

第四章(1) 图像数据处理及 分析

Lianghai Xiao

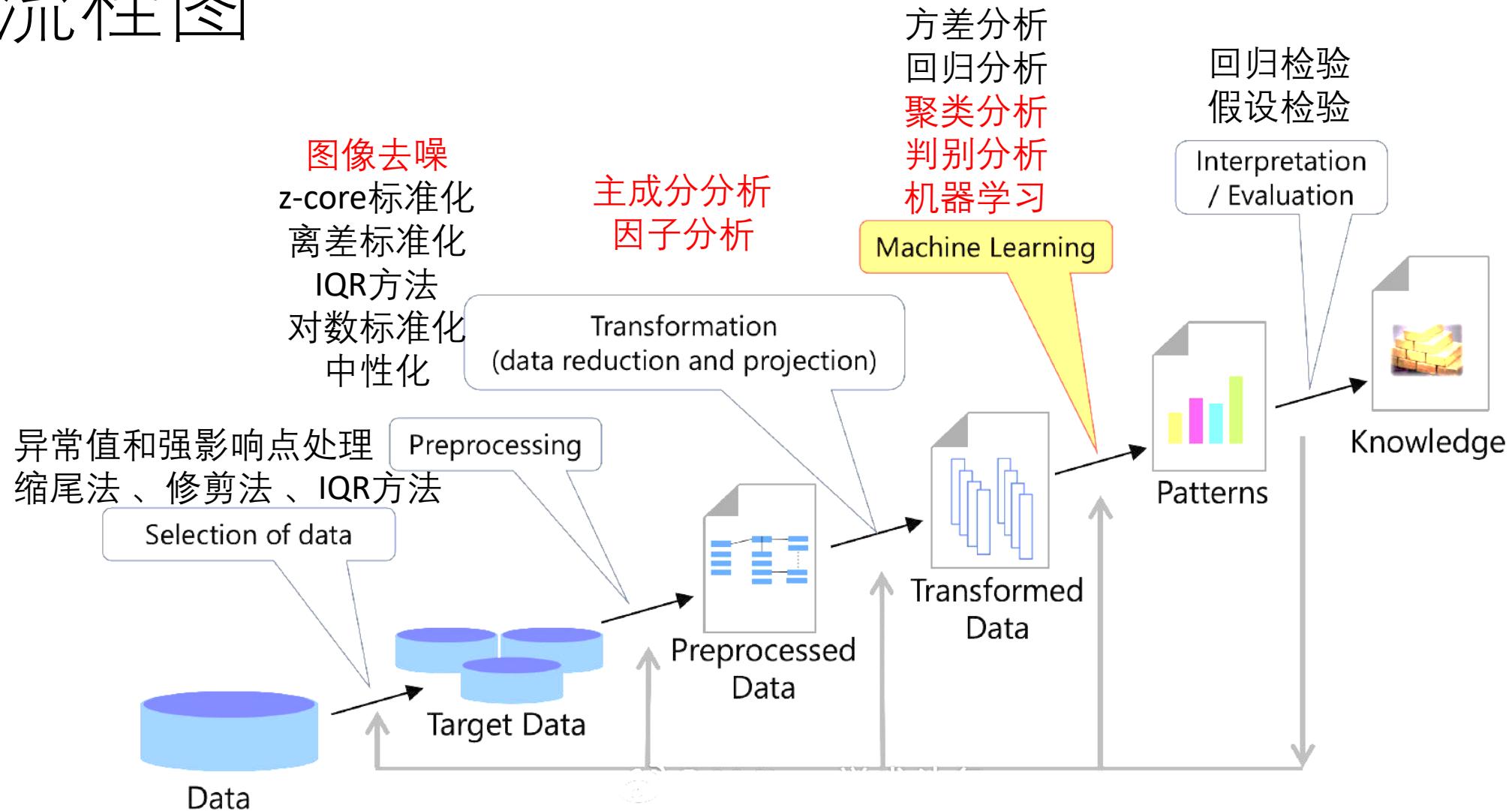
<https://github.com/styluck/mlb>

作业邮箱： alswfx@126.com

目录

- 数据特征的提取与描述
 - 主成分分析
 - 因子分析
- 数据的分类分析方法
 - 聚类分析
 - 判别分析
 - 支持向量机

流程图



图像的预处理

- **去除噪声**

- **均值滤波**: 计算图像中每个像素点邻域内像素的平均值来代替该像素点的原始值。
- 去除图像中的颗粒噪声，但会使图像边缘变得模糊。
- **中值滤波**: 将像素点邻域内的像素值进行排序，然后用中间值（中值）来替换该像素点的原始值。
- 去除椒盐噪声方面效果显著，能够在一定程度上保留图像的边缘和细节。

图像的预处理

- **去除噪声**

- **高斯滤波：**认为像素点的邻域像素对其影响是符合高斯分布的。离中心像素越近的像素点，其权重越大；离得越远，权重越小。在滤波过程中，每个像素点的值是其邻域像素值的加权平均，权重由高斯函数确定。
- 对于去除高斯噪声效果较好，但也会对图像的边缘有一定的模糊作用，不过通常比均值滤波的模糊程度要小。
- **小波去噪：**基于小波变换将图像分解为不同尺度和方向的小波系数，通过设定阈值，将小于阈值的小波系数置零或进行收缩处理，然后再进行小波逆变换得到去噪后的图像。
- 可以根据图像的特点选择合适的小波基函数进行分解，对多种类型的噪声都有较好的去除效果，并且能够在一定程度上保留图像的细节和特征，但阈值的选择比较关键。

图像的预处理

- 归一化

- 离差标准化（归一化）
- 计算公式为：

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

- 将图像的像素值从原始范围线性映射到一个新的范围，比如将像素值从 [0, 255] 映射到 [0, 1]。

- Z-score 标准化

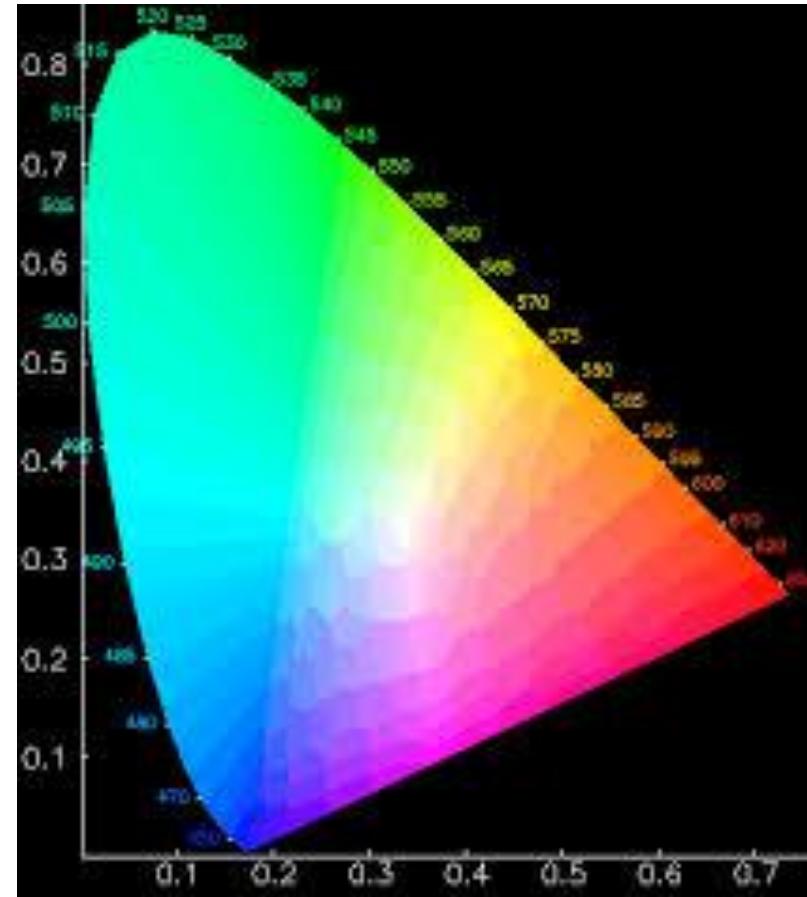
- 计算公式为：

$$Z = \frac{X - \mu}{\sigma}$$

- 使处理后的图像数据均值为 0，标准差为 1。这样可以使不同数据集在同一尺度下进行比较和处理。

图像的预处理

- 归一化
 - RGB 与灰度转换
 - 将彩色图像（RGB 颜色空间）转换为灰度图像。使用NTSC 公式将RGB 值转换为灰度：
$$\text{灰度} = 0.299 \cdot (\text{红色}) + 0.587 \cdot (\text{绿色}) + 0.114 \cdot (\text{蓝色})。$$
 - 此公式可非常准确地反映一般人对红色、绿色和蓝色光源的相对感知。



主成分分析

主成分分析

- 尽管收集和分析大量的多变量数据可以为我们提供宝贵的信息资源，但这同时也加大了数据收集的工作负担。
- 主成分分析（Principal Component Analysis，简称 PCA）的主要目的是在尽可能保留原始数据信息的情况下，对高维数据进行降维处理。
- 在一个有多个变量（如身高、体重、血压等）的数据集中，这些变量之间可能存在相关性。主成分分析可以找到一组新的、相互无关的变量（主成分）来代替原始变量。

主成分分析

- 主成分分析的目的：
 - **降维**：原始数据集可能包含大量特征，但其中许多特征可能是冗余的或高度相关的。PCA可以减少特征的数量，同时保留大部分信息。
 - **降维的好处：**
 - **加速收敛**：减少输入的数据量，可以加速优化算法收敛，加快计算速度；
 - **修正过拟合**：降低样本集对参数的影响，减轻出现过拟合的概率；
 - **噪声过滤**：去除数据中的噪声，因为它只保留最重要的变化模式。
 - **数据恢复**：PCA常用于特征提取，以恢复缺失的数据。
 - **可视化**：通过将多维数据映射到较低维度（通常是二维或三维），PCA可以帮助我们更直观地理解数据。

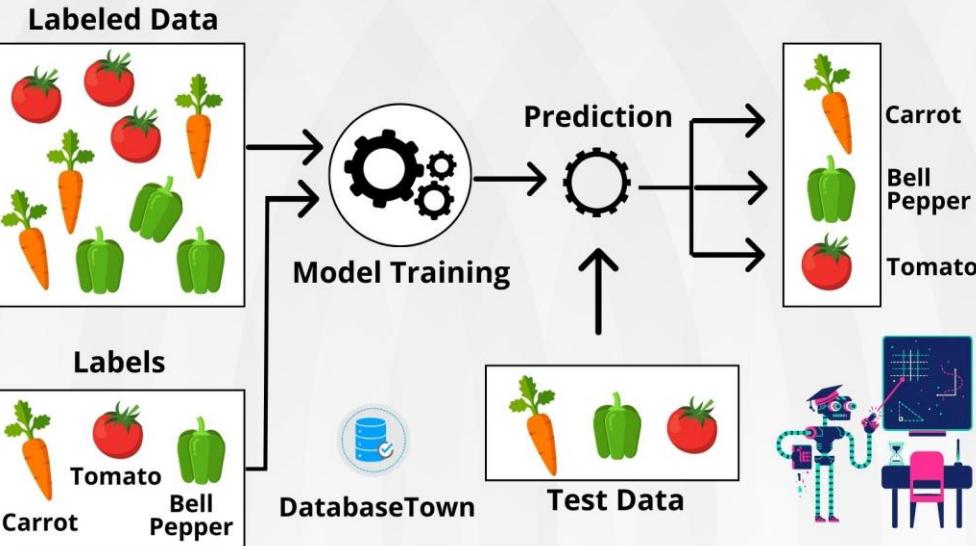
监督学习和无监督学习

- 在**监督学习**中，我们有一个包含输入特征（ x ）和对应的目标输出（ y ）的训练数据集。模型通过学习输入和输出之间的映射关系，以便在给定新的输入时能够预测输出。
- **无监督学习**是在没有给定明确的目标输出的情况下，让模型自动从数据中发现模式和结构。也就是说，数据集中只有输入特征 X ，没有对应的 y 。
- PCA是一种**无监督学习方法**，它不依赖于数据的类别标签，旨在找到数据中方差最大的方向，以这些方向作为新的特征空间，从而实现降维。

监督学习和无监督学习

SUPERVISED LEARNING

Supervised machine learning is a branch of artificial intelligence that focuses on training models to make predictions or decisions based on labeled training data.



UNSUPERVISED LEARNING

Unsupervised learning is a type of machine learning where the algorithm learns from unlabeled data without any predefined outputs or target variables.

Input Raw Data



Unlabeled Data

DatabaseTown

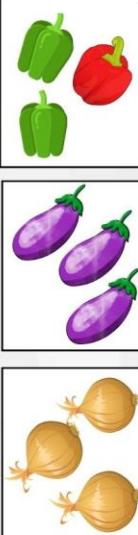
Interpretation



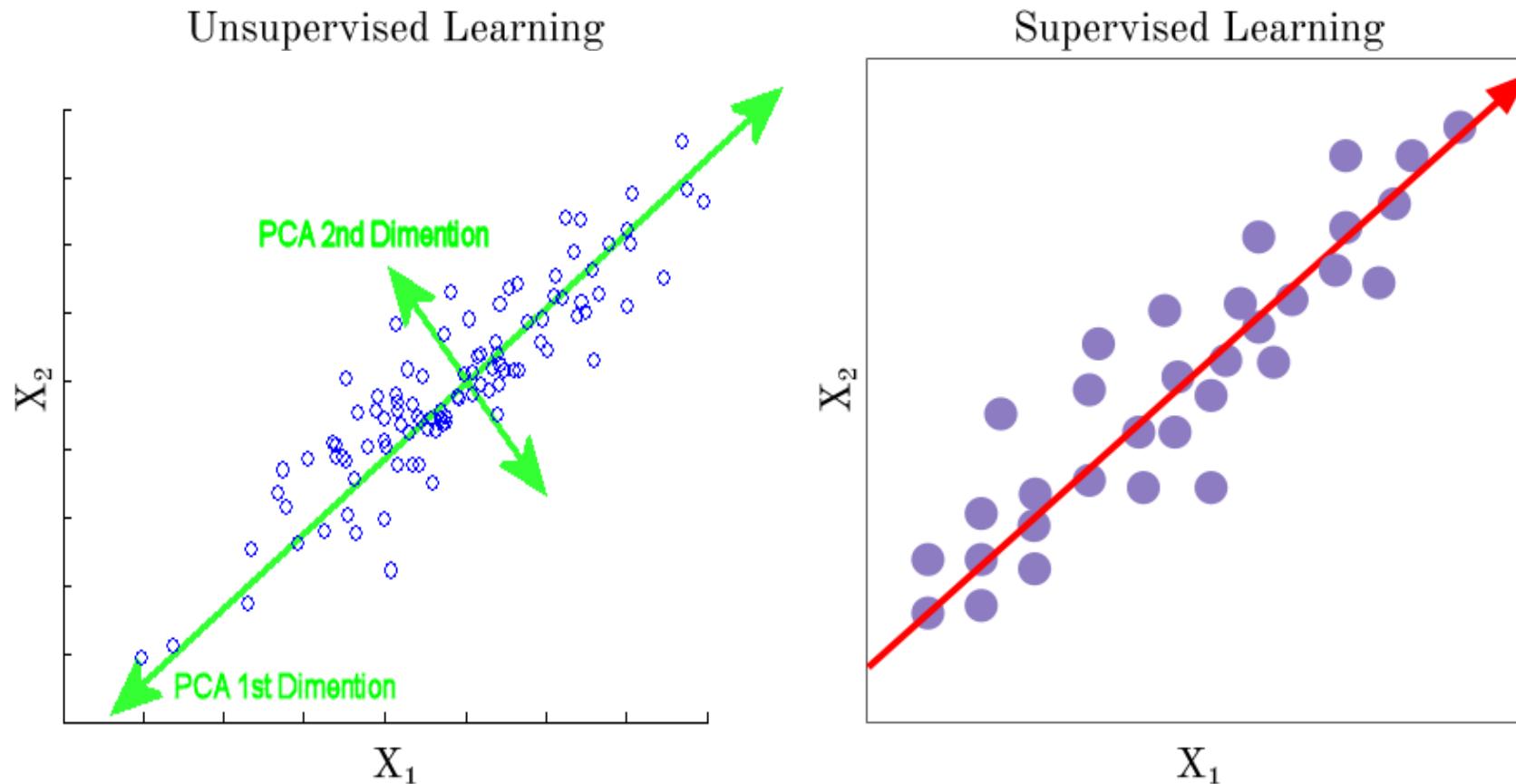
Algorithms



Processing



监督学习和无监督学习



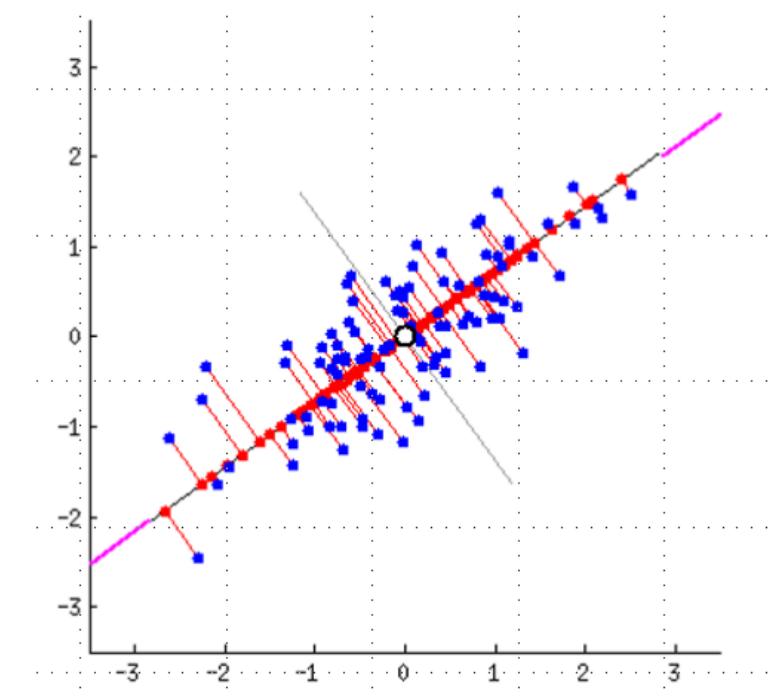
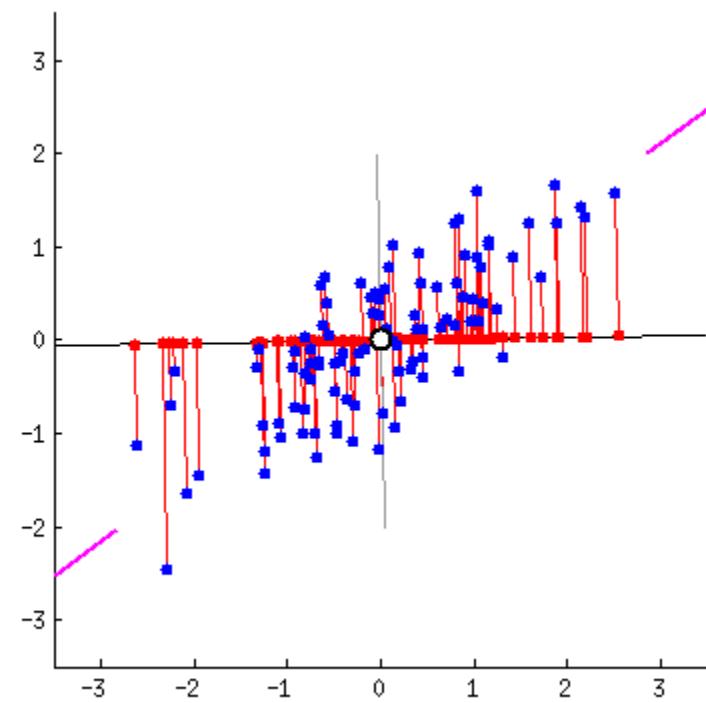
主成分分析

- PCA的数学定义：
- 找到一个正交化线性矩阵，从而得到一个线性映射，把高维的数据映射到一个低维的线性坐标系统中（把原始数据变换到一个新的坐标系统中）
- 使得这一数据的任何投影的第一大差异性在第一个坐标（称为第一主成分）上，第二大差异性在第二个坐标（第二主成分）上，以此类推。

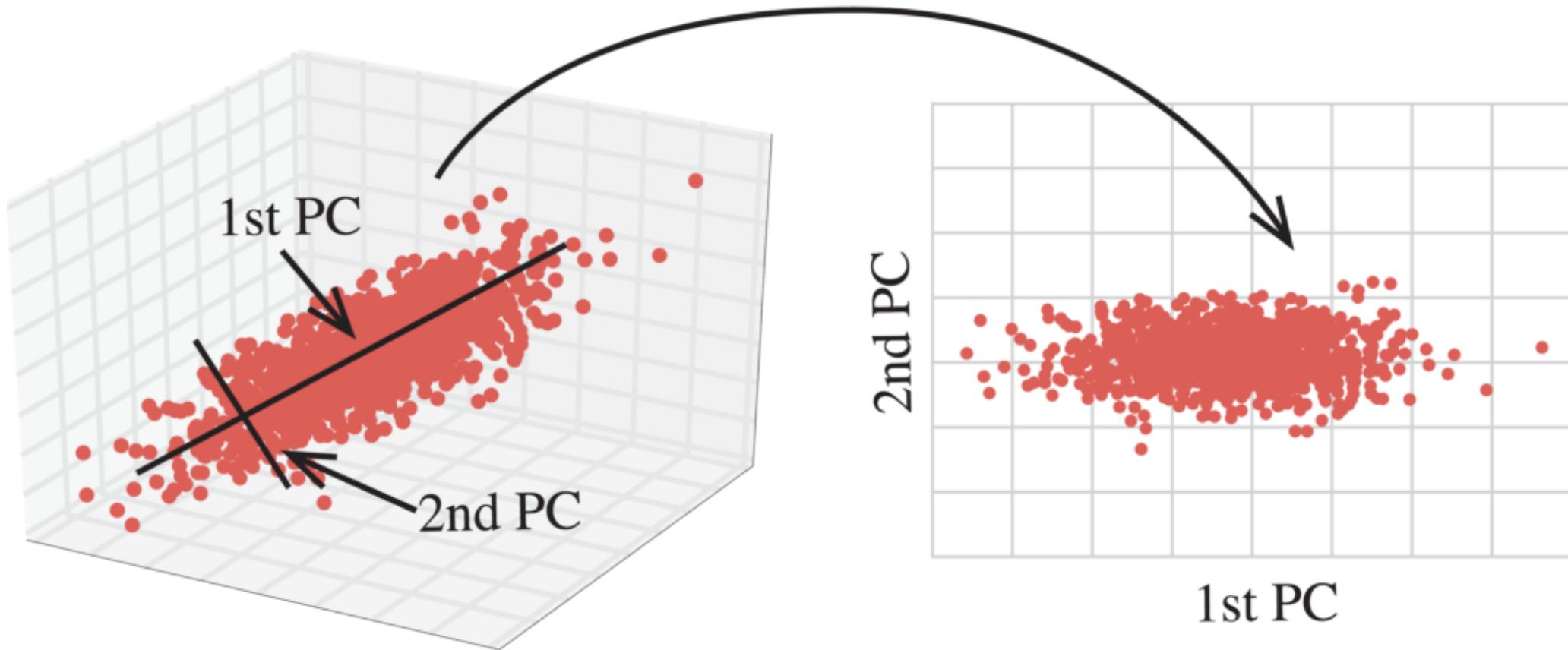
PCA优化目标

- PCA推导有两种主要思路：
 - 1.最大化数据投影后的的方差（让数据更分散）
 - 2.最小化投影造成的损失
- 采用第一种思路完成推导过程，下图中旋转的是新坐标轴，每个数据点在改坐标轴上垂直投影，最佳的坐标轴为数据投影后的数据之间距离最大。

主成分分析的优化目标



主成分分析的优化目标



PCA推导

- 换句话说，PCA的目标是找到一组新的正交基 $\{u_1, u_2, \dots, u_k\}$ (从 m 维下降到 k 维)，使得数据点在该正交基构成的平面上投影后，数据间的距离最大，即数据间的方差最大。
- 如果数据在每个正交基上投影后的方差最大，那么同样满足在正交基所构成的平面上投影距离最大。

PCA推导

- 设正交基 u_j ，数据点 x_i 在该基底上的投影距离为 $x_i u_j$ ，所以所有数据在该基底上投影的方差 J_j 为：

$$J_j = \frac{1}{n} \sum_{i=1}^n (x_i u_j - x_{center} u_j)^2$$

- 由于在数据运算之前对数据 x 进行0均值初始化，上式可写作

$$\begin{aligned} J_j &= \frac{1}{n} \sum_{i=1}^n (x_i u_j)^2 \\ &= \frac{1}{n} \sum_{i=1}^n (u_j^T x_i^T \cdot x_i u_j) = u_j^T \cdot \frac{1}{n} \sum_{i=1}^n (x_i^T x_i) \cdot u_j \end{aligned}$$

PCA推导

- 写成矩阵形式，有

$$J_j = \frac{1}{n} u_j^T X^T X u_j$$

- 则优化问题为

$$\begin{aligned} \max_{u_j} J_j &= \frac{1}{n} u_j^T X^T X u_j \\ s.t. \quad u_j^T u_j &= 1 \end{aligned}$$

PCA推导

- 只提取一个主成分：

$$\begin{aligned}\max_{u_j} J_j &= \frac{1}{n} u_j^T X^T X u_j \\ s.t. \quad u_j^T u_j &= 1\end{aligned}$$

- 提取 k 个主成分：

$$\begin{aligned}\max_{U_k} J_k &= \frac{1}{n} \text{tr}(U_k^T X^T X U_k) \\ s.t. \quad U_k^T U_k &= I_k\end{aligned}$$

主成分分析

- 如何得到包含最大差异性的主成分方向？
- 通过计算数据矩阵的协方差矩阵，然后得到协方差矩阵的特征值特征向量，**选择特征值最大(即方差最大)的k个特征所对应的特征向量组成的矩阵**。这样就可以将数据矩阵转换到新的空间当中，实现数据特征的降维。
- 两种实现方法：基于特征值分解协方差矩阵实现PCA算法、基于SVD分解协方差矩阵实现PCA算法。

特征值分解

- 如果一个向量 v 是矩阵 A 的特征向量，将一定可以表示成下面的形式：

$$Av = \lambda v$$

- 其中， λ 是特征向量 v 对应的特征值，一个矩阵的一组特征向量是一组正交向量。
- 对于矩阵 A ，有一组特征向量 v ，将这组向量进行正交化单位化，就能得到一组正交单位向量。特征值分解就是将矩阵A分解为如下式：

$$A = V\Sigma V^{-1}$$

- 其中， V 是矩阵 A 的特征向量组成的矩阵， Σ 则是一个对角阵，对角线上的元素就是特征值。

SVD分解

- 奇异值分解是一个能适用于任意矩阵的一种分解的方法，对于任意矩阵 X 总是存在一个奇异值分解：

$$\bullet X = W\Sigma U^T$$

- 设 X 是一个 $n \times m$ 的矩阵，那么得到的 U 是一个 $m \times m$ 的方阵， U 里面的正交向量被称为左奇异向量。
- Σ 是一个 $n \times m$ 的矩阵， Σ 除了对角线其它元素都为0，对角线上的元素称为奇异值。
- 一般来讲，算法会将 Σ 上的值按从大到小的顺序排列。

主成分分析

- 输入：数据集 $X = \{x_1, x_2, \dots, x_n\} \in R^{n \times m}$ ，需要从 m 维降到 k 维。
- 1) 去平均值(即去中心化)，即每一位特征减去各自的平均值。
- 2) 计算协方差矩阵 $\frac{1}{n} X^T X$ ，这里除或不除样本数量 n 或 $n - 1$ ，对求出的特征向量没有影响。
- 3) 用特征值分解方法求协方差矩阵 $\frac{1}{n} X^T X$ 的特征值与特征向量。
- 4) 对特征值从大到小排序，选择其中最大的 k 个。然后将其对应的 k 个特征向量分别作为行向量组成特征向量矩阵 U 。
- 5) 将数据转换到 k 个特征向量构建的新空间中，即 $Y = XU$ 。

主成分分析

- 数据集 $X = \{x_1, x_2, \dots, x_n\} \in R^{n \times m}$, 其行为数据样本, 列为数据类别, 并经过了去平均值处理。则 X 的奇异值分解为

$$\bullet X = W\Sigma U^T$$

- 据此,

$$\begin{aligned} Y &= XU \\ &= W\Sigma U^T U \\ &= W\Sigma \end{aligned}$$

- 我们可以利用 U_k 把 X 映射到一个只应用前面 k 个向量的低维空间中去:

$$\bullet Y = XU_k = W\Sigma U^T U_k = W\Sigma_k$$

选择降维后的维度K(主成分的个数)

- x_i :原始数据, y_i :PCA降维后的数据
- average squared projection error:

$$\frac{1}{n} \sum_{i=1}^n \|x_i - y_i\|^2$$

- total variation in the data:

$$\frac{1}{n} \sum_{i=1}^n \|x_i\|^2$$

选择降维后的维度K(主成分的个数)

- 选择不同的K值，然后用下面的式子不断计算，选取能够满足下列式子条件的最小K值即可。

$$\frac{\frac{1}{n} \sum_{i=1}^n \|x_i - y_i\|^2}{\frac{1}{n} \sum_{i=1}^n \|x_i\|^2} \leq t$$

- 其中t值可以由自己定，比如t值取0.01，则代表了该PCA算法保留了99%的主要信息。当你觉得误差需要更小，你可以把t值设置的更小。

因子分析

因子分析

- 因子分析：用少数几个假想变量（称为因子）来表示其基本的数据结构，研究众多变量之间的内部依赖关系，探求观测数据中的基本结构。
- 因子分析常用于对不能直接观测的变量，称为**隐变量**（latent variable）。
- 例如，在研究学生的学习成绩时，可能会有语文成绩、数学成绩、英语成绩、物理成绩等多个变量。因子分析可以帮助我们找到隐藏在这些成绩背后的潜在因素，比如可能是“语言能力因子”和“数理逻辑能力因子”。

正交因子模型

- 如果有 p 维的观测变量： $X = (X_1, X_2, \dots, X_p)^T$ ，其均值和协方差矩阵分别为 μ 和 Σ 。
- 因子分析模型假定观测变量 X 可以表示为
 - m 维的公共因子(common factors) $F = (F_1, F_2, \dots, F_m)^T$ ($m < p$)，和
 - 特殊因子 (unique factors) $\epsilon = (\epsilon_1, \epsilon_2, \dots, \epsilon_p)^T$ 的线性组合：
$$X_1 - \mu_1 = a_{11}F_1 + a_{12}F_2 + \dots + a_{1m}F_m + \epsilon_1$$
$$X_2 - \mu_2 = a_{21}F_1 + a_{22}F_2 + \dots + a_{2m}F_m + \epsilon_2$$
$$\dots$$
$$X_p - \mu_p = a_{p1}F_1 + a_{p2}F_2 + \dots + a_{pm}F_m + \epsilon_3$$
- $A = (a_{ij})$ 是 $p \times m$ 的因子载荷矩阵。

正交因子模型

- 表达成矩阵形式: $X_{p \times 1} = \mu + A_{p \times m}F + \epsilon$, A 称为因子载荷阵。
上述模型和线性模型形式非常相像, 但是注意右边每个量我们均不能直接观测到, 即公共因子和特殊因子均为随机变量且不能直接观测到。因此需要假定一些结构才能进行推断。
- 正交因子模型假设包括:
- (1) $EF = 0, Var(F) = E(FF') = I_m$
- (2) $E\epsilon = 0, Var(\epsilon) = E(\epsilon\epsilon') = \Psi = diag\{\psi_1, \dots, \psi_p\}$
- (3) $cov(F, \epsilon) = 0$

正交因子模型

- 在正交因子模型假设下，有：

$$Cov(X_i, F_j) = a_{ij}$$

$$\Sigma = Var(X) = Var(AF + \epsilon) = AA^T + \epsilon$$

$$\sigma_{ii} = Var(X_i) = \sum_{j=1}^m a_{ij}^2 + \psi_i \coloneqq h_i^2 + \psi_i$$

$$\sigma_{ij} = cov(X_i, X_j) = \sum_{k=1}^m a_{ik} a_{jk}, \quad i \neq j$$

模型的基本形式

- 若观测变量进行了标准化，因子分析的模型可以简单写为：

$$X = AF + \epsilon.$$

- **公共因子** F ：能够同时影响多个观测变量的潜在因素。这些因子是隐藏在观测数据背后的共同成分，它们反映了变量之间的内在相关性。例如，在研究学生的各科学习成绩（包括语文、数学、英语等）时，可能存在“学习能力因子”和“学习努力程度因子”等公共因子。
- **特殊因子** ϵ ：指只对单个观测变量起作用的因素。它代表了每个观测变量中不能被公共因子解释的部分。特殊因子通常包括测量误差、变量的特殊性质或其他只与该变量自身相关的因素。例如，学生可能对数学科目的特别偏好，这种只影响数学成绩而与其他科目关系不大的因素就可以看作是数学成绩这个观测变量的特殊因子。

因子载荷矩阵

- 因子载荷矩阵 $A = (a_{ij})$ 中, a_{ij} 表示第 i 个观测变量 X_i 在第 j 个公共因子 F_j 上的载荷, 即 $a_{ij} = \text{Cov}(X_i, F_j)$, 在标准化变量 X 和 F 的情况下, 也等于 $\text{corr}(X_i, F_j)$ 。
- 可以将因子载荷矩阵看作是从原始观测变量空间 X 映射到因子空间 F 的一种坐标变换矩阵。
- 因子载荷 a_{ij} 的绝对值大小反映了第 i 个观测变量 X_i 与第 j 个公共因子 F_j 之间关系的紧密程度; 正负表示观测变量 X_i 与公共因子 F_j 之间的相关方向。

因子载荷矩阵

- 因子载荷矩阵 A 中的元素 a_{ij} 满足以下性质：
- 在公共因子方差被标准化为1的情况下， $\sum_{i=1}^p a_{ij}^2$ 表示第 j 个公共因子 F_j 对所有观测变量 X 的方差贡献。
- $\sum_{j=1}^m a_{ij}^2$ 表示观测变量 X_i 的共性方差 h_i^2 。
- 例如，对于因子载荷矩阵 $A = \begin{pmatrix} 0.7 & 0.3 \\ 0.5 & 0.6 \\ 0.8 & 0.2 \end{pmatrix}$ ，对于第一个公共因子 F_1 ，其对所有观测变量的方差贡献为 $0.7^2 + 0.5^2 + 0.8^2$ 。对于第一个观测变量 X_1 ，其共性方差为 $0.7^2 + 0.3^2$ 。

方差的分解

- 由于 a_{ij} 表示第*i*个观测变量 X_i 在第*j*个公共因子 F_j 上的载荷，而公共因子的方差假设为1。对于观测变量 X_i ，其方差 $Var(X_i)$ 可以分解为两部分：
- 共性方差：由公共因子引起的方差 $h_i^2 = \sum_{j=1}^m a_{ij}^2$ ，这部分方差反映了公共因子对观测变量 X_i 的贡献程度。
- 特殊方差：由特殊因子引起的方差 $\psi_i = Var(\epsilon_i)$ 。
- 所以有： $Var(X_i) = h_i^2 + \psi_i$ 。
- 例如，假设变量是标准化的，对于变量 X_1 ，如果 $a_{11} = 0.7$, $a_{12} = 0.3$ ，则共性方差 $h_1^2 = 0.7^2 + 0.3^2 = 0.58$ ，假设特殊方差 $\psi_i = 0.42$ ，那么 $Var(X_1) = 0.58 + 0.42 = 1$ 。

例：

- 假设有3个观测变量 X_1, X_2, X_3 和2个公共因子 F_1, F_2 ，则模型可写为：

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{pmatrix} \begin{pmatrix} F_1 \\ F_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \end{pmatrix}$$

- 观测变量：
- 公共因子：
- 因子载荷矩阵：
- 特殊因子：

例：

- 假设：

$$A = \begin{pmatrix} 0.7 & 0.3 \\ 0.5 & 0.6 \\ 0.8 & 0.2 \end{pmatrix}$$

- 因子载荷矩阵A的第 j 列元素 $a_{1j}, a_{2j}, \dots, a_{pj}$, $\sum_{i=1}^p a_{ij}^2$ 表示第 j 个公共因子 F_j 对所有观测变量 X 的方差贡献。贡献越大的公共因子，重要性越高。
- 因子载荷矩阵A的第 i 行元素 $a_{i1}, a_{i2}, \dots, a_{im}$, $\sum_{j=1}^m a_{ij}^2$ 表示观测变量 X_i 的共性方差 h_i^2 。共性方差越大说明公共因子对观测变量 X 的解释力度越高。

参数估计

- 常见的估计方法包括
 - 主成分法
 - 迭代主因子法
 - 极大似然法(假设正态)
- 主成分法对方差关注更多，迭代主因子法和极大似然法关注如何使用公共因子的波动来描述观测变量之间的相关性。

主成分法

- 设观测变量向量 $X = (X_1, X_2, \dots, X_p)^T$, 其中 p 为观测变量的个数。
- 假设 X 的协方差矩阵为 Σ , 其特征值为 $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$, 对应的单位特征向量分别为 e_1, e_2, \dots, e_p 。
- 首先对协方差矩阵 Σ 进行特征分解, 得到特征值和特征向量。
- 当提取 m 个公共因子 ($m < p$) 时, 因子载荷矩阵 $A = (a_{ij})$ 的计算方法如下:
 - $a_{ij} = \sqrt{\lambda_j} e_{ij}$, 其中 $i = 1, 2, \dots, p$; $j = 1, 2, \dots, m$ 。
 - 这里 λ_j 是第 j 个特征值, e_{ij} 是第 j 个特征向量的第 i 个分量。

主成分法(The principal component method)

- 由 Σ 的非负定性, 可以得到正交分解

$$\Sigma = \lambda_1 e_1 e_1^T + \cdots + \lambda_p e_p e_p^T := e \Lambda e^T$$

- 其中 $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ 为特征根, e_1, \dots, e_p 为相应的特征向量
- 记 $Y = e^T(X - \mu)$, 则 Y 为总体主成分, 有

$$X - \mu = eY = \sum_{j=1}^m e_j Y_j + \sum_{j=m+1}^p e_j Y_j := AF + \epsilon$$

- 其中 $A = [\sqrt{\lambda_1}e_1, \dots, \sqrt{\lambda_m}e_m]$, $F = \left(Y_1 / \sqrt{\lambda_1}, \dots, Y_m / \sqrt{\lambda_m} \right)^T$, $\epsilon = \sum_{j=m+1}^p e_j Y_j$.

主成分法

- 可以验证正交因子模型的假设条件 (1) 和 (3) 成立, 但是 (2) 未必成立。
- 相应地,

$$\Sigma = AA^T + Var(\epsilon)$$

- 若特殊因子 ϵ 对协方差的贡献很小, 则 $\Sigma \approx AA^T$, 从而对协方差的一个良好近似为

$$\Sigma \approx AA^T + diag\{\psi_1, \dots, \psi_p\}$$

- 其中 $\psi_i = \sigma_{ii} - \sum_{j=1}^m a_{ij}^2 = \sigma_{ii} - h_i^2$, $i = 1, \dots, p$.

主成分法

- 使用样本数据 x_1, \dots, x_n 估计上述参数。记 $(\hat{\lambda}_j, \hat{e}_j), j = 1, \dots, p$ 为样本协方差矩阵 S 的特征根和特征向量对，且 $\hat{\lambda}_1 \geq \dots \geq \hat{\lambda}_p \geq 0$
- 因子模型的主成分分解：

$$\hat{A} = \left[\sqrt{\hat{\lambda}_1} \hat{e}_1, \dots, \sqrt{\hat{\lambda}_m} \hat{e}_m \right],$$

$$\widehat{\Psi} = diag(S - \hat{A}\hat{A}^T) = diag\{\widehat{\Psi}_1, \dots, \widehat{\Psi}_p\}, \quad \widehat{\Psi}_i = s_{ii} - \sum_{j=1}^m \hat{a}_{ij}^2$$

因子个数 m 的确定

- 在一些问题里，因子个数 m 是事先取定的。
- 若 m 不能事先取定：
- **解释方差比例累积法：**类似主成分个数的选择方法，选择 m 使得

$$\frac{\sum_{i=1}^m \hat{\lambda}_i}{\hat{\lambda}_1 + \cdots + \hat{\lambda}_p}$$

- 较大。
- **Kaiser 准则：**当特征值大于 1 时，意味着该因子解释的方差比单个变量的平均方差大。因此选择特征根大于 1 的个数。
- **碎石图：**将特征值按照从大到小的顺序排列，并将其绘制成折线图，会看到特征值从大到小下降的趋势。通常在坡度变缓的转折点之前的因子被认为是有意义的因子，因为这些因子能够解释相对较多的方差，而转折点之后的因子解释方差的能力急剧下降，可能是不太重要的因子。

因子旋转

- 出于降维的需要，我们常常希望 m 要比 p 小得多，这样分解式 $\Sigma = AA^T + \Psi$ 通常只能近似成立。一般来说， m 选取得越小，上述近似效果就越差，即因子模型拟合得越不理想。拟合得太差的因子模型是没有什么实际意义的。
- 注意：公共因子 F 和负荷阵 A 不唯一：对任意正交矩阵 T 有
$$X - \mu = AF + \epsilon = (AT)(T^T F) + \epsilon = A^*F^* +$$
- A^* , F^* 满足正交因子模型的所有假设。因此因子载荷阵
 - $A^* = AT$ 和 A
- 在解释协方差 Σ 时候是等价的。

因子旋转

- 初始得到的因子载荷矩阵可能不容易解释公共因子的实际意义。
- 由正交矩阵的性质，称正交变换 AT , T^TF 为因子旋转。在合适的准则下对因子载荷阵 A 进行旋转，以期得到更易解释的结果。
- 因子旋转的目的：
 - 因子旋转在不改变公共因子对观测变量方差解释程度的前提下，对因子载荷矩阵进行变换，使每个变量在一个因子上的负荷量尽可能高，而在其他因子上的负荷量尽可能低，这让因子载荷在某些变量上更加集中，从而使公共因子的意义更加清晰，便于解释。

因子旋转

- 前面我们已经看到，在正交因子模型下，对因子载荷阵乘以一个正交矩阵 T 可以得到对协方差矩阵同样的逼近。
- 这意味着，我们使用 \hat{A} 或 $\hat{A}T$ 来估计因子载荷阵都是可以的，其中 T 为任意正交矩阵。
- 估计的残差矩阵 $S - (\hat{A}\hat{A}^T + \Psi) = S - ((\hat{A}T)(\hat{A}T)^T + \Psi)$ 保持不变，而且估计的特殊方差和共性方差均不变。
- 我们希望通过旋转因子来更好对结果进行解释：使因子载荷的平方两极分化，要么接近 0，要么接近 1。
- 因子旋转方法主要有正交旋转和斜交变换两类。

因子分析、回归分析

- **因子分析**: 探索和揭示观测变量 x 之间潜在的结构关系，通过少数几个不可观测的公共因子 F 来解释众多观测变量之间的相关性。要求观测变量之间存在一定程度的相关性，因为如果变量之间相互独立，就无法提取公共因子。
- **回归分析**: 研究预测变量与被预测变量之间的定量关系，建立一个能够根据自变量 x 的值来预测因变量值 y 的数学模型。对于自变量之间，一般要求不存在严重的多重共线性（即自变量之间不能有高度的线性相关性），否则会影响回归系数的估计。

因子分析、主成分分析

- **因子分析：**重点在于找出这些潜在的公共因子，并且这些因子是具有实际意义的潜在变量，用于解释观测变量之间的相关性。因子分析更侧重于对数据结构的解释，即寻找隐藏在观测变量背后的公共因子，并且要使这些因子具有明确的实际意义，以便更好地理解数据生成的内在机制。
- **主成分分析：**一种数据降维的技术。这些主成分是按照能够解释原始数据方差的大小依次排列的。它是在尽可能保留原始数据信息（方差）的前提下，将高维数据投影到低维空间，重点在于数据的降维。

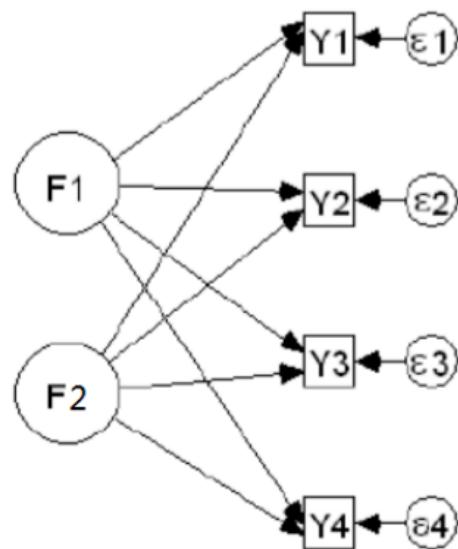
因子分析的分类

- 因子分析常包括探索性因子分析(Exploratory FA) 和验证性因子分析(Confirmatory FA) 两类。
- **探索性因子分析**常用来降维，降维的方式是试图用少数几个潜在的，不可观测的随机变量（因子）来描述原始变量间的协方差关系。“探索性”是指在没有对可观测变量之间，以及可观测变量与因子之间的线性关系赋予任何结构，而只指定隐变量 (latent variables) 的个数。
- **验证性因子分析**则常用来研究一个假设的因子模型对一组新的样本拟合程度如何，其允许对模型中的参数进行限制。

EFA vs CFA

Exploratory FA

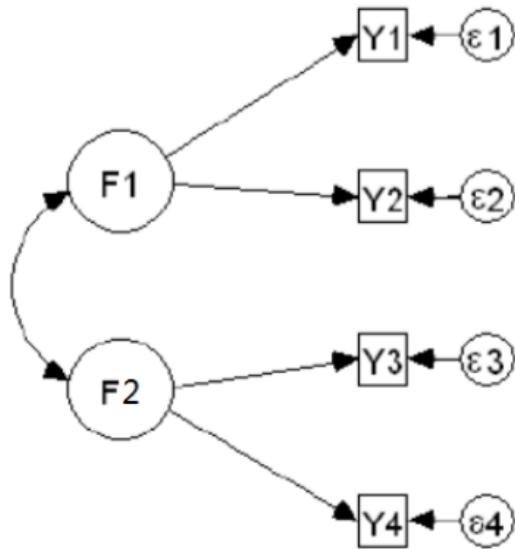
因子个数未知
载荷模式未知
(例如所有的因子在所有变量)
可以不是可识别的
因子是不相关的



Exploratory FA

Confirmatory FA

因子个数根据理论来设置
载荷模式由理论决定
(e.g. $Y_1 = l_1F_1 + \epsilon$)
一般是可识别的
因子可以是相关的



Confirmatory FA

智能手机用户满意度研究

- 一家智能手机制造商想要深入了解用户对其产品的满意度，但不确定影响满意度的潜在因素有哪些。
- **探索性因子分析阶段：**在线问卷收集用户对智能手机的评价，包括但不限于以下变量：电池续航时间满意度、屏幕分辨率满意度、色彩准确性满意度、相机拍照质量满意度、相机功能多样性满意度、外观设计美感满意度、机身材质质感满意度、系统响应速度满意度、软件兼容性满意度、价格接受程度。
- 进行探索性因子分析：根据特征值大于 1 的准则和碎石图，提取出可能的因素。通过方差最大旋转后的因子载荷矩阵发现，第一个因子在电池续航时间满意度、系统响应速度满意度和软件兼容性满意度上有较高载荷，可初步解释为“手机性能体验因子”；第二个因子在相机拍照质量满意度和相机功能多样性满意度、外观设计美感满意度和机身材质质感满意度上载荷较大，解释为“手机外观与相机功能因子”；第三个因子在价格接受程度上有突出载荷，定义为“价格接受因子”。

智能手机用户满意度研究

- **验证性因子分析阶段：**基于探索性因子分析的结果，制造商提出了一个理论模型，即用户对智能手机的满意度主要由“手机性能体验”、“手机外观与相机功能”和“价格接受”这三个潜在因子决定。现在要验证这个模型是否正确。
- 再次收集一组新用户的数据，并且按照之前的潜在因子设定构建验证性因子分析模型。
- 通过探索性因子分析初步挖掘出影响智能手机用户满意度的潜在因子结构，然后通过验证性因子分析在新的数据集中验证了这个结构的合理性。

词嵌入

