简答题合集

一、主成分简答题

1. 主成分分析的基本思想及其作用？

主成分分析就是把原有的多个指标转化成少数几个代表性较好的综合指标，这少数几个指标能够反映原来指标大部分的信息（85%以上），并且各个指标之间保持独立，避免出现重叠信息。主成分分析主要起着降维和简化数据结构的作用。

2. 请阐述主成分分析法的基本步骤。

1. 求样本均值$\bar{X}=(\bar{x}_1,\bar{x}_2)$和样本协方差矩阵 S；
2. 求 S 的特征根

求解特征方程$|S-\lambda I|=0$，其中 I 是单位矩阵，解得 2 个特征根$\lambda_1,\lambda_2(\lambda_1\geq\lambda_2)$

3. 求特征根所对应的单位特征向量
4. 写出主成分的表达式

3. 用主成分分析法进行综合评价时，如何构建综合评价函数？

第一种方法，通过主成分分析得到综合指标

$$F_1=a_{11}x_1+a_{21}x_2+\ldots+a_{p1}x_p$$

利用 F1作为评估指标，根据F1得分对样本点进行排序比较。但有3个前提条件：

1. F1与全体原变量都正相关，即$a_{i1}=a_{i1}\sqrt{\lambda_1}>0$ (i=1,2,…,p）。
2. x_i (i=1,2,…,p）在数值上的分布较为均匀。
3. F1的方差贡献率较大。

4. 简述主成分分析的应用。

主成分分析主要用于对数据的降维主成分分析用于系统评估，主成分回归能够解决回归分析中的多重共线性问题

5. 简述提取样本主成分的原则。

方差最大化原则： 样本主成分分析的目标是通过线性变换将原始数据映射到新的坐标系中，使得数据在新坐标系下的方差最大。通过最大化方差，可以保留数据中最重要的信息，降低信息丢失。

正交性原则： 样本主成分之间应该是正交的，即它们之间应该是相互独立的。这是为了确保每个主成分提供的信息是不重复的，不会出现冗余。

排序原则： 样本主成分按照它们对方差的贡献大小进行排序，第一个主成分包含最大的方差，第二个主成分包含次大的方差，依此类推。这样可以根据需求选择保留的主成分数量。

解释总方差的原则： 在降维的过程中，要确保选择的主成分能够解释大部分原

始数据的方差。通常，通过计算累积方差贡献率，来确定选择多少个主成分可以满足需求。

标准化数据原则： 在进行主成分分析之前，通常需要对数据进行标准化，确保不同特征的尺度一致。这是因为主成分分析是基于协方差矩阵计算的，尺度不一致可能导致主成分分析结果不准确。

6. 简述主成分分析的适用范围。

数据降维： 主成分分析可以用于降低数据维度，保留数据中最重要的信息。这在处理高维数据时特别有用，可以减少计算复杂性、消除冗余信息、提高模型训练效率。

特征提取： PCA 可以用于提取数据的主要特征，将数据投影到主成分上，从而减少特征的数量。这对于去除噪声、减少过拟合、改善模型性能很有帮助。

数据可视化： 将数据映射到主成分上可以方便地将高维数据可视化到二维或三维空间中。这有助于更好地理解数据的结构、发现数据之间的关系，并进行更直观的分析。

去除共线性： 在多重共线性存在的情况下，主成分分析可以用来减少特征之间的相关性，提高模型的稳定性和可解释性。

模型简化： 在某些情况下，原始数据中可能包含一些冗余信息，主成分分析可以帮助简化模型，提高模型的泛化能力。

图像处理： 在图像处理中，主成分分析可以用于图像压缩、特征提取和去噪等方面，有助于减小图像数据的规模同时保留重要信息。

7. 简述量纲对主成分分析的影响及消除方法。

从标准化的数据提取的主成分，实际上只包含了各指标间相互影响这一部分信息，不能准确反映原始数据所包含的全部信息

改进原始数据的无量纲化方法

◆ **均值化方法**

$$x_{ji} \to x'_{ji} = \frac{x_{ji}}{\bar{x}_i}$$

均值化后，数据的协方差矩阵S 中的元素

$$u_{jk} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_{ij}}{\bar{x}_j}-1\right)\left(\frac{x_{ik}}{\bar{x}_k}-1\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\frac{(x_{ij}-\bar{x}_j)(x_{ik}-\bar{x}_k)}{\bar{x}_j\bar{x}_k} = \frac{s_{jk}}{\bar{x}_j\bar{x}_k}$$

均值化后，数据的协方差矩阵

$$S=\begin{pmatrix} \frac{s_{11}}{\bar{x}_1^2} & \frac{s_{12}}{\bar{x}_1\bar{x}_2} & \cdots & \frac{s_{1p}}{\bar{x}_1\bar{x}_p} \\ \frac{s_{21}}{\bar{x}_2\bar{x}_1} & \frac{s_{22}}{\bar{x}_2^2} & \cdots & \frac{s_{2p}}{\bar{x}_2\bar{x}_p} \\ \vdots & \vdots & & \vdots \\ \frac{s_{p1}}{\bar{x}_p\bar{x}_1} & \frac{s_{p2}}{\bar{x}_p\bar{x}_2} & \cdots & \frac{s_{pp}}{\bar{x}_p^2} \end{pmatrix}$$

对角线上是原变量标准差系数的平方，其他位置上是变量两两之间的相互关系。

均值化处理后的协方差矩阵不仅消除了指标量纲与数量级的影响，还能包含原始数据的全部信息。

二、因子分析简答题

1. 简述因子分析的基本思想。

因子分析是根据相关矩阵内部的依赖关系，把一些具有错综复杂关系的变量综合为数量较少的几个因子。通过不同因子来分析决定某些变量的本质及其分类的一种统计方法。简单地说，就是根据相关性大小把变量分组，使得同组内的变量之间相关性较高，不同组的变量相关性较低。每组变量代表一个基本结构，这个基本结构称为因子。

2. 简述因子载荷矩阵的含义、统计特征及其意义。

1、因子载荷 a_{ij} 的统计意义

因子载荷 a_{ij} 是第i个变量与第j个公共因子的相关系数

模型为 $X_i=a_{i1}F_1+\cdots+a_{im}F_m+\varepsilon_i$

$$\mathrm{cov}(X_i,F_j)=\mathrm{cov}(\sum_{k=1}^{m}a_{ik}F_k+\varepsilon_i,F_j)=\mathrm{cov}(\sum_{k=1}^{m}a_{ik}F_k,F_j)+\mathrm{cov}(\varepsilon_i,F_j)$$
$$=a_{ij}$$

根据公共因子的模型性质，有

$\gamma_{x_iF_j}=a_{ij}$（载荷矩阵中第i行，第j列的元素）反映了第i个变量与第j个公共因子的相关性。绝对值越大，相关的密切程度越高。

■ 因子载荷不是惟一的

设T为一个p×p 的正交矩阵，令

$A^*=AT$，$F^*=T'F$，则模型可以表示为

$$X=AF+\varepsilon=(AT)(T'F)+\varepsilon=A^*F^*+\varepsilon$$

且满足因子模型的条件

$$E(F^*)=E(T'F)=T'E(F)=0$$
$$Var(F^*)=Var(T'F)=T'Var(F)T=I$$
$$\mathrm{cov}(\mathbf{F}^*,\boldsymbol{\varepsilon})=E(\mathbf{F}^*\boldsymbol{\varepsilon}')=\mathbf{0}$$
$$E(\varepsilon)=0,Var(\varepsilon)=diag(\sigma_1^2,\sigma_2^2,\cdots,\sigma_p^2)$$

3. 比较因子分析和主成分分析，说明它们的相似和不同之处。

主成分分析分析与因子分析也有不同，主成分分析仅仅是变量变换，而因子

分析需要构造因子模型。

主成分分析：原始变量的线性组合表示新的综合变量，即主成分。

因子分析：潜在的假想变量和随机影响变量的线性组合表示原始变量。

4. 因子模型和回归模型相比较有何异同？

因子分析与回归分析不同，因子分析中的因子是一个比较抽象的概念，而回归因子有非常明确的实际意义。

5. 因子分析中对因子载荷矩阵进行旋转的目的是什么？常用的旋转方法有哪些？

因子分析的数学目的不仅仅要找出公共因子以及对变量进行分组，更重要的要知道每个公共因子的含义，以便进行进一步的分析。如果每个公共因子的含义不清，则不便于进行实际背景的解释。由于因子载荷阵是不惟一的，所以应该对因子载荷阵进行旋转。目的是使因子载荷阵的结构简化，使载荷矩阵每列或行的元素平方值向0和1两极分化。主要的正交旋转法有方差最大法和四次方最大法。

正交旋转：由因子载荷矩阵 A 右乘一正交阵而得到，经过旋转后的新的公因子仍然保持彼此独立的性质。

1、方差最大法

2、四次方最大旋转

斜交旋转：放弃了因子之间彼此独立这个限制，可达到更简洁的形式，实际意义也更容易解释。

不论是正交旋转还是斜交旋转，都应该在因子旋转后，使每个因子上的载荷尽可能拉开距离，一部分趋近 1，一部分趋近 0，使各个因子的实际意义能更清楚地表现出来。

6. 阐述运用因子分析进行综合评价时，综合评价函数的构造方法。

用因子分析方法进行综合评价

通过因子分析，取m个公共因子 $F_1,F_2,\cdots,F_m$ ，

以每个公共因子 F_j 的方差贡献率 $\alpha_j=\frac{g_j^2}{p}$

为权，构造综合评价函数

$$F=\alpha_1F_1+\alpha_2F_2+\cdots\alpha_mF_m$$

按F值的大小对样品进行排序比较或分类。

7. 阐述主成分分析和因子分析用于对变量降维时，两种方法在基本思想和做法上的差异。

从二者表达的含义上看，主成分分析法和因子分析法都寻求少数的几个变量（或因子）来综合反映全部变量（或因子）的大部分信息，变量虽然较原始变量少，但所包含的信息量却占原始信息量的 85%以上，用这些新变量来分析问题，其可信程度仍然很高，而且这些新的变量彼此间互不相关，消除了多重共线性。这两种分析法得出的新变量，并不是原始变量筛选后剩余的变量。在主成分分析中，最终确定的新变量是原始变量的线性组合，如原始变量为 x1，x2，……，x3，经过坐标变换，将原有的 p 个相关变量 xi 作线性变换，每个主成分都是由原有 p 个变量线性组合得到。在诸多主成分 Zi 中，Z1 在方差中占的比重最大，说明它综合原有变量的能力最强，越往后主成分在方差中的比重也小，综合原信息的能力越弱。

因子分析是要利用少数几个公共因子去解释较多个要观测变量中存在的复杂关系，它不是对原始变量的重新组合，而是对原始变量进行分解，分解为公共因子与特殊因子两部分。公共因子是由所有变量共同具有的少数几个因子；特殊因子是每个原始变量独自具有的因子。

三、聚类分析简答题

1. 如何测度样品和变量间的相似性？计算样品之间的距离有哪些公式？它们各有什么特点？

（1）距离：测度样品之间的亲疏程度。将每一个样品看作 p 维空间的一个点，并用某种度量测量点与点之间的距离，距离较近的归为一类，距离较远的点应属于不同的类。相似系数：测度变量之间的亲疏程度

（2）明氏距离、标准化的欧氏距离、马氏距离、兰氏距离、

（3）①明氏距离的数值与指标的量纲有关

②没有考虑各个变量之间相关性的影响

马氏距离又称为广义欧氏距离。

马氏距离考虑了观测变量之间的相关性。如果假定各变量之间相互独立，即观测变量的协方差矩阵是对角矩阵，此时马氏距离就是标准化的欧氏距离。

马氏距离不受指标量纲及指标间相关性的影响

兰氏距离不受量纲影响

对大的奇异值不敏感，特别适合于高度偏倚的数据

没有考虑指标之间的相关性

2. Q 型聚类法和 R 型聚类法有什么异同？

R 型分析----对变量进行分类

Q 型分析----对样品进行分类

3. 简述系统聚类法的基本思想及主要步骤。

基本思想：

先将 n 个样品各自看成一类，然后规定样品之间的“距离”和类与类之间的距离。选择距离最近的两类合并成一个新类，计算新类和其它类（各当前类）的距离，再将距离最近的两类合并。这样，每次合并减少一类，直至所有的样品都归成一类为止。

主要步骤：1.计算 n 个样品两两间的距离d_{ij}，记作 D=$\{d_{ij}\}$。

2.构造 n 个类，每个类只包含一个样品。

3.合并距离最近的两类为一新类。

4.计算新类与各当前类的距离。

5.重复步骤 3、4，合并距离最近的两类为新类，直到所有的类并为一类为止。

6.画聚类谱系图。

7.决定类的个数和类。

4. 简述系统聚类分析的优缺点。

简单，直观；

当样本点数量十分庞大时，则是一件非常繁重的工作，且聚类的计算速度也比较慢。

5. Q 型系统聚类法包括哪几种方法，有什么特点。

最短距离法 (Single linkage)

最长距离法 (Complete method)

中间距离法（Median method）
重心法（Centroid method）
类平均法 （Average linkage）
离差平方和法（Ward method
6. 简述动态聚类法的基本思想与步骤。

基本思想：

选取若干个样品作为凝聚点，计算每个样品和凝聚点的距离，进行初始分类，然后根据初始分类计算其重心，再进行第二次分类，一直到所有样品不再调整为止。

步骤：

第一，选择凝聚点；

第二，初始分类；

对于取定的凝聚点，视每个凝聚点为一类，将每个样品根据定义的距离向最近的凝聚点归类。

第三，修改分类

得到初始分类，计算各类的重心，以这些重心作为新的凝聚点，重新进行分类，重复步骤 2，3，直到分类的结果与上一步的分类结果相同为止，表明分类已经合理。

四、判别分析简答题

1. 简述判别分析的基本思想。

根据已知类别的样本所提供的信息，总结出分类的规律性，建立判别公式和判别准则，判别新的样本点所属类型，是判别个体所属群体的一种统计方法。

2. 简述判别分析与聚类分析的区别与联系。

区别：

判别分析：已知研究对象分为若干个类别，并且已 经取得每一类别的一批观测数据，在此基础上寻求出分类的规律性，建立判别准则，然后对未知类别的样品进行判别分类。

聚类分析：一批样品划分为几类事先并不知道，正需要通过聚类分析来给以确定类型。

联系：

聚类分析：用不同的聚类方法可能得到不同的结果，保留共性的聚类结果；对于用不同方法归类不同的少数样品，再结合判别分析加以判断归类。

判别分析：将共性的聚类结果，作为已知类别的样本的信息（训练样本），对未知类别的样品（测试样本）进行判别分类。

3. 简述距离判别法的基本思想，并写出两总体协方差矩阵相等时的判别函数和判别规则。

距离判别的最直观的想法：计算样品 x 到第 i 个类的距离$d^2(x,G_i)$，哪个距离最小，就将它判归哪个总体。所以，首先考虑，是否能够构造一个恰当的距离函数，通过计算样本点与某类别之间距离的大小，判别其所属类别。

$$W(x) = (x - \frac{\mu_1 + \mu_2}{2})'\Sigma^{-1}(\mu_1 - \mu_2) = (x - \bar{\mu})' \alpha$$

$$= \alpha'(x - \bar{\mu})$$

$$= a_1(x_1 - \bar{\mu}_1) + \cdots + a_p(x_p - \bar{\mu}_p)$$

则前面的判别法则表示为 $f_1(x) - f_2(x)$

$$\begin{cases} x \in G_1, & \text{如} W(x) > 0, \\ x \in G_2, & \text{如} W(x) < 0。 \\ \text{待判}, & \text{如} W(x) = 0 \end{cases}$$

4. 距离判别法中的线性判别函数在什么情况下适用？

两总体协方差矩阵相等时

距离判别只要求知道总体的数字特征，不涉及总体的分布函数，当总体均值和协差阵未知时，用样本的均值和协方差矩阵来估计。

距离判别方法简单实用，但没有考虑到每个总体出现的机会大小，即先验概率，没有考虑错判的损失。

5. 简述 Fisher 判别的基本思想、判别步骤及判别规则

费歇判别的基本思想是投影，将 k 组 p 维数据投影到某一个方向，使其投影的组与组之间尽可能地分开。Fisher 判别法由 Fisher 在 1936 年提出，是根据方差分析的思想建立起来的一种能较好区分各个总体的线性判别法，该判别方法对总体的分布不做任何要求。

6. 简述距离判别和 Fisher 判别的异同。

距离判别和 Fisher 判别都是模式识别和统计学中用于分类的方法

异同点总结：

目标不同：

距离判别的目标是通过测量距离直接判断样本间的相似性。

Fisher 判别的目标是通过找到一个投影方向，最大化类别间差异，最小化类别内差异。

特征权重考虑：

距离判别不考虑特征之间的权重，仅仅通过距离度量来进行分类。

Fisher 判别考虑了特征之间的权重，通过最优化投影方向来增强分类性能。

应用范围：

距离判别适用于简单的、线性可分的问题。

Fisher 判别更适用于多特征、非线性可分的问题，对于提高类别间的可分性更有优势。

对异常值敏感：

距离判别对异常值比较敏感，因为它直接依赖于样本之间的距离。

Fisher 判别在一定程度上对异常值有一些鲁棒性，因为它更关注类别间的整体

分布。

五、对应分析简答题

1. 对应分析的基本思想是什么？

对应分析，也称关联分析、R-Q 型因子分析，是由法国人 Beozecri 于 1970 年提出来的。通过分析由定性变量构成的交互汇总表来揭示变量之间的联系。可以揭示同一变量的各个类别之间的差异，以及不同变量各个类别之间的对应关系。它是一种视觉化的数据分析方法，它能够将几组看不出任何联系的数据，通过视觉上可以接受的定位图展现出来。

2. 简述对应分析与因子分析的联系。

在因子分析中人们通常只是分析原始变量的因子结构，找出决定原始变量的公共因子，从而使问题的分析简化和清晰。这种研究对象是变量的因子分析称为 R 型因子分析。对有些问题来说，我们还需要研究样品的结构，若对于样品进行因子分析，称为 Q 型因子分析。当我们对数据同时进行 R 和 Q 型因子分析，并分别保留两个公共因子，则是对应分析的初步。对应分析综合了 R 型和 Q 型分析的优点，将两者统一起来；更重要的是可以把变量和样品的载荷反映在相同的公因子轴上。这就把变量和样品联系起来，便于解释和推断。对应分析的基本思想是将一个列联表的行和列中各元素的比例结构以点的形式在低维空间表示出来。它最大特点是能把众多的样品和众多的变量同时作到一张图上，直观展示。

3. 简述对应分析的基本步骤。

1、获取对应分析数据

首先要规定研究的目的，然后选择对应分析中所需数据，应该包括的背景资料。

2、建立列联表

3、对应分析

4、对应图并解释结果的意义。

4. 对应分析有什么特点？

优点

（1）定性变量划分的类别越多，这种方法的优越性越明显

（2）揭示行变量类间与列变量类间的联系

（3）直观地表变量所含类别间的关系

缺点：

（1）不能用于相关关系的假设检验

（2）受极端值的影响

5. 对应分析图可以做哪些方面的分析？

对应分析是一种数据分析技术，它能够帮助我们研究由定性变量构成的交互汇总表来揭示变量间的联系。

交互表的信息以图形的方式展示。

主要适用于有多个类别的定类变量 Category Data，可以揭示同一个变量的各个类别之间的差异，以及不同变量各个类别之间的对应关系。适用于两个或多个定类变量。也可以，揭示样品和变量间的内在联系

6. 对应分析具有哪些实际应用？

概念发展 （Concept Development）

新产品开发 (New Product Development)

市场细分 (Market Segmentation)

竞争分析（Competitive Analysis）
广告研究（Advertisement Research）

六、典型相关简答题

1. 阐述典型相关分析的基本思想及其应用。

典型相关分析（Canonical Correlation）是研究两组变量之间相关关系的一种多元统计方法。它能够揭示两组变量之间的内在联系。典型相关分析的目的是识别并量化两组变量之间的联系，将两组变量相关关系的分析，转化为一组变量的线性组合与另一组变量线性组合之间的相关关系分析。目前，典型相关分析已被应用于心理学、市场营销等领域。如用于研究个人性格与职业兴趣的关系，市场促销活动与消费者响应之间的关系等问题的分析研究。

2. 简述典型相关分析与相关分析有何异同点？

3. 什么是典型变量？它具有哪些性质？

在典型相关分析中，在一定条件下选取系列线性组合以反映两组变量之间的线性关系，这被选出的线性组合配对被称为典型变量；

三、典型变量的性质

1、**同一组变量的典型变量之间互不相关**

$$u_k = \mathbf{a}_k'\mathbf{x} \quad v_k = \mathbf{b}_k'\mathbf{y} \qquad k,l = 1,2,\cdots,r; k \neq l$$

因为特征向量之间是正交的。故

X组的典型变量之间是相互独立的：

$$\mathrm{cov}(u_k,u_l) = \mathrm{cov}(a_k'X, a_l'X) = a_k'\,\Sigma_{11}\,a_l = 0$$

Y组的典型变量之间是相互独立的：

$$\mathrm{cov}(v_k,v_l) = \mathrm{cov}(b_k'Y, b_l'Y) = b_k'\,\Sigma_{11}\,b_l = 0$$

2、不同组变量的典型变量之间的相关性

不同组内一对典型变量之间的相关系数为：

$$\mathrm{cov}(u_i,v_j) = \mathrm{cov}(\mathbf{a}_i'\mathbf{x},\mathbf{b}_j'\mathbf{y})$$

$$= \mathbf{a}_i\,\mathrm{cov}(\mathbf{x},\mathbf{y})\mathbf{b}_j' = \mathbf{a}_i'\boldsymbol{\Sigma}_{12}\mathbf{b}_j$$

$$= \begin{cases} \lambda_i, i = j \\ 0, i \neq j \end{cases}$$

同对相关系数为λ_i，不同对则为零。

3、原始变量与典型变量之间的相关系数
（典型载荷分析）

原始变量相关系数矩阵 $R = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}$

X典型变量系数矩阵

$$\mathbf{A} = [\mathbf{a}_1 \quad \mathbf{a}_2 \quad \cdots \quad \mathbf{a}_r]_{p\times r} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1r} \\ a_{21} & a_{22} & \cdots & a_{2r} \\ \vdots & \vdots & & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pr} \end{bmatrix}$$

4. 简述典型相关分析中冗余分析的内容与作用。

8.5 简述典型相关分析中冗余分析的内容及作用。

在进行样本典型相关分析时，我们也想了解每组变量提取出的典型变量所能解释的该组样本总方差的比例，从而定量测度典型变量所包含的原始信息量的大小。因此，典型相关分析中冗余分析的作用就是解释典型变量所包含原始变量信息量的大小。主要内容有：

对于经标准化变换处理的样本数据协差阵就等于相关系数矩阵，因而，第一组变量样本的总方差为 $tr(\mathbf{R}_{11})=p$，第二组变量样本的总方差为 $tr(\mathbf{R}_{22})=q$。

那么如何计算前 r 个典型变量对样本总方差的贡献呢？从前述的典型载荷可知，$\hat{\mathbf{A}}_z^*$ 和 $\hat{\mathbf{B}}_z^*$ 是样本典型相关系数矩阵，典型系数向量是矩阵的行向量，$\hat{\mathbf{U}}=\hat{\mathbf{A}}_z^*\mathbf{Z}^{(1)}$，$\hat{\mathbf{V}}=\hat{\mathbf{B}}_z^*\mathbf{Z}^{(2)}$。

那么：