

第13章：连续变量关系探索与变量压缩

《Python数据科学：技术详解与商业实践》

讲师：Ben

自我介绍

- 天善商业智能和大数据社区 讲师 – Ben
- 天善社区 ID - Ben_Chang
- <https://www.hellobi.com> – 学习过程中有任何相关的问题都可以提到技术社区数据挖掘版块。

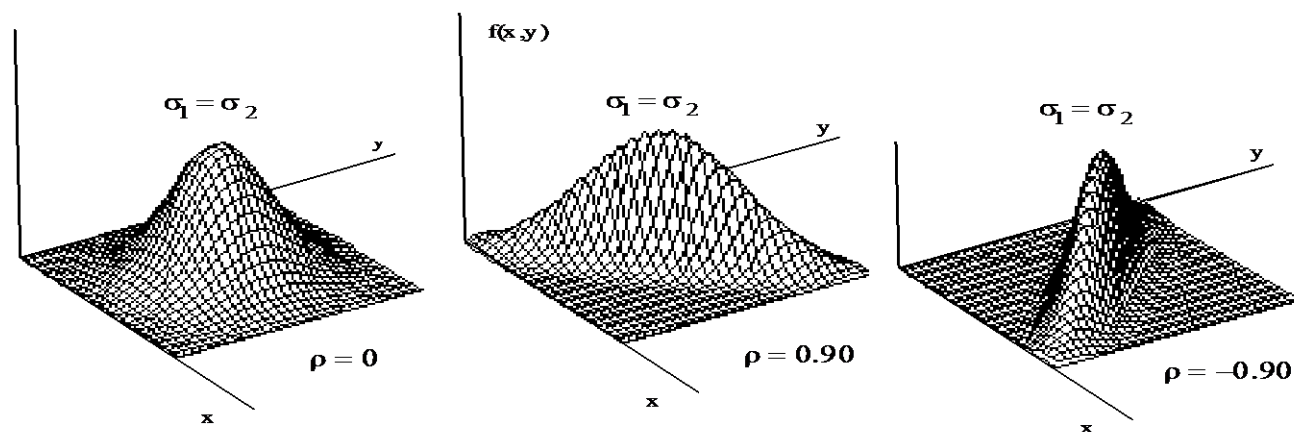
- 多元统计基础与变量约减的思路
- 主成分分析
- 因子分析
- 稀疏主成分分析
- 变量聚类

多元统计基础与变量约减的思路

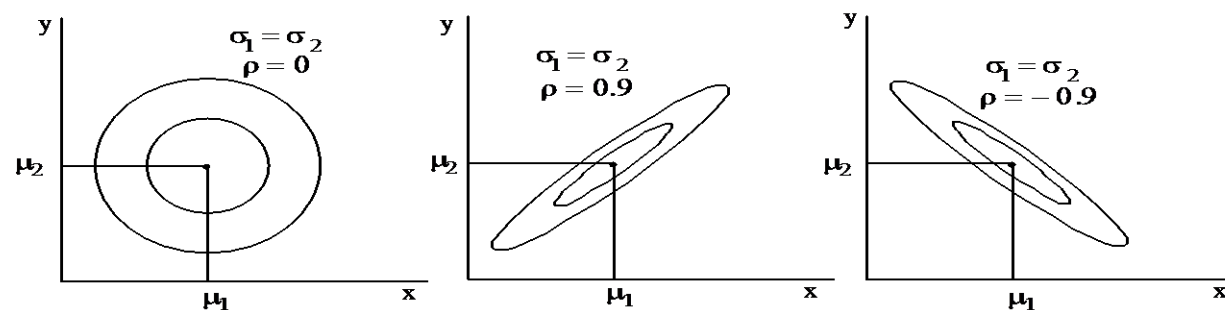


两正态分布变量之间的关系

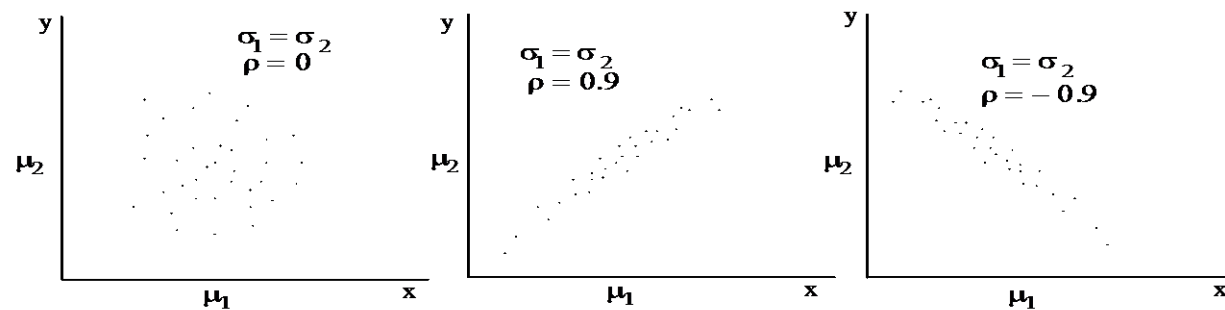
The Bivariate Normal Distribution



Contour Plots of the Bivariate Normal Distribution



Scatter Plots of data from the Bivariate Normal Distribution



- 使用散点图来查看两个连续变量间的关系。
- 使用相关性统计来量化两个连续变量的相关性。
- 描述一下错误使用相关系数的可能情形。
- 使用相关（Correlations）任务获得皮尔森相关系数。

三个连续变量-短信量、微信和Web登陆



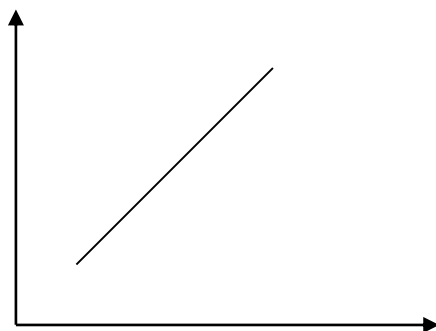
变量之间的依赖关系有两种不同的类型：

- 函数关系，即当一个或多个变量的数值确定以后，另一个变量的数值按照某种关系也随之被确定；
- 相关关系，即变量之间不存在确定的函数关系，只是存在某种非确定性的联系，这种依赖关系我们将用相关分析来研究。

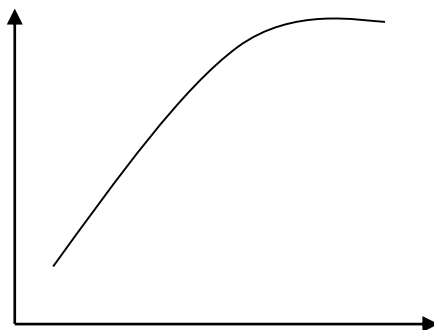
简单相关分析

相关关系是一种不完全确定的随机关系，当一个或几个变量的数值被确定后，与之相应的另一个变量的值虽然不能确定，但是仍按照某种依赖关系在一定的范围内变化。

简单相关分析是研究两个变量之间相关关系的方法。按照变量性质的不同，所采用的相关分析方法也不同。对于连续变量，通常使用Pearson相关系数来描述变量间的相关关系；对于有序变量，则常使用Spearman相关系数。



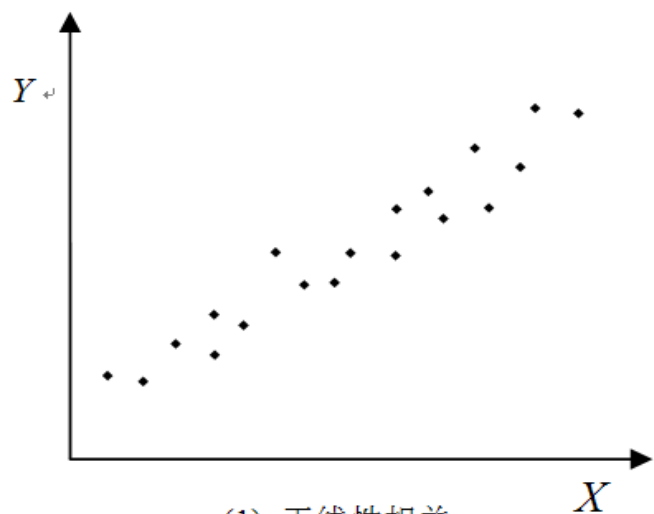
Pearson



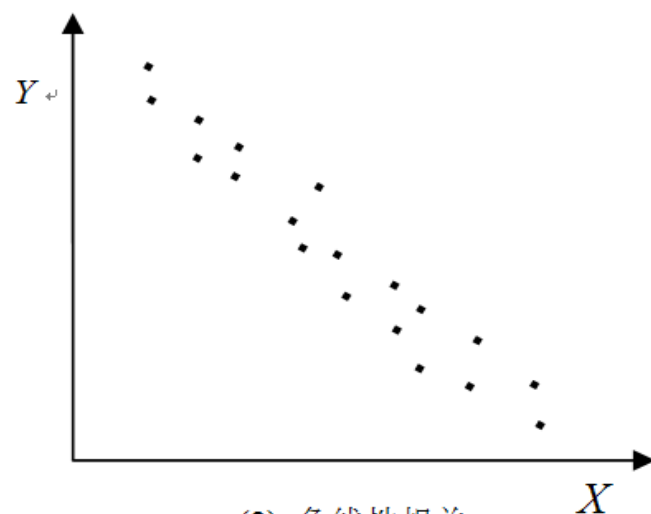
Spearman

简单相关分析

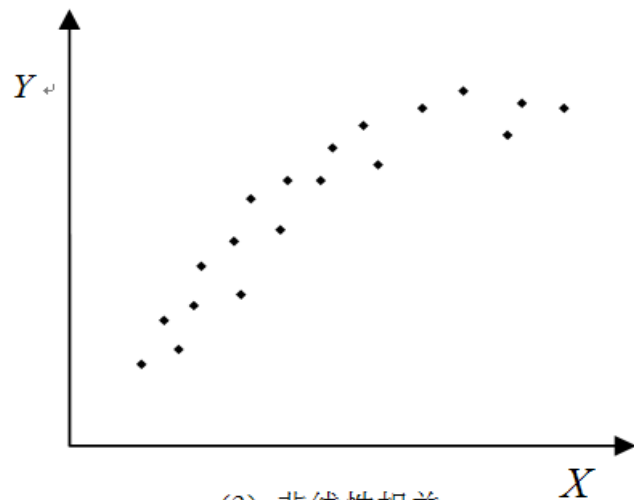
两个变量之间的相关关系也可以通过散点图来进行直观描述：



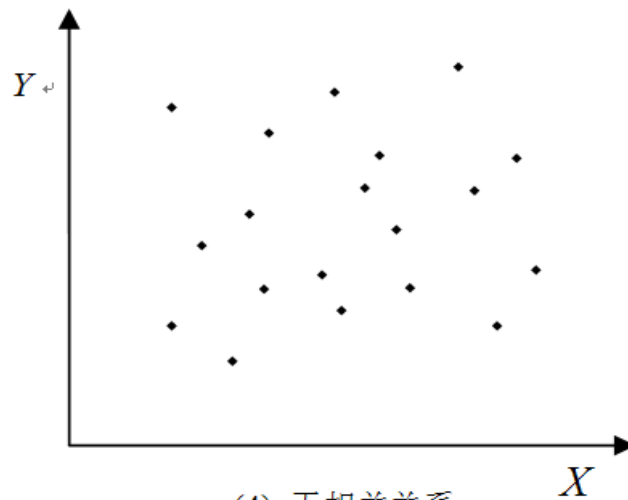
(1) 正线性相关



(2) 负线性相关



(3) 非线性相关



(4) 无相关关系

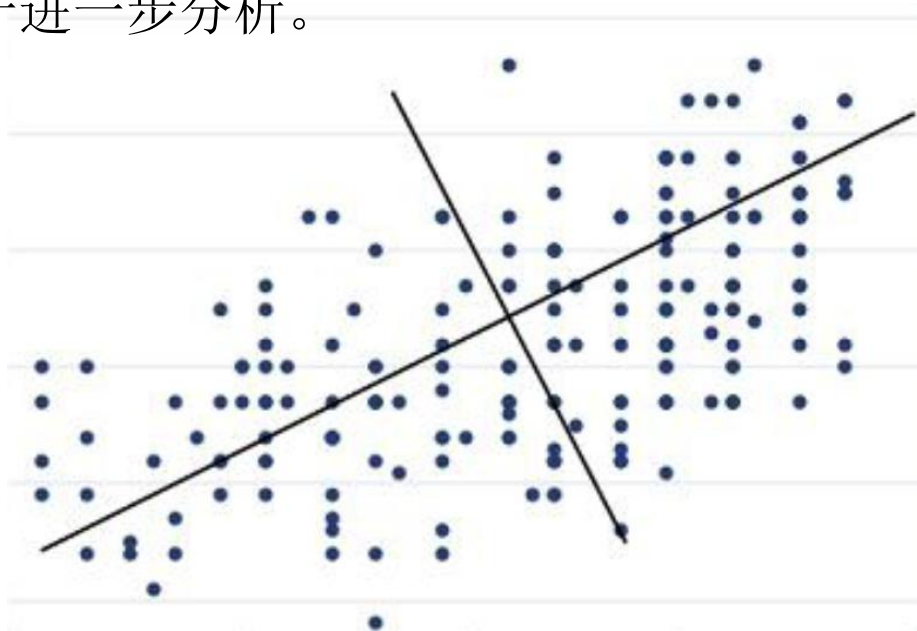


- 使用散点图和相关系数矩阵，对表“PROFILE_Bank”的连续变量之间的关系进行描述。

主成分分析

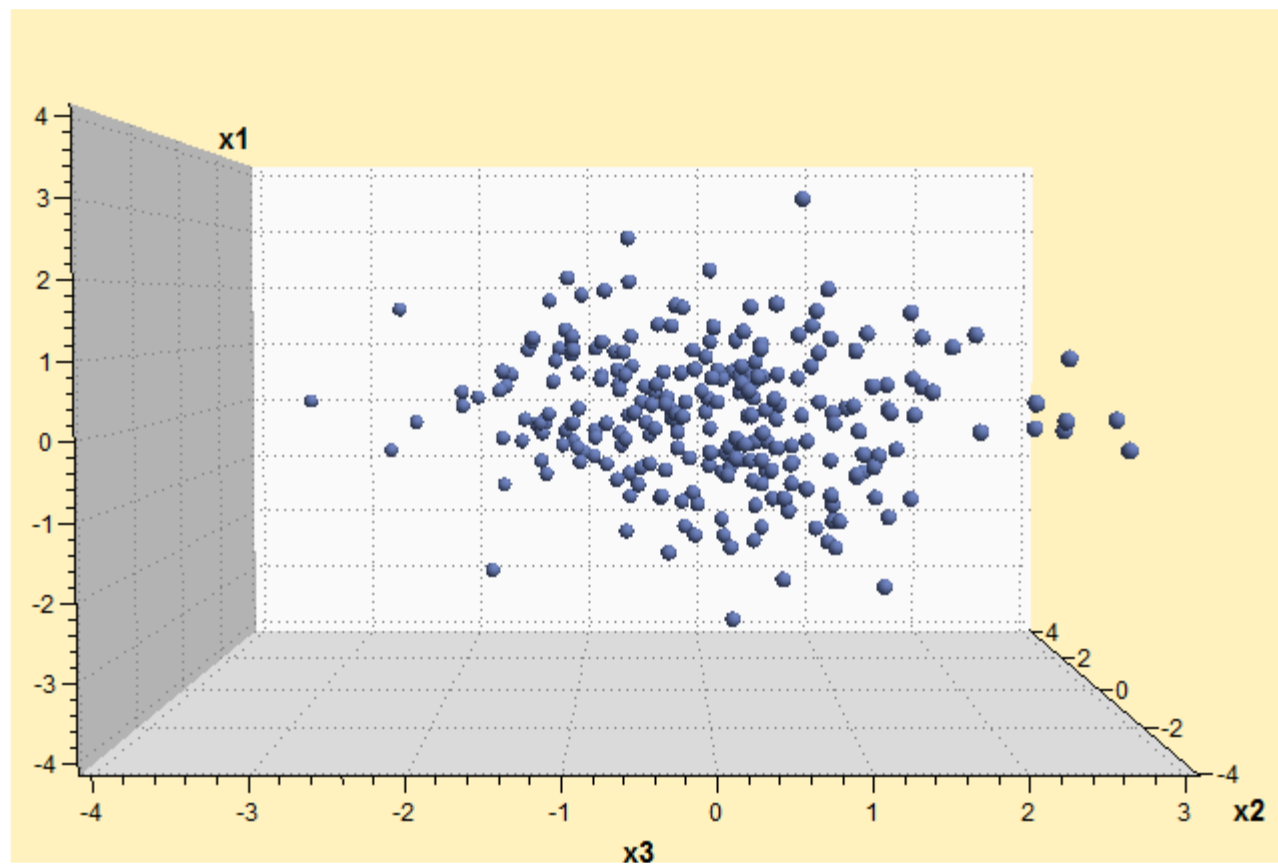
主成分分析的思路

- 主成分分析的目的在于构造输入变量的少数线形组合，尽量能解释数据的变异性。这些线形组合被称为主成分，它们形成的降维数据可用于进一步分析。



- 第一个主成分由图中比较长的直线代表，在这个方向上能够最多地解释数据的变异性，即方差最大；
- 第二个主成分由图中比较短的直线代表，与第一个主成分正交，能够最多的解释数据中剩余的变异性；
- 一般而言，每个主成分都需要与之前的主成分正交，并且能够最多的解释数据中剩余的变异性。

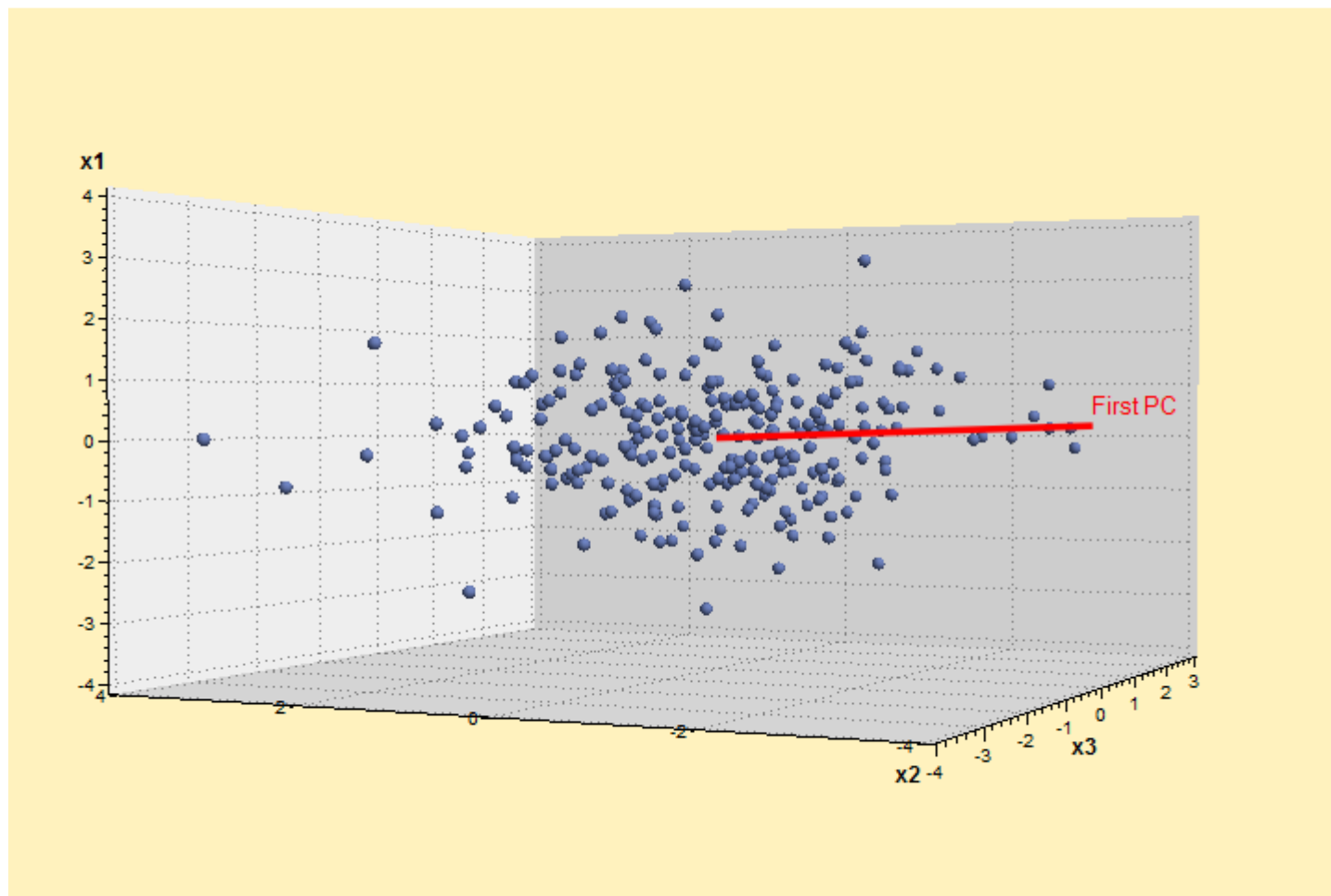
三维变量之间的关系



- 三维空间上的相关连续变量呈椭球状分布。只有这样的分布才可以做主成分分析。如果呈球形分布，这说明变量间没有相关关系，没有必要做主成分分析，也不能做变量的压缩。

提取第一个主成分

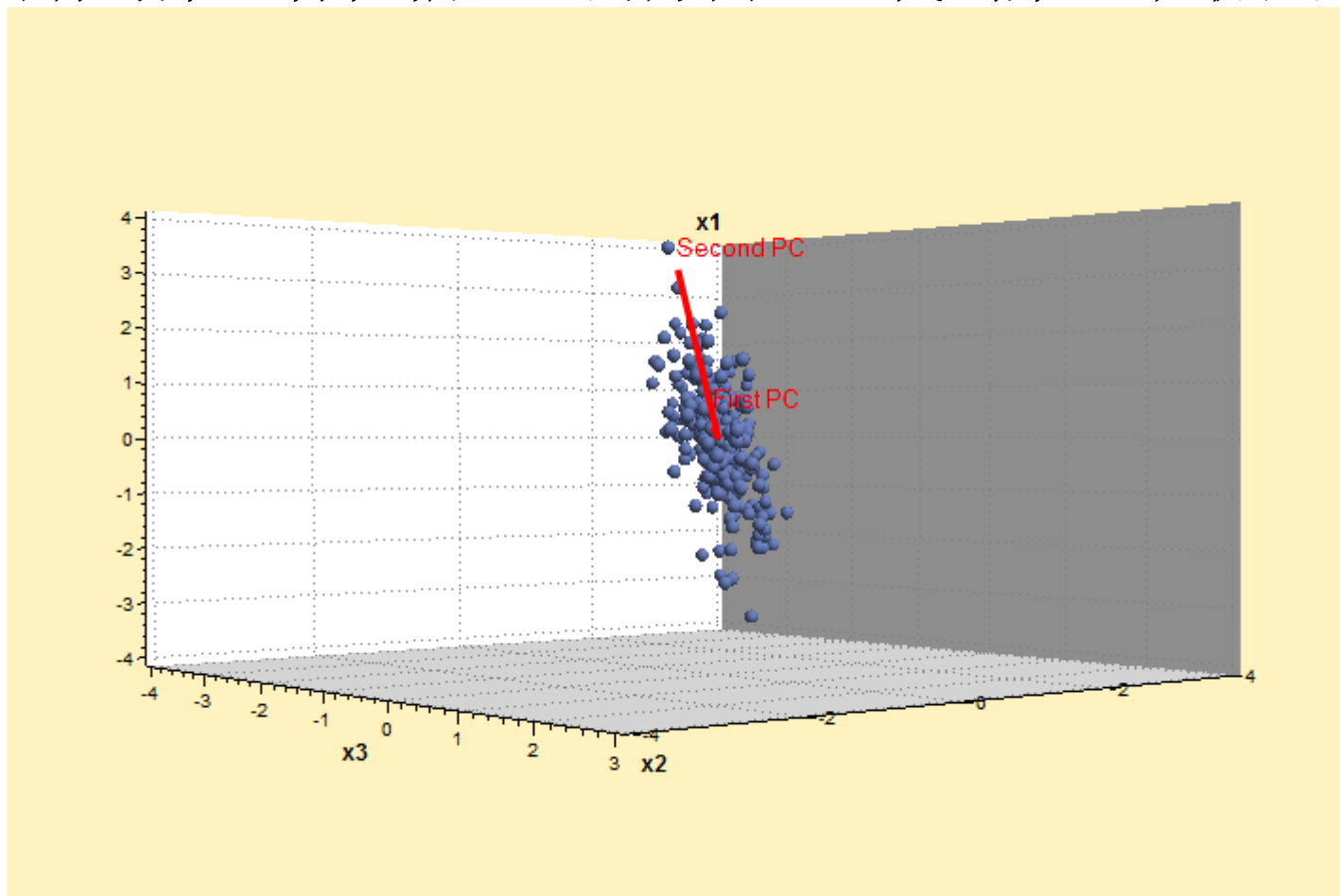
首先找到这个空间椭球的最长轴，即数据变异最大的轴



第一特征根=1.94

提取第二个主成分

在所有与第一特征根垂直的方向上，找到第二个最长的轴



第一特征根=1.94

第二特征根=1.02

公式化表述1-主成分建模

用 $\underline{X}' = (X_1, X_2, \dots, X_p)$ 表示随机向量，它的方差-协方差矩阵为 Σ

$$Z_1 = \underline{a}_1' \underline{X} = a_{11}X_1 + a_{12}X_2 + \dots + a_{1p}X_p$$

$$Z_2 = \underline{a}_2' \underline{X} = a_{21}X_1 + a_{22}X_2 + \dots + a_{2p}X_p$$

\vdots

$$Z_p = \underline{a}_p' \underline{X} = a_{p1}X_1 + a_{p2}X_2 + \dots + a_{pp}X_p$$

则 \mathbf{Z} 的方差为： $Var(Z_i) = \underline{a}_i' \Sigma \underline{a}_i$

Z_1, Z_2, \dots, Z_p 就是需要寻找的主成分，我们要求每个主成分两两之间是正交的。

- 有多少个变量就会有多个正交的主成分；
- 主成分的变异（方差）之和等于原始变量的所有变异；
- 前若干个主成分的变异（方差）解释了绝大多数的变异（方差）；
- 如果原始变量不相关，即没有协方差，则不需要做主成分。

公式化表述2-特征值与特征向量

令 $|\lambda I - \Sigma| = 0$ 为特征根方程, 对其求解, 得到特征根, 按大小排序, $\lambda_1 > \lambda_2 > \dots > \lambda_p$, $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_p$ 是分别属于它们的特征向量, 特征向量线性无关。这样的变换在线形代数中称为线形变化。

特征向量 λ_1 是线性无关的。

对 Σ 进行进行上述正交分解, 便得到这样的 $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \dots, (\lambda_p, \mathbf{e}_p)$ 这样的特征根-特征向量对。

则第 i 主成分为:

$$Z_i = \mathbf{e}_i' \mathbf{X} = e_{i1}X_1 + e_{i2}X_2 + \dots + e_{ip}X_p, \quad i = 1, 2, \dots, p$$

主成分的方差-协方差为:

$$\text{Var}(Z_i) = \mathbf{e}_i' \Sigma \mathbf{e}_i = \lambda_i, \quad i = 1, 2, \dots, p$$

$$\text{Cov}(Z_i, Z_k) = \mathbf{e}_i' \Sigma \mathbf{e}_k = 0, \quad i \neq k.$$

公式化表述3-主成分的个数选取

令 $\sigma_1, \sigma_2, \dots, \sigma_p$ 表示原始变量的方差序列，它们之和等于主成分之和。它们之间的区别在于主成分是从大到小排序的。

$$\sigma_{11} + \sigma_{22} + \dots + \sigma_{pp} = \sum_{i=1}^p \text{Var}(X_i) = \lambda_1 + \lambda_2 + \dots + \lambda_p = \sum_{i=1}^p \text{Var}(Z_i)$$

每个主成分解释的变异为： $\frac{\lambda_k}{\lambda_1 + \lambda_2 + \dots + \lambda_p}$

原始变量单位不一致情况下，原始变量需要进行学生标准化，则所有原始变量的方差为1。

主成分个数的选取原则：

- 单个主成分解释的变异不因该小于1，比如选取3个主成分，第3个主成分解释的变异相当于一个原始变量的变异；
- 选取主成分累积的解释变异达到80%-90%。

基于相关系数矩阵的主成分分析

因为主成分是通过最大化线性组合的方差来得到的，所以它对变量的测量尺度非常敏感。

- ▶ 例如，若一个输入变量是“企业销售额（元）”，最大观测和最小观测可以相差几千万，而另一个变量是“企业雇员数”，最大观测和最小观测只相差几千，因为“企业销售额”的方差比“企业雇员数”的方差大得多，所以它会主导主成分分析的结果，使得第一个主成分可能几乎等于“企业销售额”，而忽略了输入变量之间的关系。
- ▶ 使用“万元”作为测量单位和使用“元”作为测量单位得到的主成分分析的结果会相差很大。

在实际应用中，通常首先将各输入变量进行标准化，使每个变量均值为0，方差为1，这等价于使用相关系数矩阵 \mathbf{R} 替代协方差矩阵 $\mathbf{\Sigma}$ 来进行主成分分析。

主成分的解释

我们可以从两个方面来解释所得的第 i 个主成分:

1. 考察第 i 个主成分对应的系数 (即 e_{i1}, \dots, e_{ip}), 根据系数绝对值较大的输入变量来解释第 i 个主成分。值得注意的是, 系数的正负本身没有意义, 这是因为 Σ 或 R 的任意特征向量 \mathbf{e} 取负之后仍然是特征向量。但是, 系数之间的正负对比是有意义的。
2. 计算第 i 个主成分与各输入变量的相关系数, 根据那些对应相关系数的绝对值较大的输入变量来解释第 i 个主成分。

主成分分析的三种运用场景

- 1、做一个综合打分：这种情况在日常中经常遇到，比如高考成绩的加总、员工绩效的总和排名。这类情况要求只出一个综合打分，因此主成分分析比较适合。相对于讲单项成绩简单加总的方法，主成分分析会赋予区分度高的单项成绩以更高的权重，分值更合理。不过当主成分分析不支持取一个主成分时，就不能使用该方法了。
- 2、对数据进行描述：描述产品情况，比如著名的波士顿矩阵，子公司业务发展状况，区域投资潜力等，需要将多变量压缩到少数几个主成分进行描述，如果压缩到两个主成分是最理想的。这类分析一般做主成分分析是不充分的，做到因子分析更好。
- 3、为聚类或回归等分析提供变量压缩：消除数据分析中的共线性问题，消除共线性常用的有三种方法，分别是：1) 同类变量中保留一个最有代表性的；2) 保留主成分或因子；3) 从业务理解上进行变量修改。这主成分是三种方法的基础。

演示一：做一个综合打分



- 使用“**Loan_aply**”数据对客户信用进行评级打分。

某金融服务公司为了了解贷款客户的信用程度，评价客户的信用等级，采用信用评级常用的**5C**方法，说明客户违约的可能性。

- 品格：指客户的名誉；
- 能力：指客户的偿还能力；
- 资本：指客户的财务实力和财务状况；
- 担保：指对申请贷款项担保的覆盖程度；
- 环境：指外部经济、政策环境对客户的影响。

每个单项都是由专家打分给出的。

演示一：做一个综合打分

步骤一：变量之间相关系数，多数变量之间有显著的强线性相关，这表明做主成分分析是有意义的。

Pearson 相关系数, N = 10 Prob > r under H0: Rho=0					
	X1	X2	X3	X4	X5
X1	1.00000	0.72666	0.82534	0.67631	0.68556
品格		0.0173	0.0033	0.0318	0.0286
X2	0.72666	1.00000	0.92908	0.93838	0.84141
能力	0.0173		0.0001	<.0001	0.0023
X3	0.82534	0.92908	1.00000	0.88346	0.73348
资本	0.0033	0.0001		0.0007	0.0158
X4	0.67631	0.93838	0.88346	1.00000	0.76256
附带担保品	0.0318	<.0001	0.0007		0.0103
X5	0.68556	0.84141	0.73348	0.76256	1.00000
环境条件	0.0286	0.0023	0.0158	0.0103	

可以看出，能力与资本、附带担保品有着较强的相关性，表明客户的偿还能力与其财务实力、财务状况和抵押资产有着重要的关系。

演示一：做一个综合打分

结果分析1：

- 总方差：原始变量总的变异；
- 特征值：每个主成分解释变异的数量；
- 比例：每个特征根解释的变异占原始数据总变异的比率；
- 累积：累积到当前的主成分，总共解释总变异的比率。

总方差 485.3147778

协方差矩阵的特征值				
	特征值	差分	比例	累积
1	410.505594	367.241611	0.8459	0.8459
2	43.263983	22.594102	0.0891	0.9350
3	20.669881	12.599063	0.0426	0.9776
4	8.070819	5.266318	0.0166	0.9942
5	2.804500		0.0058	1.0000

可以看出第一个主成分解释了**84.6%**的变异，根据选择主成分个数的第二个原则，超过了**80%**，这表明使用第一个主成分作为每家贷款企业的信用打分是适宜的。

演示一：做一个综合打分

结果分析2:

		特征向量				
		PRIN1	PRIN2	PRIN3	PRIN4	PRIN5
X1	品格	0.468814	-.830612	0.021406	0.254654	-.158081
X2	能力	0.484876	0.329916	0.014801	-.287720	-.757000
X3	资本	0.472744	-.021174	-.412719	-.588582	0.509213
X4	附带担保品	0.461747	0.430904	-.240845	0.706283	0.210403
X5	环境条件	0.329259	0.122930	0.878054	-.084286	0.313677

特征向量提供了由原始变量到每个主成分的转换系数（权重）。

$$Z_i = \underline{e}_i' \underline{X} = e_{i1}X_1 + e_{i2}X_2 + \cdots e_{ip}X_p, \quad i = 1, 2, \dots, p$$

第一个主成分的计算公式为：

$$P1 = 0.469 * \text{品格} + 0.485 * \text{能力} + 0.473 * \text{资本} + 0.462 * \text{担保品} + 0.329 * \text{环境条件}$$

利用特征向量的取值也可以对主成分进行解释，对第一主成分而言，各变量所占比重大致相等，且均为正数，说明第一主成分是对所有指标的一个综合测度，作为综合的信用等级指标，可以用于排序。

演示一：做一个综合打分

结果分析3:

获取打分结果:

⑫ 客户ID	⑫ PRIN1
7	35.877457883
3	25.091224289
6	13.619783398
1	3.165810157
4	-4.356665011
5	-6.407140081
2	-9.00917313
8	-10.34402785
10	-13.83321647
9	-33.80405318

在正确评估了客户的信用等级后,就能正确制定出对其信用期限、收款等政策,用于加强应收装款的管理工作。

演示二：做样本特征描述



• “**cities_10**”记录了十个沿海省份的经济指标，如何对这些省份的经济发展情况进行表述？。

省份	GDP	人均GDP	工业增加值	第三产业增加值	固定资产投资	基本建设投资
辽宁	5458.2	13000	1376.2	2258.4	1315.9	529.6
山东	10550.0	11643	3502.5	3851.0	2288.7	1070.1
河北	6076.6	9047	1406.7	2092.6	1161.6	591.5
天津	2022.6	22068	822.8	960.0	703.7	361.9
江苏	10636.3	14397	3536.3	3967.2	2320.0	1141.1
上海	5408.8	40627	2196.2	2755.8	1970.2	779.5
浙江	7670.0	16570	2356.5	3065.0	2296.6	1180.1
福建	4682.0	13510	1047.1	1859.0	964.5	397.1
广东	11769.7	15030	4224.6	4793.6	3022.9	1271.1

演示二：做样本特征描述

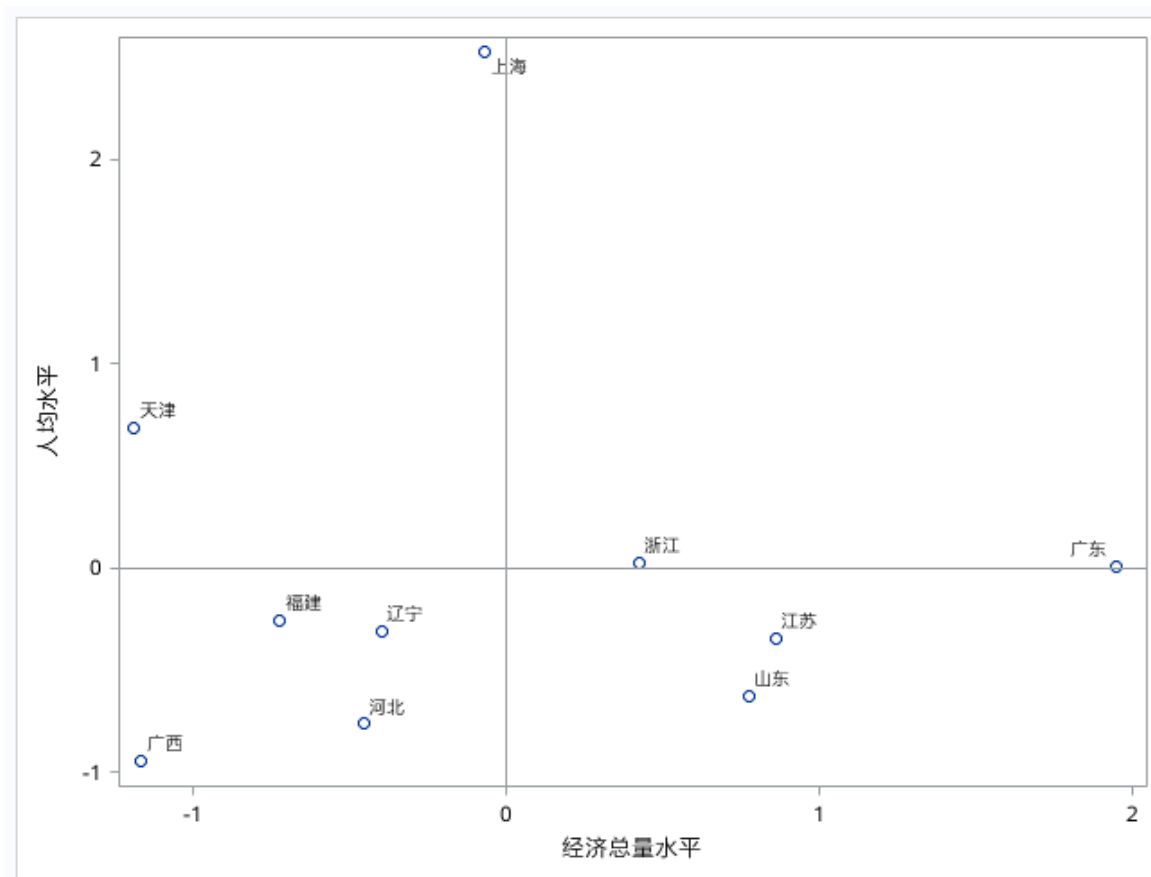
主成分结果：

		特征向量								
		PRIN1	PRIN2	PRIN3	PRIN4	PRIN5	PRIN6	PRIN7	PRIN8	PRIN9
X1	GDP	0.353682	-.212192	-.247627	0.384183	0.039060	0.165180	0.523242	0.414956	-.382345
X2	人均GDP	0.040555	0.942778	-.127315	0.121282	0.043348	0.229653	0.081859	0.116571	0.060800
X3	工业增加值	0.364148	-.009845	-.183606	0.341020	0.610942	-.238835	-.534805	0.000131	0.031377
X4	第三产业增加值	0.367584	-.045377	-.154498	0.301899	-.200596	-.012882	0.301955	-.612127	0.491456
X5	固定资产投资	0.365917	0.095213	-.165382	-.283114	-.220270	0.126849	-.223228	-.452498	-.656003
X6	基本建设投资	0.352119	-.023027	-.315878	-.736801	0.253840	-.049248	0.208173	0.187494	0.295764
X7	社会消费品零售总额	0.364419	-.135241	0.150223	0.023784	-.375304	0.545687	-.441122	0.329374	0.289954
X8	海关出口总额	0.297565	0.048047	0.802794	-.081281	0.385563	0.172984	0.233196	-.140819	-.072588
X9	地方财政收入	0.355405	0.183830	0.265924	0.007328	-.425018	-.718717	-.013739	0.267018	-.021050

第一个主成分在表达经济总量的指标上的权重相当，而第二个主成分只在人均GDP上权重很高，因此可以为每个变量取一个名字。

演示二：做样本特征描述

作结果展现



注：如果一个数据的变量可以被压缩为两个主成分，则通过展现在二维图形上已经可以完成样本聚类的工作。如果因子多于两个，则需要使用聚类算法进行样本分类。

思考题:

“PROFILE_BANK”记录了银行客户产品使用频数的信息,我们希望使用这个数据作银行客户的客户画像, 首先如何对这些信息进行约减?

“CITIES_10” 记录了十个沿海省份的经济指标, 希望用于做聚类分析。

请回答:

- 1、是否可以尝试着给每个主成分取一个名字, 用以表达这个主成分所测量的维度?
- 2、主成分分析是否可以达到变量分类、维度分析的目的? 什么情况下可以完成, 什么情况下不能完成。

练习解答:

PROFILE_BANK

		特征向量			
		PRIN1	PRIN2	PRIN3	PRIN4
CNT_TBM	柜面次数	0.303020	0.834245	0.445132	0.118622
CNT_ATM	ATM机次数	0.555131	-.377566	0.135542	0.728630
CNT_POS	POS机次数	0.559520	-.315486	0.386716	-.661708
CNT_CSC	有偿服务次数	0.535673	0.248894	-.796201	-.131035

CITIES_10

		特征向量								
		PRIN1	PRIN2	PRIN3	PRIN4	PRIN5	PRIN6	PRIN7	PRIN8	PRIN9
X1	GDP	0.353682	-.212192	-.247627	0.384183	0.039060	0.165180	0.523242	0.414956	-.382345
X2	人均GDP	0.040555	0.942778	-.127315	0.121282	0.043348	0.229653	0.081859	0.116571	0.060800
X3	工业增加值	0.364148	-.009845	-.183606	0.341020	0.610942	-.238835	-.534805	0.000131	0.031377
X4	第三产业增加值	0.367584	-.045377	-.154498	0.301899	-.200596	-.012882	0.301955	-.612127	0.491456
X5	固定资产投资	0.365917	0.095213	-.165382	-.283114	-.220270	0.126849	-.223228	-.452498	-.656003
X6	基本建设投资	0.352119	-.023027	-.315878	-.736801	0.253840	-.049248	0.208173	0.187494	0.295764
X7	社会消费品零售总额	0.364419	-.135241	0.150223	0.023784	-.375304	0.545687	-.441122	0.329374	0.289954
X8	海关出口总额	0.297565	0.048047	0.802794	-.081281	0.385563	0.172984	0.233196	-.140819	-.072588
X9	地方财政收入	0.355405	0.183830	0.265924	0.007328	-.425018	-.718717	-.013739	0.267018	-.021050

对于第一个例子，第一个主成分是对所有指标的一个综合测度，作为综合的信用等级指标。第二个主成分有正有负，是一个调和指标。不能说第一、二个主成分分别解释哪个变量，因此不能做到变量分类，也不能为每个主成分起名字。第二个例子就有所不同，第一个主成分在表达经济总量的指标上的权重相当，而第二个主成分只在人均GDP上权重很高，因此可以为每个变量取一个名字

总结:

说明：仅提取变量的主要信息，无法完成维度分析的功能。像“**CITIES_10**”这样变量本身就具有很好的分类表现的数据是很少见的。完成变量聚类的主要方法下面介绍的因子分析。

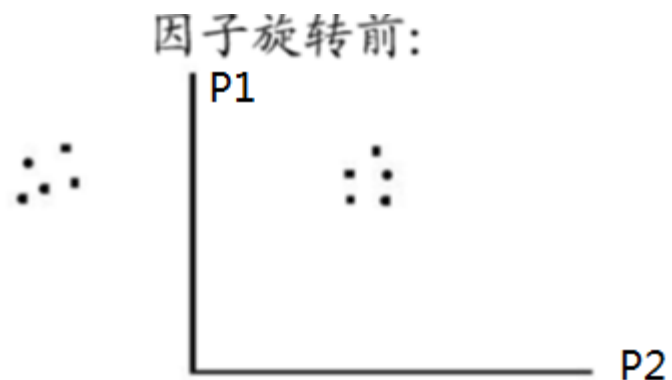
主要用途：**1**、对个体的情况或表现进行打分；**2**、一种简单省力的综合信息的手段，降低变量之间的关系，作为预测类模型的输入变量。



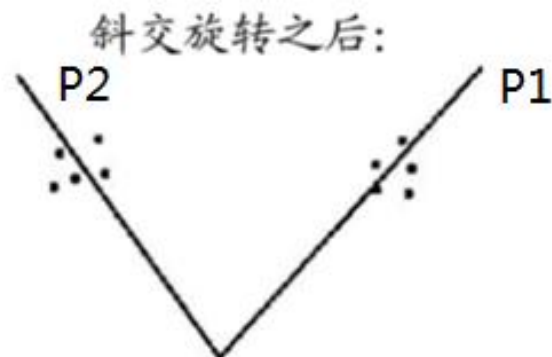
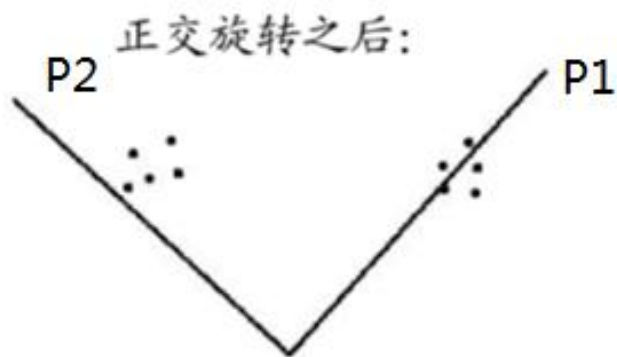
因子分析

因子分析的思路

- 继续主成分分析的思路，就象之前例子中呈现的那样，一般得到的主成分中，第一个主成分是综合指标，第二个主成分是调和指标。下图是以每个变量在这两个主成分上的权重作的散点图。



- 如果可以将主成分的坐标轴进行旋转，使得一些变量的权重的绝对值在一个主成分上达到最大，而在其他主成分上绝对值最小，这样就达到了变量分类的目的。变量旋转分为正交和非正交两种，一般使用前者。



公式化表述1-正交因子模型

- ▶ 假设 $\mathbf{X} = (X_1, \dots, X_p)^\top$ 是一个 p 维随机向量。
- ▶ \mathbf{X} 的均值向量为 $\boldsymbol{\mu} = (\mu_1, \dots, \mu_p)^\top$ 。
- ▶ \mathbf{X} 的协方差矩阵为 $\boldsymbol{\Sigma}$ ，其对角线上的值 σ_k^2 给出了 X_k 的方差 ($k = 1, \dots, p$)。
- ▶ 令 F_1, \dots, F_q ($q \leq p$) 表示 q 个公共因子。
- ▶ 令 $\varepsilon_1, \dots, \varepsilon_p$ 表示特殊因子。

$$X_1 - \mu_1 = l_{11}F_1 + l_{12}F_2 + \dots + l_{1q}F_q + \varepsilon_1$$

$$X_2 - \mu_2 = l_{21}F_1 + l_{22}F_2 + \dots + l_{2q}F_q + \varepsilon_2$$

$$\dots \quad \dots$$

$$X_p - \mu_p = l_{p1}F_1 + l_{p2}F_2 + \dots + l_{pq}F_q + \varepsilon_p$$

写成矩阵的形式是: $\mathbf{X} - \boldsymbol{\mu} = \mathbf{LF} + \boldsymbol{\varepsilon}$, 其中:

- ▶ $\mathbf{F} = (F_1, \dots, F_q)^\top$ 是 q 维随机向量;
- ▶ $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_p)^\top$ 是 p 维随机向量;
- ▶ \mathbf{L} 称为因子载荷矩阵, 其第 k 行第 i 列的值 l_{ki} 表示 X_k 在因子 F_i 上的载荷。

因子载荷矩阵的估计是因子分析的主要问题之一。令 Ψ 表示对角元素为 ψ_k 的 p 维对角矩阵。将前面的结论1和结论2写成矩阵形式可以得出：

$$\Sigma = LL^T + \Psi。 \quad (1)$$

- ▶ LL^T 是 Σ 中能被公共因子解释的部分；
- ▶ Ψ 是 Σ 中不能被公共因子解释而归结于特殊因子的部分。
- ▶ 注意，尽管开始 Ψ 被定义为一个对角矩阵，但实际上却不一定能够找到一个对角矩阵正好满足上式。

公式化表述2-主成分法

- ▶ 令 $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$ 表示 Σ 的特征值, $\mathbf{e}_1, \cdots, \mathbf{e}_p$ 表示对应的特征向量。统计理论证明 Σ 可以拆分为

$$\Sigma = \lambda_1 \mathbf{e}_1 \mathbf{e}_1^\top + \lambda_2 \mathbf{e}_2 \mathbf{e}_2^\top + \cdots + \lambda_p \mathbf{e}_p \mathbf{e}_p^\top。$$

- ▶ 对任意 $i = 1, \cdots, q$, 令因子载荷矩阵 $\tilde{\mathbf{L}}$ 的第 i 列为 $\sqrt{\lambda_i} \mathbf{e}_i$, 那么 $\tilde{\mathbf{L}} \tilde{\mathbf{L}}^\top = \lambda_1 \mathbf{e}_1 \mathbf{e}_1^\top + \lambda_2 \mathbf{e}_2 \mathbf{e}_2^\top + \cdots + \lambda_q \mathbf{e}_q \mathbf{e}_q^\top。$
- ▶ 对任意 $1 \leq k \leq p$, 令 $\tilde{\Psi}_k = \sigma_k^2 - \sum_{i=1}^q \tilde{l}_{ki}^2$, 令 $\tilde{\Psi}$ 为对角元素为 Ψ_k 的 p 维对角矩阵。
- ▶ $\tilde{\mathbf{L}}$ 和 $\tilde{\Psi}$ 就是 \mathbf{L} 和 Ψ 的估计值。

公式化表述3-最大方差旋转 (varimax rotation)

应用最广泛的因子旋转方法:

- ▶ 它是一种正交旋转;
- ▶ 目的是使载荷平方的方差最大化, 即最大化

$$\sum_{k=1}^p \sum_{i=1}^q \left(l_{ki}^2 - \frac{1}{pq} \sum_{k'=1}^p \sum_{i'=1}^q l_{k'i'}^2 \right)^2。$$

因子分析演示一：



” **cities_10**”记录了十个沿海省份的经济指标，希望用于做聚类分析。

省份	GDP	人均GDP	工业增加值	第三产业增加值	固定资产投资	基本建设投资
辽宁	5458.2	13000	1376.2	2258.4	1315.9	529.6
山东	10550.0	11643	3502.5	3851.0	2288.7	1070.1
河北	6076.6	9047	1406.7	2092.6	1161.6	591.1
天津	2022.6	22068	822.8	960.0	703.7	361.1
江苏	10636.3	14397	3536.3	3967.2	2320.0	1141.1
上海	5408.8	40627	2196.2	2755.8	1970.2	779.1
浙江	7670.0	16570	2356.5	3065.0	2296.6	1180.1
福建	4682.0	13510	1047.1	1859.0	964.5	397.1
广东	11769.7	15030	4224.6	4793.6	3022.9	1221.1

因子分析演示一：

步骤一：变量之间相关系数；作主成分分析，知道保留因子的数量（略）。

步骤二：进行“因子分析”，将参与分析的连续变量放入对应的角色中。

- *选择估计方法。一般使用主成分分析方法。

- *选择合适的因子数量，这需要前期的主成分分析的经验。因子个数的确定标准较宽，比如特征根大于**0.7**就可以考虑保留。

因子分析演示一：

因子旋转之前

因子模式			
		Factor1	Factor2
X1	GDP	0.94970	-0.22248
X2	人均GDP	0.10890	0.98850
X3	工业增加值	0.97780	-0.01032
X4	第三产业增加值	0.98703	-0.04758
X5	固定资产投资	0.98255	0.09983
X6	基本建设投资	0.94550	-0.02414
X7	社会消费品零售总额	0.97853	-0.14180
X8	海关出口总额	0.79901	0.05038
X9	地方财政收入	0.95433	0.19274

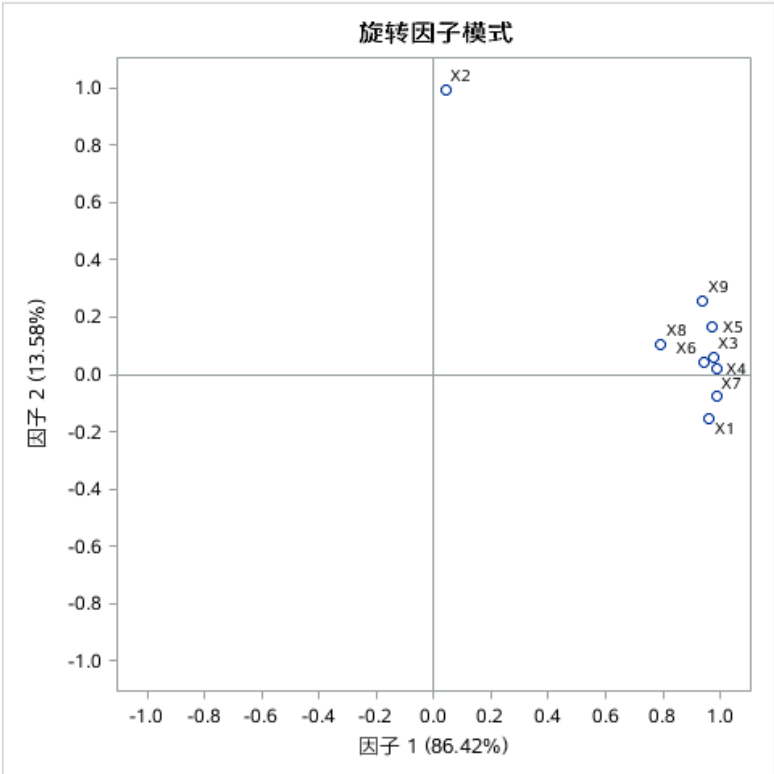
结果分析1:

因子旋转使得原始变量在两个因子上的权重更加两极分化。从右图可以看出，变量被很好的分为两类，也可以尝试着为每个因子其名字：

- 因子1：经济总量水平
- 因子2：人均水平

因子旋转之后

旋转因子模式			
		Factor1	Factor2
X1	GDP	0.96273	-0.15675
X2	人均GDP	0.04077	0.99364
X3	工业增加值	0.97620	0.05683
X4	第三产业增加值	0.98797	0.02030
X5	固定资产投资	0.97338	0.16705
X6	基本建设投资	0.94493	0.04083
X7	社会消费品零售总额	0.98596	-0.07428
X8	海关出口总额	0.79367	0.10512
X9	地方财政收入	0.93884	0.25781






因子分析演示一：

结果分析2：

对样本进行打分。

得到样本
的因子得
分。

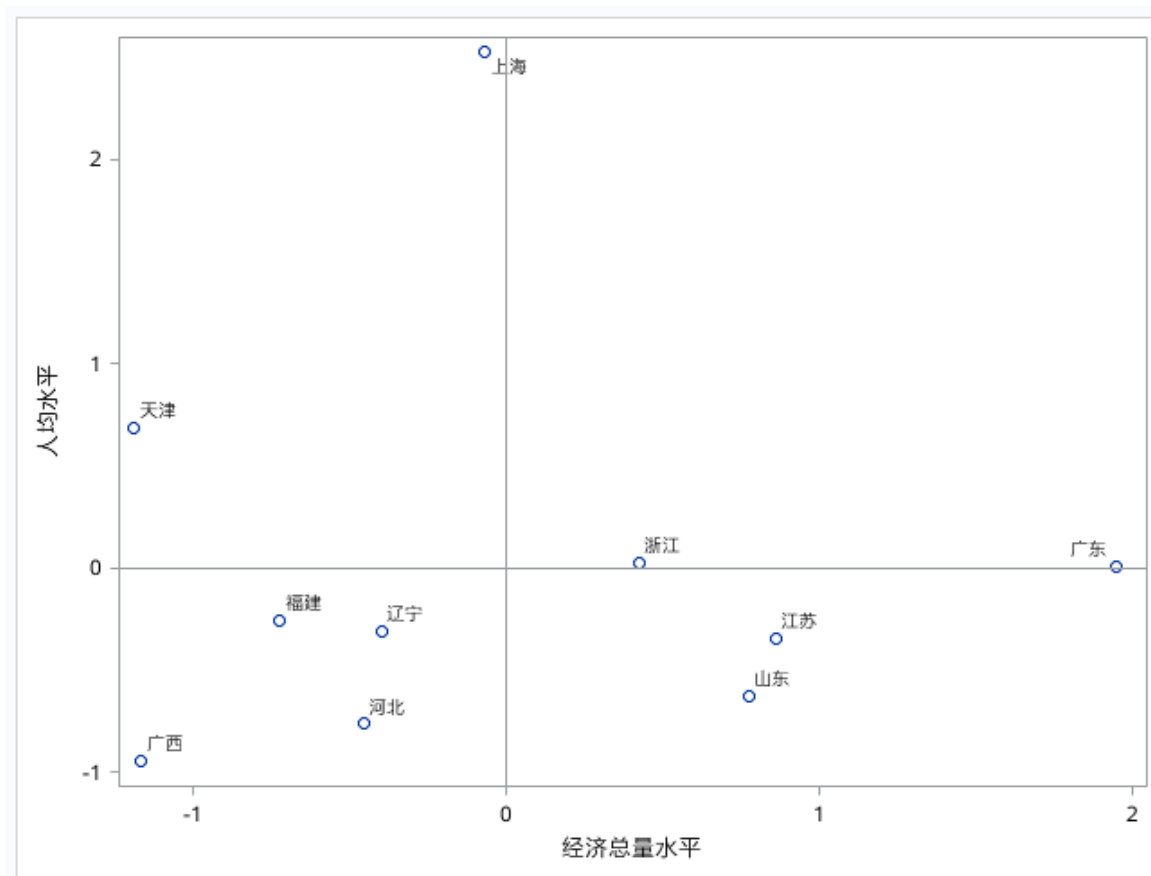
 省份	 Factor1	 Factor2
辽宁	-0.39993162	-0.309512056
山东	0.7740543407	-0.62561796
河北	-0.456630584	-0.763053711
天津	-1.19112364	0.684575144
江苏	0.8626085032	-0.34405354
上海	-0.071276086	2.527887949
浙江	0.426868848	0.0235858902
福建	-0.724931948	-0.256639477
广东	1.9477766663	0.004570315
广西	-1.167414479	-0.941742554

因子分析演示一：

结果分析3：

结果展现

为了在散点图上加上数据标签：



注：如果一个数据的变量可以被压缩为两个因子，则通过展现在二维图形上已经可以完成样本聚类的工作。如果因子多于两个，则需要使用聚类算法进行样本分类。

练习:

- 1、上市公司按行业统计的各项财务比率指标数据存放在表"indu_index "中，尝试着对变量进行因子分析，并给每个因子取名字，如果可以保留两个因子的话，在二维散点图上表述每个行业所处的位置；
- 2、“PROFILE_BANK”记录了银行客户产品使用频数的信息,我们希望使用这个数据作银行客户的客户画像，请分离出数据的维度；
- 3、“PROFILE_TELECOM”记录了电信运营商客户手机业务的使用情况，我们希望使用这个数据作银行客户的客户画像，请分离出数据的维度。

总结:

- 1、因子分析是主成分方法的拓展，可以很好地满足维度分析的需求；
- 2、对于没有业务经验的数据分析人员来讲，是通过观察每个原始变量在因子上的权重绝对值来给因子取名称的。而对于业务知识丰富的数据分析人员，已经对变量的分类有一个预判，并通过进行不同的变量转换（标准化）方式和旋转方式使得预判别为同一组的原始变量在共同的因子上权重绝对值最大化。所以因子分析的要点在于选择变量转换方式。
- 3、因子分析作为维度分析的手段，是构造合理的聚类模型和稳健的分类模型的必然步骤。在这方面，主成分分析、脊回归、**LASSO**算法回归只是在建模时间紧张和缺乏业务经验情况下的替代办法。

稀疏主成分分析

稀疏主成分分析

目标函数：

$$(U^*, V^*) = \arg \min_{U, V} \frac{1}{2} \|X - UV\|_2^2 + \alpha \|V\|_1$$

subject to $\|U_k\|_2 = 1$ for all $0 \leq k < n_{components}$

X-原始数据，N*P

U-主成分，N*R

V-权重矩阵，P*R

Alpha是惩罚系数，取值越小，则越接近普通的主成分分析；取值越大，权重矩阵中的0值越多。

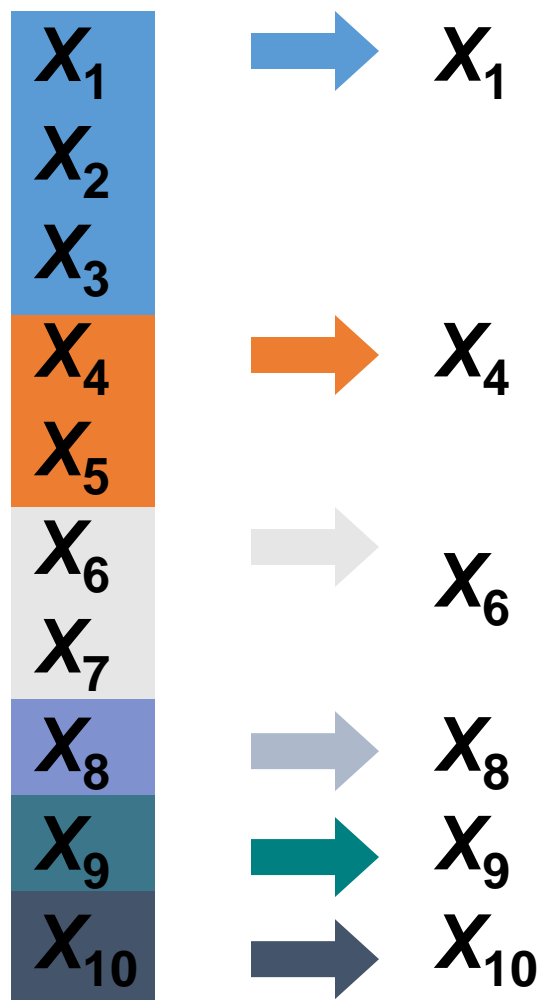
Rodolphe Jenatton, 2009, Structured Sparse Principal Component Analysis

&

<http://scikit-learn.org/stable/modules/decomposition.html#sparse-principal-components-analysis-sparsepca-and-minibatchsparsepca>

变量聚类

变量聚类思路



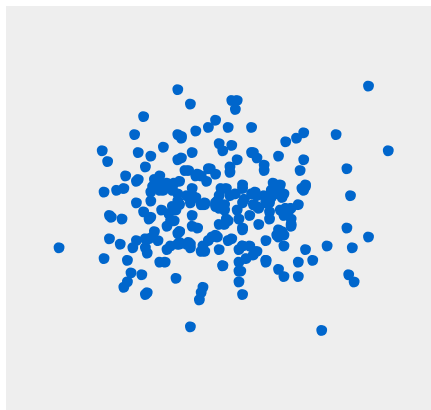
变量聚类之后，选择输入变量：

- 对聚类的代表性
- 专家指定
- 与被解释变量的相关性

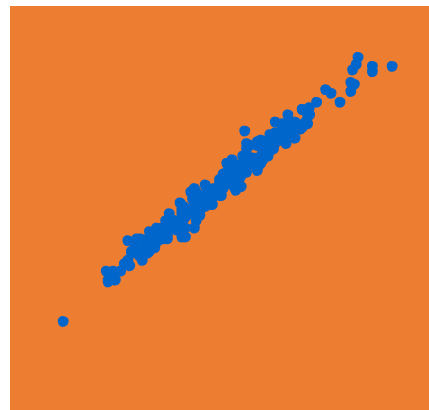
对每组组内的代表性

第一个
主成分

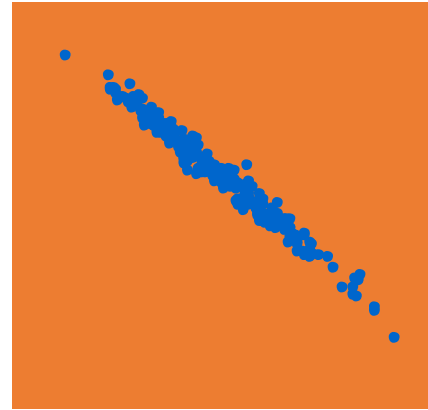
X_1



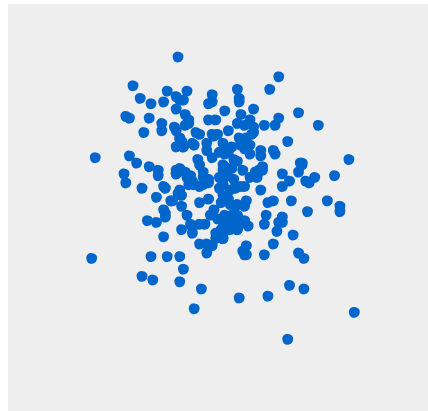
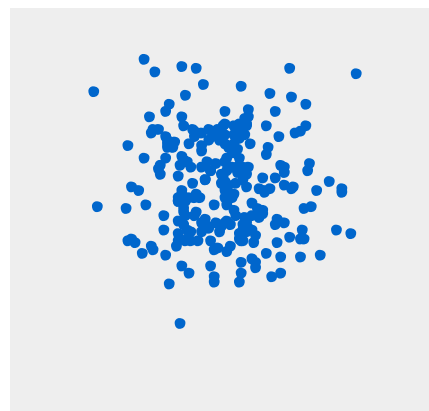
X_2



X_3



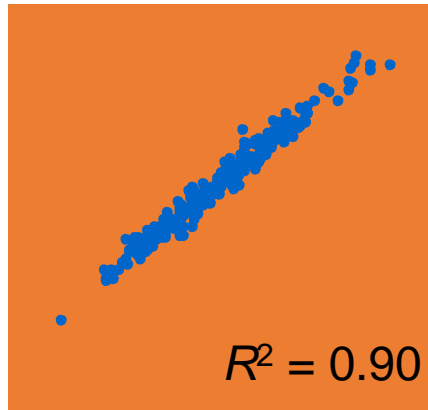
第二个
主成分



对每组组内的代表性

$1 - R^2$ Ratio x_2

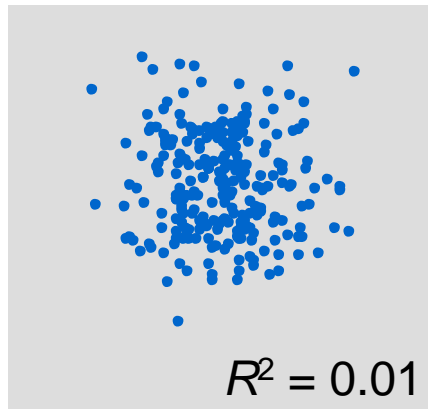
所在类
的主成分



$$1 - R^2 \text{ Ratio} = \frac{1 - R^2_{\text{own-cluster}}}{1 - R^2_{\text{next-closest}}}$$

该值越
小越好: $\frac{1 - \uparrow}{1 - \downarrow} = \frac{\downarrow}{\uparrow} \Rightarrow \downarrow$

其他类
的主成分



$$\frac{1 - R^2_{\text{own cluster}}}{1 - R^2_{\text{next closest}}} = \frac{1 - 0.90}{1 - 0.01} = 0.101$$

借助稀疏主成分分析实现变量聚类

步骤一:进行SparsePCA计算,选择合适的惩罚项 α ,当恰巧每个原始变量只在一个主成分上有权重时,停止循环

指定数量的主成分

原始的
变量

Index	0	1	2
0	0	311.228	0
1	-253.956	0	0
2	-253.955	0	0
3	0	0	311.228

借助稀疏主成分分析实现变量聚类

步骤二、根据上一步得到的惩罚项的取值，估计SparsePCA，并得到稀疏主成分得分

获取指定数量的主成分

0	1	2
0.00235852	-0.00106719	0.000603825
0.000876155	-0.000754573	0.00295041
-0.00247144	0.00168387	0.00764357
0.00207914	-0.000817097	-0.00148202
0.00134069	-0.00150486	-0.00122129
-0.00349909	0.00046465	-0.000960562
-0.00248072	-0.00109846	0.00268967

借助稀疏主成分分析实现变量聚类

步骤三:每个主成分中, 选出原始变量的1-R方比值最小的。

原始的变量

指定数量的主成分			
Index	0	1	2
0	0.330923	0	0.320155
1	0.109806	0.332301	0.313795
2	0.109807	0.331002	0.315073
3	0.305158	0.320155	3.70074e-16

借助稀疏主成分分析实现变量聚类

步骤四:在原始数据中，选取被筛选出的变量。

Index	CNT_ATM	CNT_TBM	CNT_CSC
0	3	34	9
1	17	44	18
2	26	122	36
3	3	42	1
4	15	20	2
5	20	83	3
6	9	33	17
7	5	22	1
8	27	31	1

为回归分析提供变量压缩

•计划使用” CREDITCARD_EXP”数据，通过线形回归构造客户价值预测模型，但是发现解释变量之间具有强相关性。使用这样的数据进行构造的预测模型稳健型差，需要事先进行处理。这里考虑使用主成分分析的方法。



Index	Income	age2	high_avg	dist_home_val	Selfempl
0	16.0352	1600	0.102361	99.93	1
1	15.8475	1024	0.0511842	49.88	0
2	8.4	1296	0.91	88.61	0

秦路主讲

七周成为数据分析师

七周为期，Get一条数据分析师职业黄金通道！



Python

数据分析与挖掘

集Python爬虫、数据采集、数据处理、数据分析与数据挖掘于一体，打造Python全栈工程师

主讲老师：韦玮

VIP会员群+在线答疑+录播复习+1年反复观看

参团课程

案例为师，实战为王

开启Python机器学习之路

科学规划全套课程体系，从入门到进阶，从理论到技巧，嵌入丰富课程案例讲解，逐步推进

讲师：唐宇迪 深度学习领域多年一线实践研究专家

独一无二的数据库建模指南系列教程升级版

- 从企业视角进行数据规划以及数据库模型的搭建
- 高质量的数据库模型和技巧，以及丰富的例子
- 数据库架构理论和实践要领

资深讲师：BAO胖子 15年+BI从业经验
涉足电力、快消品、医药、信息服务行业的BI老兵

业务知识一站通

技术+业务，挣钱有门路！

讲师：陈文



自己动手 丰衣足食

Python3网络爬虫实战案例

一循序渐进，案例为王，诠释全面，思路制胜一

讲师：崔庆才 北航硕士，百万级热度爬文博主



讲师 丘祐玮

人人都爱数据科学家

Python数据科学精华实战课程

数据分析报告制作

秘籍升级版

讲师：陈丹奕 知乎大神，前百度资深数据分析师

先机致胜 破冰AI

深度学习模型/框架与实战

讲师：唐宇迪 同济大学硕士
深度学习领域多年一线实践研究专家



BI、商业智能
数据挖掘 大数据
数据分析师
R语言 Python
机器学习
深度学习
人工智能
Hive Hadoop
Tableau
BIEE ETL
数据科学家
PowerBI