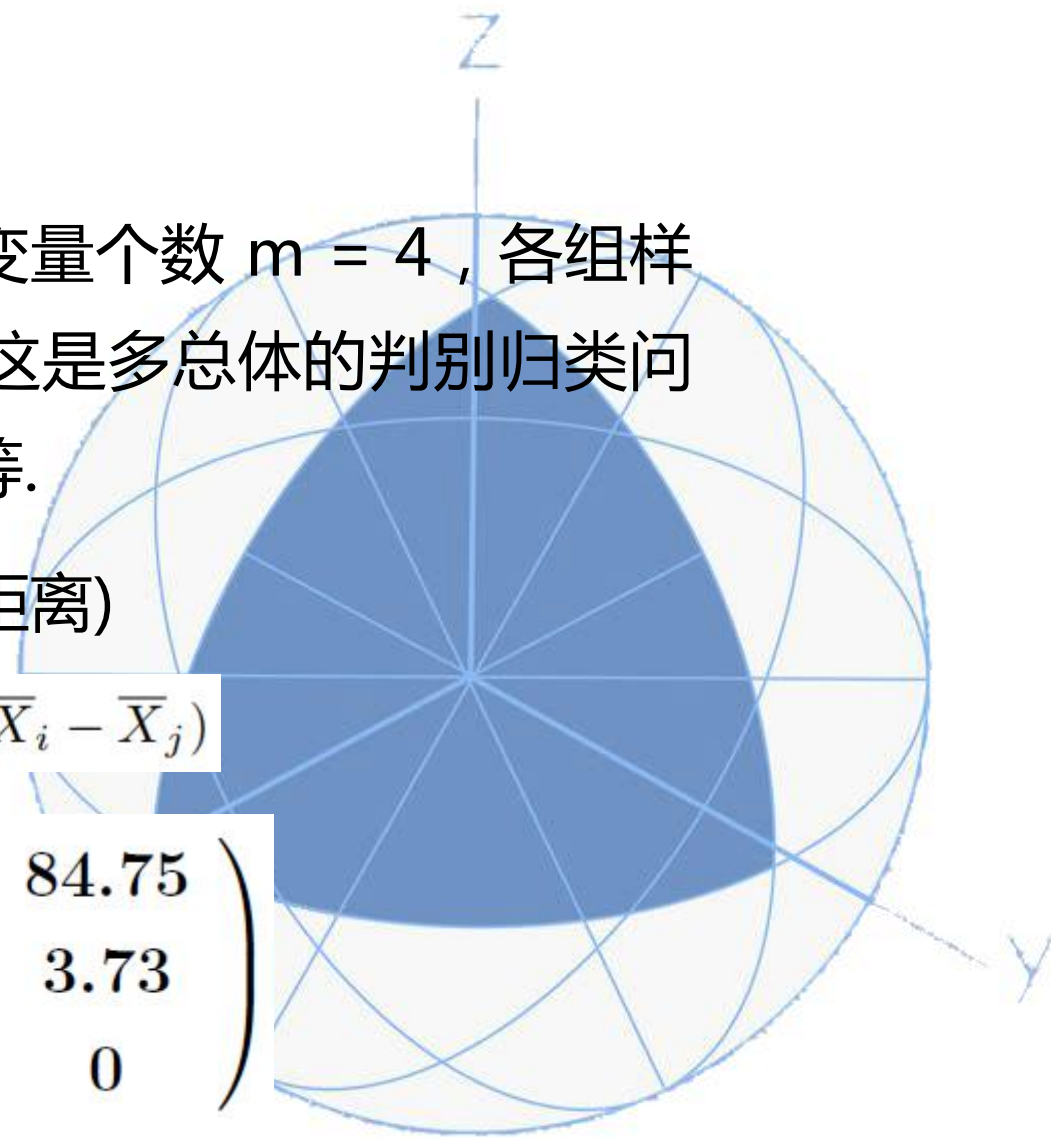


(2) 问题求解此例中总体个数  $k = 3$  , 变量个数  $m = 4$  , 各组样品个数为 :  $n_1 = n_2 = n_3 = 5$  ( $n = 15$ ). 这是多总体的判别归类问题. 假定先验概率相等 , 协方差阵不相等.

计算两两配对的组间平方距离 (即马氏距离)

$$d^2(i, j) = (\bar{X}_i - \bar{X}_j)' S_j^{-1} (\bar{X}_i - \bar{X}_j)$$

$$d = \begin{pmatrix} 0 & 486.03 & 84.75 \\ 22.12 & 0 & 3.73 \\ 19.10 & 284.78 & 0 \end{pmatrix}$$





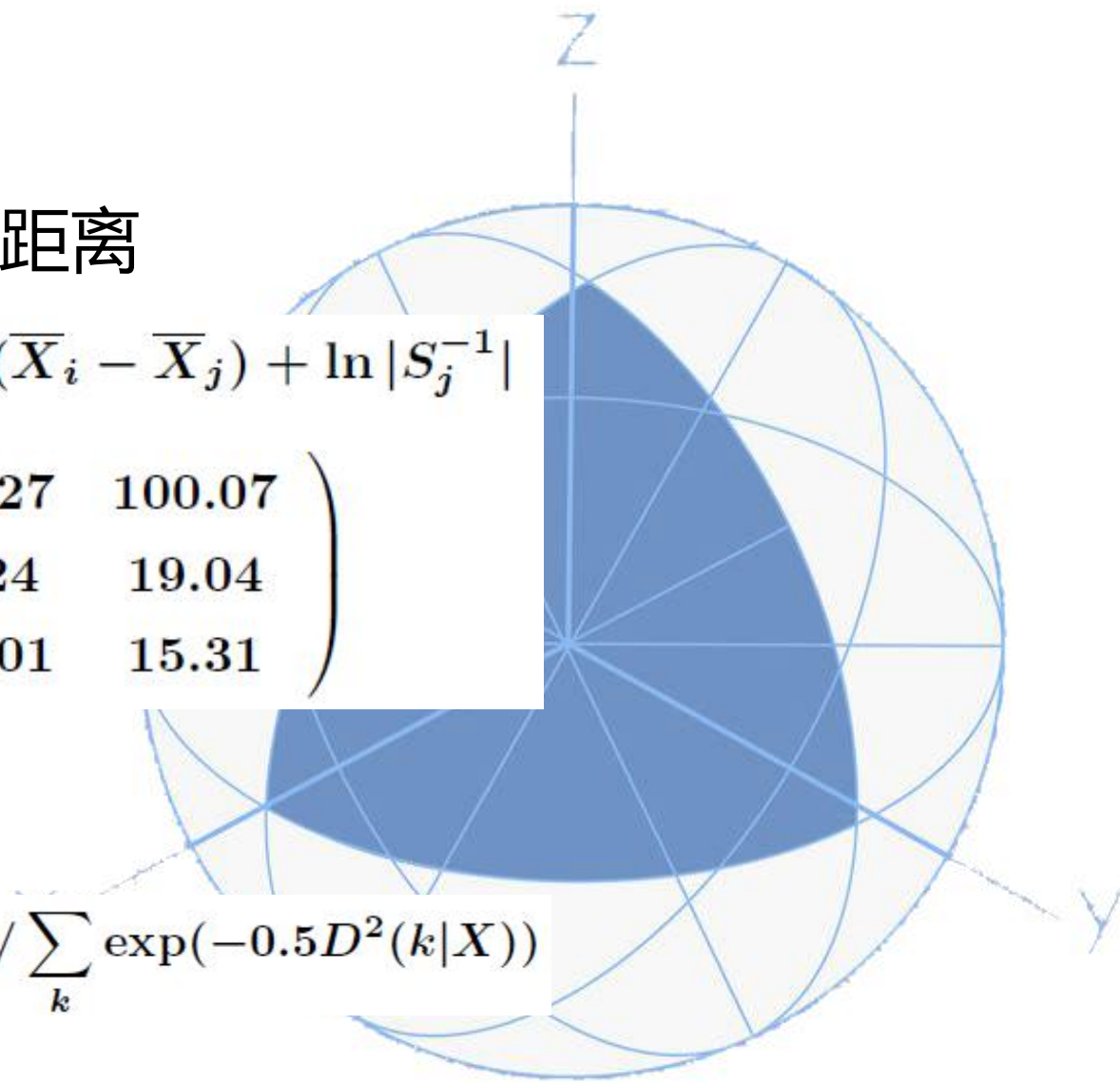
计算两两配对的组间广义平方距离

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

$$D = \begin{pmatrix} 20.95 & 498.27 & 100.07 \\ 43.06 & 12.24 & 19.04 \\ 40.04 & 297.01 & 15.31 \end{pmatrix}$$

计算后验概率

$$P(j|X) = \exp(-0.5D^2(j|X)) / \sum_k \exp(-0.5D^2(k|X))$$

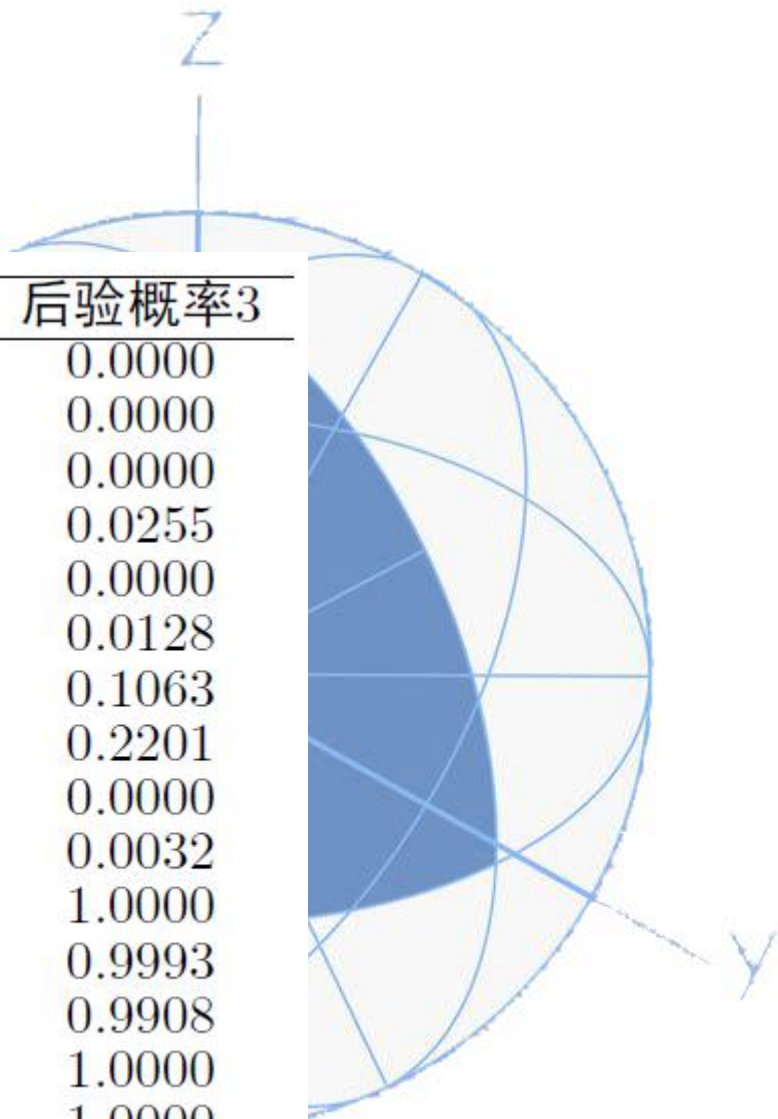






得到结果如下：

序号	原始组	分类组	后验概率1	后验概率2	后验概率3
1	1	1	1.0000	0.0000	0.0000
2	1	1	1.0000	0.0000	0.0000
3	1	1	0.9525	0.0475	0.0000
4	1	1	0.9745	0.0000	0.0255
5	1	1	1.0000	0.0000	0.0000
6	2	2	0.0047	0.9824	0.0128
7	2	2	0.0000	0.8937	0.1063
8	2	2	0.0000	0.7799	0.2201
9	2	2	0.0000	1.0000	0.0000
10	2	2	0.0000	0.9968	0.0032
11	3	3	0.0000	0.0000	1.0000
12	3	3	0.0007	0.0000	0.9993
13	3	3	0.0092	0.0000	0.9908
14	3	3	0.0000	0.0000	1.0000
15	3	3	0.0000	0.0000	1.0000





廈門大學  
XIAMEN UNIVERSITY

THANK YOU

