第一章矩阵代数：

1.正交矩阵：正交矩阵A：p×p的p个列向量和p个行向量都是一组正交单位向量。2.矩阵的逆：(***A***′)−1=(***A***−1)′；|***A***−1|=|***A***|−1；若***A***是正交矩阵，则***A***−1=***A***′；A和A′有相同的特征值；3.特征值/向量和迹：若A和B分别是p×q和q×p矩阵，则AB和BA有相同的非零特征值;若A为实对称矩阵，则A的特征值全为实数，若λi≠λj，特征向量xi和xj必正交，即xi′xj=0;***A***的行列式等于其特征值的乘积；对称矩阵谱分解；tr(***A***)=*λ*1+*λ*2+⋯+*λp* ；4.正定矩阵：若对一切x，有x′Ax≥0，则称A为非负定矩阵，记作A≥0。(1) ***A***>0(或≥0)则有***A***′=***A***，*λi* >0(或≥0)，*i=*1,2,⋯,*p*。(2)设***A***≥0，则***A***的秩等于***A***的正特征值个数。(3)若***A***>0，则***A***−1>0。(4)设***A***≥0，则***A***>0，当且仅当|***A***|≠0。(5)若***A***>0(或≥0)，则|***A***|>0(或≥0)。(6)***BB***′≥0，对一切矩阵***B***成立。(7)若***A***>0(或≥0)，则存在***A***1/2 >0 (或≥0)，使得***A***=***A***1/2***A***1/2，***A***1/2称为***A***的平方根矩阵。

第二章随机向量：

1.数学期望：设***A***,***B***,***C***为常数矩阵，则*E*(***AXB***+***C***)=***A****E*(***X***)***B***+***C***

2.协方差矩阵：；两个独立的随机变量必然不相关，但两个不相关的随机变量未必独立；当x=y时，协方差即为方差，也就是

若Cov(x,y)=0，则称x和y不相关。*V*(***x***)亦记作***Σ***=(*σij*)，其中*σij*=Cov(*xi*,*xj*)，*σii*=*σi*2=*V*(*xi*)。

(1)***Σ***≥0。推论 若|***Σ***|≠0，则***Σ***>0。(2)设***A***为常数矩阵，***b***为常数向量，则；当*p*=1时，退化为熟知的；的分量之间存在线性关系（以概率1）。(3)设A和B为常数矩阵，则。(4)方差相等的两个随机变量的和与差是不相关的。

3.相关矩阵： 若ρ(x,y)=0，则表明x和y不相关。***R***=(*ρij*)和***Σ*** =(*σij*)之间有关系式：***R***=***D***−1***ΣD***−1 其中

另外：

4.标准化变换：

标准化后的协方差矩阵正好是原始向量的相关矩阵。

5.马氏距离：

马氏距离对下列形式的p维向量x度量单位的改变具有不变性：y=Cx+b；当各分量不相关时马氏距离为各分量经标准化后的欧式距离；协差阵为单位阵时马氏距离退化为欧氏距离。

第三章多元正态分布：

1.多元正态分布的定义：一元正态分布N(μ,σ2)的概率密度函数为。

若随机向量的概率密度函数为



则称***x***服从*p*元正态分布，记作***x***~*Np*(***μ***, ***Σ***)，其中，参数***μ***和***Σ***分别为***x***的均值和协差阵。

设***x***~*N*2(***μ***, ***Σ***)，这里易见，*ρ*是*x*1和*x*2的相关系数。当|*ρ*|<1时，可得***x***的概率密度函数为

2.多元正态分布的性质：多元正态分布的任何边缘分布仍为多元正态分布，反之未必；对于多元正态变量而言，其子向量之间互不相关和相互独立是等价的；对于多元正态变量，其子向量的条件分布仍是（多元）正态的。

3.μ和Σ的极大似然估计：



其中称为样本均值向量（简称为样本均值），称为样本协方差矩阵。

4.相关系数ρij的极大似然估计为：

称*rij*为样本相关系数、为样本相关矩阵。

5.无偏性：如果，则称估计量是被估参数***θ***的一个无偏估计，否则就称为有偏的。是***Σ***的有偏估计。*E*(***S***)=***Σ***。

6.复相关系数：

则*y*和***x***的线性函数***l***′***x***（***l*** ≠**0**）间的最大相关系数称为*y*和***x***间的复（或多重）相关系数，记作*ρy*·***x***或*ρy*·1,2,…,*p*，它度量了一个变量*y*和一组变量*x*1,*x*2,…,*xp*间的相关程度。



若*x*1,*x*2,⋯,*xp*互不相关，则有



样本复相关系数即将公式中协方差矩阵和相关矩阵替换成样本协方差矩阵和样本相关矩阵。

7.偏相关系数：称为***x***2给定时***x***1的偏协方差矩阵。称为偏协方差，称为偏方差。

第四章多元正态总体的统计推断：

1.一元情形：均值向量检验取自, 检验

1.已知:





2.未知:





***μ***的置信度为1−*α*的置信区域为：



2.两个总体均值的比较判断：

(1). 两个独立样本：从两个独立样本各抽取一个样本，检验



霍特林T2检验统计量：



拒绝规则为：

(2)成对试验的统计量：di=xi−yi检验：*H*0：***μ***1=***μ***2即*H*0：***δ***=***μ***1−***μ***2=**0**检验统计量为：

拒绝规则为：

3.多个总体均值的比较检验：*k*个总体分布分别是*Np*(***μ***1,***Σ***), 检验*H*0：***μ***1=***μ***2=⋯=***μ****k ；H1：****μ****i≠****μ****j,至少存在一对i≠j*



***T***=***E***+***H ；***相应的自由度：*n*−1=(*n*−*k*)+(*k*−1)

若*Λ*≤*Λ*1−*α*(*p*,*k*−1,*n*−*k*) ，则拒绝*H*0

4.相关系数的检验(简单相关性)

检验*H*0：*ρij*=0 统计量：

拒绝规则为：

第五章判别分析：

1.距离判别 (1)两组距离判别 ①Σ1=Σ2时



误判概率：

从样本出发：

误判概率非参数估计：回代法、划分样本、交叉验证（留一法）

②Σ1≠Σ2时

(2)多组距离判别

*d*2(***x***,*πi*)=***x***′***Σ***−1***x***−2(***I****i*′***x***+*ci*) 其中

实践中

.

判别分类是否有效：各组均值向量之间应有明显的差异；

选择线性/二次：数据量小线性判别，反之二次判别；对Σ进行齐次检验；直觉判断是否相等；两个都用比较正确率。

2.贝叶斯判别 1.最大后验概率法

. 若为正态组：

*D*2(***x***,*πi*)=*d*2(***x***,*πi*)+*gi*+*hi*

若协差阵全相等：

若无先验信息则pi全相等：

2.最小期望代价误判法

两组的一般情形

两个正态组的情形

.多组的情形 使ECM达到最小的判别规则是：

总误判概率：

总正确概率：

3.费希尔判别：需假定,组间平方和及组内平方和:



约束条件***a***′***S****p****a***=1下,最大化*Δ*(***a***)；是Σ的联合无偏

即求***E***−1***H*** 的特征值和特征向量ti；*y*1=***t***1′***x***为费希尔第一线性判别函数。4.逐步判别： 1.附加信息检验 



*拒绝规则为：F*≥*Fα*(*k*−1,*n*−*k*−*p*+1)，则拒绝*H*0

2.变量的选择方法:前进法，后退法，逐步判别法.

第六章聚类分析：

明氏距离，标准化处理.

名义尺度变量的定义:

变量的夹角余弦为,相关系数为

1.系统聚类法

类平均法

重心法

离差平方法(ward)

1.最短距离法最容易产生结，且有一种挑选长链状聚类的倾向，称为链接倾向2.最长距离法容易被异常值严重地扭曲3. 类平均法较好地利用了所有样品之间的信息，在很多情况下它被认为是一种比较好的系统聚类法4.重心法在处理异常值方面更稳健，但是在别的方面一般不如类平均法或离差平方和法的效果好5. 离差平方和法使得两个大的类倾向于有较大的距离，因而不易合并；相反，两个小的类却因倾向于有较小的距离而易于合并。这往往符合我们对聚类的实际要求。最短距离法、最长距离法和类平均法都属于连接方法，它们既可以用于样品的聚类，也能够用于变量的聚类。不过并非所有的系统聚类方法都适用于对变量的聚类

第七章主成分分析：

1.定义：设*E*(***x***)=***μ***，*V*(***x***)=***Σ，***进行如下的线性变换在约束条件||a1||=1最大化；2.解法：求***Σ***的特征值λi和特征向量ti，第一主成分为

它的方差具有最大值*λ*1；T是正交矩阵，，

主成分的协差阵*V*(***y***)=***Λ=diag(λ1,λ2,⋯,λp)*** ；主成分的总方差；主成分*yi*的贡献率；原始变量*xi*与主成分*yk*之间的相关系数***x***=***Ty***即***xi=ti1y1+ti2y2+…+tipyp*** 所以

***Cov(xi,yk)=Cov(tikyk, yk)=tikλk，***

原始变量对主成分的影响*yk*=*t*1*kx*1+*t*2*kx*2+⋯+*tpkxp*称*tik*为*yk*在*xi*上的载荷，它反映了*xi*对*yk*的重要程度。方差大的主成分与方差大的原始变量有较密切的联系，而方差小的主成分与方差小的原始变量有较强的联系。通常我们取前几个主成分，因此所取主成分会过于照顾方差大的原始变量，而对方差小的原始变量却照顾得不够。*yp*的贡献率常常很小，可视作接近于一个常数（均值）。*yp*可能揭示出原始变量之间存在着一个意外的多重共线性关系。3.从相关阵出发求主成分：用于各变量的单位不全相同的情形或变量方差的差异较大（在应用中常表现为各变量数据间的数值大小相差较大）的情形。的协差阵正是***x***的相关阵***R***；性质：*E*(***y***\*)=**0**，*V*(***y***\*)=***Λ***\*，其中；；变量与主成分之间的相关系数即有因此，在解释主成分时，从相关阵***R***出发求得的载荷和相关系数所起的作用是完全相同的，只需选其一用来作主成分解释即可。标准化后的结论完全可能会发生很大的变化，因此标准化不是无关紧要的。4.样本的主成分：用代替即可.各主成分之间的样本协方差为0.总样本方差与和计算与前文相同.中心化的主成分 从出发求主成分与上述流程一致.但是把换成了.  应用：在本身作为目标的主成分分析中，我们首先应保证所提取的前几个主成分的累计贡献率达到一个较高的水平，其次对这些被提取的主成分必须都能够给出符合实际背景和意义的解释主成分的解释其含义一般多少带有点模糊性，不像原始变量的含义那么清楚、确切，这是变量降维过程中不得不付出的代价。因此，提取的主成分个数 m通常应明显小于原始变量个数p（除非p本身较小），否则维数降低的“利”可能抵不过主成分含义不如原始变量清楚的“弊”

第八章因子分析：

1.与主成分分析的区别: 1.主成分分析涉及的只是一般的变量变换，它不能作为一个模型来描述，本质上几乎不需要任何假定；而因子分析需要构造一个因子模型，并伴有几个关键性的假定;2. 主成分是原始变量的线性组合；而在因子分析中，原始变量是因子的线性组合，但因子却一般不能表示为原始变量的线性组合; 3.在主成分分析中，强调的是用少数几个主成分解释总方差；而在因子分析中，强调的是用少数几个因子去描述协方差或相关关系主成分的解是唯一的（除非含有相同的特征值或特征向量为相反符号）；4.而因子的解可以有很多，表现得较为灵活（主要体现在因子旋转上），这种灵活性使得变量在降维之后更易得到解释，这是因子分析比（需对主成分作出解释的）主成分分析有更广泛应用的一个重要原因;5. 主成分不会因其提取个数的改变而变化，但因子往往会随模型中因子个数的不同而变化

2.正交因子模型：***x***=***μ***+***Af***+***𝛆*** f为公共因子向量，***𝛆***为特殊因子向量， 称为因子载荷矩阵。通常假定

3.正交因子模型的性质1. x的协差阵∑的分解:2. 因子载荷不为一,作正交变换,. ***x***=***μ***+***A***\****f***\*+***ε*** E(f\*)=T′E(f)=0；V(f\*)=T′V(f)T=T′T=I；Cov(f\*,ε)= E(f\*ε′)=T′E(fε′)=0仍满足模型条件。***Σ***也可分解为***Σ***=***A***\****A***\*′+***D***

4.因子载荷矩阵的统计意义①A的元素：

②A的行元素平方和：

反映了公共因子对*xi*的影响，可以看成是公共因子*f*1,*f*2,⋯,*fm*对*xi*的方差贡献，称为共性方差；是特殊因子*εi*对*xi*的方差贡献，称为特殊方差。当***x***为各分量已标准化了的随机向量时，*σii*=1；③A的列元素的平方和： 反映了公共因子*fj*对*x*1,*x*2,⋯,*xp*的影响，是衡量公共因子*fj*重要性的一个尺度，可视为公共因子*fj*对*x*1,*x*2,⋯,*xp*的总方差贡献。*fj*所解释的总方差的比例（或称贡献率）为，如果各原始变量已作了标准化，则该比例就简化为。④A的元素平方和：或这是*f*1, *f*2,⋯, *fm*对总方差的累计贡献，*f*1, *f*2,⋯, *fm*所解释的总方差的累计比例（或称累计贡献率）为对于标准化了的原始变量可简化为5.参数估计主成分法协方差矩阵的特征值λ和正交单位特征向量t，其中 这里的A和D就是因子模型的一个主成分解。

主因子法

由此我们可以重新估计特殊方差，的最终估计为极大似然法



 6.因子旋转(正交旋转) 对公共因子



 7.因子得分回归法



第九章对应分析：

1.行、列独立性检验 统计量为当独立性的原假设为真，且样本容量*n*充分大，期望频数拒绝规则为则拒绝独立性。

2.总惯量度量行列变量关联性

总惯量也可度量行轮廓之间的总变差和列轮廓之间的总变差。行和列之间的关联性越强，行（列）轮廓之间的差异性就越大；反之亦然。总惯量为零与以下三种情形的任一种等价：或；所有的行轮廓相等；所有的列轮廓相等；所以，如果行变量与列变量相互独立，则我们可以期望（由样本数据构成的）列联表中所有的行有相近的轮廓，所有的列亦有相近的轮廓。总惯量的分解： 

3. 行列轮廓的坐标将中心化的行轮廓和中心化的列轮廓放在同一k维坐标系中前者的坐标为(*xi*1,*xi*2,⋯,*xik*)，后者的坐标为(*yj*1,*yj*2,⋯,*yjk*)，各行和列点在第*i*维坐标轴上坐标的加权平均值为0 即各行点和列点在第*i*坐标轴上的坐标平方的加权平均都等于，称之为第*i*主惯量或第*i*惯量。主惯量度量了在每一坐标轴上的变差，类似于主成分的方差。总惯量可以分解为各主惯量之和。4.行点和列点相近的意涵：在累计贡献率足够大的条件下如果一个行点和一个列点相近，则表明行、列两个变量的相应类别组合发生的实际频数一般会高于这两个变量相互独立情形下的期望频数，也就意味着该行类别与该列类别相关联。一般地，对于相近的行点和列点，它们离原点越远，其关联性就越强，也就是其类别组合的实际频数越是明显高于两变量独立情形下的期望频数。如果它们都在原点附近，则其关联性一般较弱、甚至可能几乎无关联性。

第十章典型相关分析：

1.典型相关：设***x***=(*x*1,*x*2,⋯,*xp*)′和***y***=(*y*1,*y*2,⋯,*yq*)′是两组随机变量，***Σ***21=***Σ***12′研究*u*=***a***′***x***与*v*=***b***′***y***之间的相关关系，最大化*ρ*(*u*,*v*)=***a***′***Σ***12***b***；的特征值，设相应于的正交单位特征向量为***β***1,***β***2,⋯,***β****m*，令；***α***1,***α***2,⋯,***α****m*为相应于的正交单位特征向量。***a***1,***a***2,⋯,***a****m*为相应于的特征向量。***b***1,***b***2,⋯,***b****m*为相应于的特征向量。为第一对典型变量，称*a*1，*b*1为第一对典型系数向量，称*ρ*1为第一典型相关系数。2.典型相关变量的性质：1.同一组的典型变量互不相关*V*(*ui*)=***a****i*′***Σ***11***a****i*=1，*V*(*vi*)=***b****i*′***Σ***22***b****i*=1 *ρ*(*ui*,*uj*) = Cov(*ui*,*uj*) = ***a****i*′***Σ***11***a****j*=0，*ρ*(*vi*,*vj*) = Cov(*vi*,*vj*) =***b****i*′***Σ***22***b****j*=0 2.不同组的典型变量之间的相关性 3.原始变量与典型变量之间的相关系数

4.典型相关系数也是某种复相关系数3.典型相关系数的显著性检验1.全部总体典型相关系数均为零的检验*H*0：*ρ*1=*ρ*2=⋯=*ρm*=0 统计量

若，则拒绝*H*0 认为典型变量*u*1与*v*1之间的相关性是显著的；否则，就认为第一典型相关系数不显著。

进一步检验*H*0：*ρ*2=*ρ*3=0 若原假设*H*0被接受，则认为只有第一对典型变量是显著的；

简答题补充：1. 欧氏距离与马氏距离的优缺点是什么?每个坐标对欧氏距离的贡献是同等的。它将样品的不同属性之间的差别等同看待，这一点有时不能满足实际要求.没有考虑到总体变异对距离远近的影响。如果协方差矩阵为单位矩阵，那么马氏距离就简化为欧氏距离，如果协方差矩阵为对角阵，则其也可称为正规化的欧氏距离。两点之间的马氏距离与原始数据的测量单位无关。还可以排除变量之间的相关性的干扰。缺点：夸大了变化微小的变量的作用。受协方差矩阵不稳定的影响，马氏距离并不总是能顺利计算出。2.应用判别分析应该具备什么样的条件？假设每一个判别变量不能是其他判别变量的线性组合，各组变量的协方差矩阵相等. 各判别变量之间具有多元正态分布。3.总惯量的意义：反映了行剖面集定义的各点与其重心加权距离的总和，也反映了两个属性变量各状态之间的相关关系。4.系统聚类的基本思想是：开始时将n个样品各自作为一类，并规定样品之间的距离和类与类之间的距离，然后将距离最近的两类合并成一个新类，计算新类与其他类的距离；重复进行两个最近类的合并，每次减少一类，直至所有的样品合并为一类

填空题：Q型聚类分析是对 样品 的分类，R型聚类分析是对变量\_的分类。因子分析中，把变量表示成各因子的线性组合，而主成分分析中，把主成分表示成各 变量的线性组合。类平均法的两种形式为组间联结法和组内联结法。随机向量 的协方差阵一定是对称的半正定阵。