1. 矩阵运算与高等代数

* 常用结论
* 若是正交矩阵，则也是正交矩阵，
* 若是属于的特征向量,则,且也为特征向量
* 相似矩阵特征多项式相同；与非零特征值及重数相同
* 关于实对称矩阵：特征值为实数；不同特征值的特征向量必正交；必正交相似于对角矩阵；必可对角化；对角线元素介于特征值之间；正定等价于特征值均大于0；正定等价于存在实可逆矩阵的分解；若正定则存在唯一正定矩阵使得；若正定则
* 施密特三角化分解（QR分解）

设为矩阵，则存在分解,其中为列满秩矩阵，为对角线元素非负的上三角矩阵

* 施密特正交化

设线性无关，令,则正交

* 矩阵拉直与Kronecker积
  + 拉直：设是一个矩阵,则
  + Kronecker积：设分别为的矩阵，定义的矩阵，称为矩阵的Kronecker积。

1. 数据可视化与R语言

* 函数plot()对应不同数据类型时绘制的图形

数值/数值,数值——散点图；因子/一维频数分布表——条形图；

因子,因子-脊形图；二维列联表——马赛克图；

数值,因子——箱线图；数据框——散点图矩阵

* 数据可视化的图形
  + 轮廓图（平行坐标图、多线图）：可以比较各样本在多个变量上取值的异同；也可以直观地看出单个变量数值的分散程度等。
  + 雷达图（蜘蛛图）：以二维图表形式展示多变量数据。
  + 星图：利用个边形，可以比较个样本的相似性。
  + 脸谱图：将个维度的数据用人脸部位的形状或大小来表示。
  + 散点图：可以直观看出两个变量之间相关关系及相关的程度。
  + 气泡图：三维散点图的变化，气泡大小表示第三个变量的大小.

1. 多元正态分布

* 联合分布函数



* 概率密度函数的性质

若在处连续，有:

* 边缘概率密度函数

若维随机向量作为整体，有联合分布函数，其任意一个分量也是一个维的随机向量（其中），也有自己的联合分布函数，记为：



* 随机向量的相互独立及等价条件
  + 若对一切和有下式成立，则称和相互独立:



* + 和相互独立等价于：

或

* 随机向量的条件分布

给定条件下，的条件分布函数为：

条件概率密度函数为：

其中为的边缘密度。

* 协方差
  + 单个随机向量的协方差

设是维随机向量，则的协方差矩阵为：



记为，其中

* + 两个随机向量的协方差

设和分别是维和维的两个随机向量，则和的协方差矩阵为：



* 相关系数矩阵

为的相关系数矩阵，其中：



其中为随机变量的方差。

* 均值向量与协方差矩阵的计算



* 协方差矩阵的性质
  + 当和时，
  + 当和独立时，
  + 若，则称随机向量和不相关
  + 协方差矩阵是对称非负定矩阵
  + 记为标准差对角阵，则：

,

* 随机向量的变换

定义随机向量的联合密度函数为，维空间到空间的一一对应实值函数,，随机向量为，存在逆变换，有雅克比矩阵：



其中为绝对值，那么就得到的概率密度函数：



* 标准多元正态分布的性质

设，为元向量，和为对称矩阵，则有：

* 多元正态分布的概率密度函数

若是一个维随机向量，为阶正定矩阵，则：



* 多元正态分布的其他三个等价定义
  + 基于一元正态分布作线性变换：

设为独立同分布随机变量序列，，是常数矩阵，是常数向量，则称的线性组合为元正态分布，且满足

* + 特征函数：若特征函数满足下式，则服从元正态分布：



* + 基于一元正态分布的线性组合：若的任意线性组合均服从一元正态分布，则称为元正态分布。
* 多元正态分布的运算性质
  + 线性变换：设，，则有：



* + 可加性：设是相互独立的组维随机向量，，则有：
  + 二次型：设，对于阶矩阵，二次型服从中心分布的充要条件为，且其自由度为
* 多元正态分布的边际分布

设，，将,和剖分为：



则，

* 多元正态分布的条件分布

给定时，的条件分布服从元正态分布，即：



其中，

* 多元正态分布的边际、条件分布的独立性
  + ，当且仅当时等号成立，此时与相互独立。
  + 与相互独立。
  + 若与相互独立，则存在维实数向量和维实数向量，满足和独立。
* 矩阵正态分布

设为随机矩阵，若，则称服从矩阵正态分布,记作,其中令，则有

1. 多元正态总体的抽样分布

* 二次型分布（卡方分布）
  + 定义：设的随机变量相互独立，令，且则，其中非中心参数，表示的二次型.
  + 性质4.1.1、4.1.2、4.1.3：设，则有：

，

其中，

* + 性质4.1.4、4.1.5： 为阶对称矩阵，为矩阵，则
  + 性质4.1.6、4.1.7：设且秩为

则有：

* + 性质4.1.8：为阶对称矩阵，则有：



* Wishart分布
  + 定义：设的随机变量相互独立，记，则称阶矩阵，即服从Wishart分布，其中称为自由度。
  + 均值、变换、可加性：若，则



* + 矩阵二次型：若为随机样本矩阵，为幂等矩阵，则矩阵二次型服从Wishart分布，其中；与独立的充分必要条件是.
  + 行列式、逆矩阵期望、逆矩阵性质：设,，则，其中独立,且;对于任意非零维向量,总有
  + Bartlett分解：设，作分解，其中为正对角线元素的下三角矩阵。令，则相互独立，且时；时.
* Hotelling T2分布
  + t分布：设且独立，则
  + Hotelling T2分布：设且相互独立，则，称为服从自由度为的Hotelling T2分布，该分布只与有关，与无关 。
  + 性质：设分别是正态总体的样本均值向量和样本协方差矩阵，则

1. 多元正态分布的参数估计

* 简单随机样本矩阵的统计量
  + 样本均值向量： 
  + 样本离差阵：
  + 样本协方差阵：
  + 样本相关阵：
* 样本均值向量与样本离差阵抽样分布的性质

(1) ；(2)； (3)相互独立

* 统计量的性质
  + 无偏性：
  + 充分性，相合性，完备性，有效性

1. 主成分分析

* （总体）主成分分析的基本模型

设，作线性变换得到:

其中

* 目标函数（信息最大化的准则）

对于第个主成分，限制条件下：



令最大化。（此时）

* 主成分的计算和性质

令是维随机向量的协方差矩阵，且有特征值和单位特征向量序列，其中特征值从大到小排列。令的第个主成分为，则：



* 贡献率和累计贡献率
  + 贡献率：第个主成分方差在总方差中的占比: 
  + 累计贡献率：前个主成分方差在总方差中所占的比例。
* 因子载荷量与相关系数的平方和
  + 因子载荷量：即主成分与原变量的相关系数



* + 相关系数的平方和：将因子载荷量关于求平方和，则有：

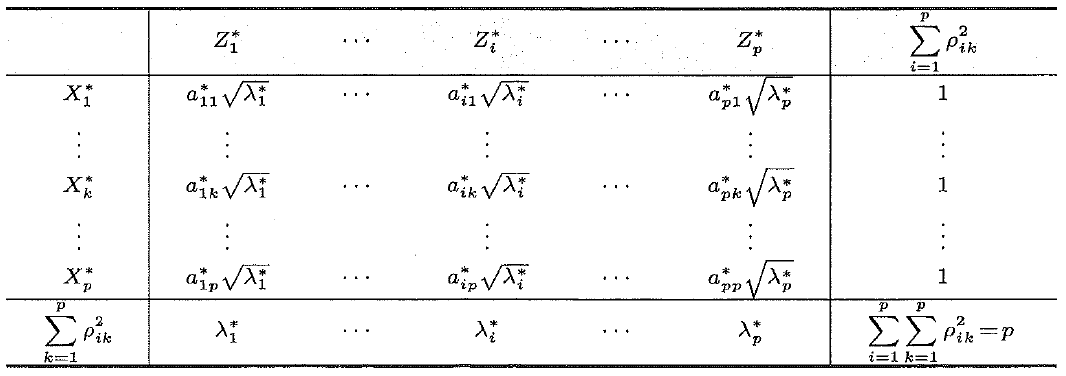


若只对前个主成分求平方和，则定义为前个主成分对原变量的贡献率：

* 基于标准化的（总体）主成分分析

对总体进行标准化：后，再进行主成分分析，等价于直接对相关系数矩阵进行谱分解得到后，基于的主成分为：，其中为标准差对角阵。

此时有：



* 样本主成分分析

将用替代，则分别替换成，性质与总体主成分相同. 其中称为在第个主成分上的得分，称为载荷。

* 基于标准化的（样本）主成分分析

将用替代，则分别替换成，性质与标准化的总体主成分相同。

* 主成分分析在图像处理中的应用
  + 奇异值分解（SVD）：

对数据矩阵作分解：，其中为（或）的个共同的非零特征值的平方根的对角矩阵。进一步地，的列分别包含了的特征向量。

* + 基于SVD的主成分分析：

记为对数据矩阵中心化后的数据矩阵，则的样本协方差矩阵为；再对作SVD分解，则为主成分得分。此时对进行近似：

* + 人脸识别与特征脸：

记维图像数据矩阵为，令，其中。令为的第个特征值，对应的特征向量为，那么对应的特征向量为

选择个主成分，记为，称为个特征脸，则对任意一个图像，其主成分得分为：。因此：

1. 因子分析

* 因子分析模型及假设

设为维随机向量，为其均值向量，若：



其中因子载荷矩阵是一个的常数矩阵，公共因子向量为维随机向量，特殊因子向量为维随机向量。且要求满足：



此时为正交因子分析模型，且

* 因子分析的统计意义
  + 载荷矩阵的统计意义：

可以证明，即，说明每个载荷反映了第个变量与第个公共因子的相关重要性。绝对值越大，相关的密切程度越高。

* + 变量共同度的统计意义：

记，称为共同度，反映了对公共因子的依赖程度。注意到，因此如果靠近，则非常小，表明因子分析的效果好。

* + 公共因子方差贡献的统计意义：

记为公共因子对的各个分量的总方差贡献，为衡量公共因子的贡献率。

* 因子分析模型的性质
  + 因子载荷矩阵不唯一：

令为任意正交矩阵，令，则新的因子载荷矩阵与公共因子向量仍然满足模型假设。

* + 因子分析具有尺度不变性：

令，令，则新的因子载荷矩阵与公共因子向量仍然满足模型假设。

* 因子载荷矩阵的估计方法
  + 主成分法：

1. 求出的协方差矩阵或无偏估计，并作谱分解：



1. 取前个较大的特征值及其对应的特征向量:



1. 因子载荷矩阵与特殊因子的方差估计为：



* + 主因子法：

1. 求出的相关系数矩阵，并假定已知特殊因子的方差，并对约相关矩阵作近似谱分解：



1. 因子载荷矩阵与特殊因子的方差估计为：



1. 给定初始值，替换的对角线元素得到，以此开始迭代更新

注：(1)要求最终求出的;

(2)初始时，主因子法与主成分法等价

* + 极大似然法（要求假定服从多元正态分布）：

1. 给定的初始值（参考主因子法）
2. 对作谱分解，求其前个特征值以及特征向量，并令



1. 令，重复迭代。

* 因子旋转（正交旋转方差最大法）

记为未旋转的的载荷矩阵，为的正交矩阵，旋转后的因子载荷矩阵为，方差最大准则要求最大化，其中：



另外还有其他方法，如四次方最大法和等量最大法。

* 因子得分
  + Thomson因子得分：

假设服从元正态分布，的先验分布为，考虑对的联合分布取条件分布的均值，就得到



其中两种形式等价，且可用估计量代替。

* + Bartlett因子得分：

假设服从元正态分布，则Bartlett因子得分为：



性质：(1)Bartlett因子得分具有无偏性；(2)当因子载荷采用主成分法估计时，Bartlett因子得分就是的前个样本主成分的标准化；(3)在已知的情况下，Thomson因子得分的预测均方误差小于Bartlett因子得分的预测均方误差。

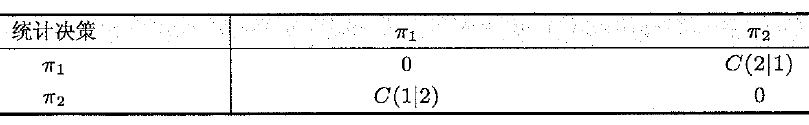
* 因子分析与主成分分析的关系
  + 主成分分析对数据来源的总体分布的协方差结构不做任何假设；而因子分析假设数据来源于因子分析模型
  + 假设特殊因子的方差为零时，二者等价。
  + 主成分分析侧重“变异量”，而因子分析更重视相关变量的“共变异量”。

1. 判别分析

* 判别准则简介

给定以一个个体的测量值向量，判断该个体来自总体或者总体，对于该问题，主要是寻找判别方法，把空间分成两个区域和，记为. 如果这个个体的观测值向量，则把他判断为来自总体，否则为来自总体.

* 判别分析的错判损失



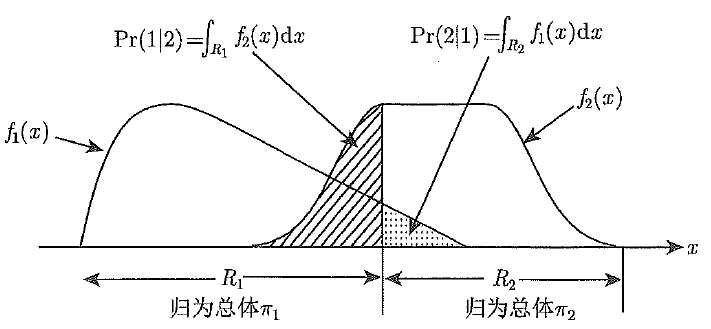
* 判别概率

假设总体的密度函数为，总体的密度函数为，则:

个体来自总体,被正确判别为的概率为:个体来自总体,被错判为的概率为:

个体来自总体,被正确判别为的概率为:

个体来自总体,被错判为的概率为:



* Bayes判别方法

令是来自总体的观测值的先验概率，是来自总体的观测值的先验概率，满足.Bayes判别方法就是找到一种的划分，使得错判的平均损失(ECM)最小化：



* 两个总体的判别准则（结合错判损失）



* 两个已知多元正态分布的判别
  + 协方差不相等的情况：

令，其中，，则最好的判别区域为：

* + 协方差相等的情况：

若，值同上，则最好的判别区域为：



此时退化为关于的一次函数。

* 参数未知时两个正态总体的判别

假设对两个总体存在两组历史样本，，，则可以用这两组样本估计并在判别区域中使用这些估计量替换未知参数，判别准则同上。注意当时，

* 一组样本的整体判别

假设有一个来自总体或新的样本，下面将样本当作整体进行判别。首先计算估计量,其中为新样本的样本离差矩阵。定义Fisher线性判别函数为：

1. 聚类分析

* 聚类分析的分类
  + R-型聚类:对变量或指标进行分类；Q-型聚类:对样本进行分类
* 数据标准化
  + 中心化变换： ;标准化变换：
  + 极差标准化变换：
  + 极差正规化变换:
  + 对数变换:
* 样本之间的距离
  + 性质：非负性；对称性；三角不等式
  + Minkowski距离：

其中依次为绝对值、欧式、切比雪夫距离

不足：(1)距离与各变量的量纲有关；(2)没有考虑指标间的相关性；(3)没有考虑各变量方差的不同

* + 马氏距离：

优点：(1)排除了各指标之间相关性的干扰；(2)不受各个指标量纲的影响；(3)将原始数据作线性变换后，距离不变

* + Canberra（兰氏）距离、斜交空间距离
* 变量之间的相似系数
  + 性质：；；当且仅当共线
  + 夹角余弦：
  + Pearson相关系数
  + Kendall相关系数、Spearman相关系数
* K均值聚类
  + 基本性质：(1)每个观测样本属于个类中至少一个类；(2)没有一个观测样本同时属于两个类或更多的类中。
  + 目标函数：
  + 算法步骤：

1. 为每个观测样本随机分配一个从到的数字，可以看作是这些观测样本的初始类；
2. 重复以下操作，直到类的分配停止位置：
3. 分别计算个类的类中心，第个类的中心是第个类中样本的维观测向量的均值向量；
4. 将每个观测样本分配到距离其最近的类中心所在的类中。

上述算法保证了下式的单调递减：



* 值的确定：绘制关于类内总平方和的碎石图

类内总平方和：

* 缺陷：(1)对椭球或椭圆形状的凸形区域的数据更适用，对其他非凸的数据集则可能失效；(2)对异常值敏感
* 系统聚类的算法步骤：

1. 首先将每个样本看作一类，计算个观测样本中所有两两样本间的距离（共）个。
2. 令：
3. 在第个类中，比较任意两类间的距离，找到距离最小的那一对类，并将他们合并起来；
4. 计算剩下的个新类中，每两个类之间的距离，同样把距离最小的两个类合并。