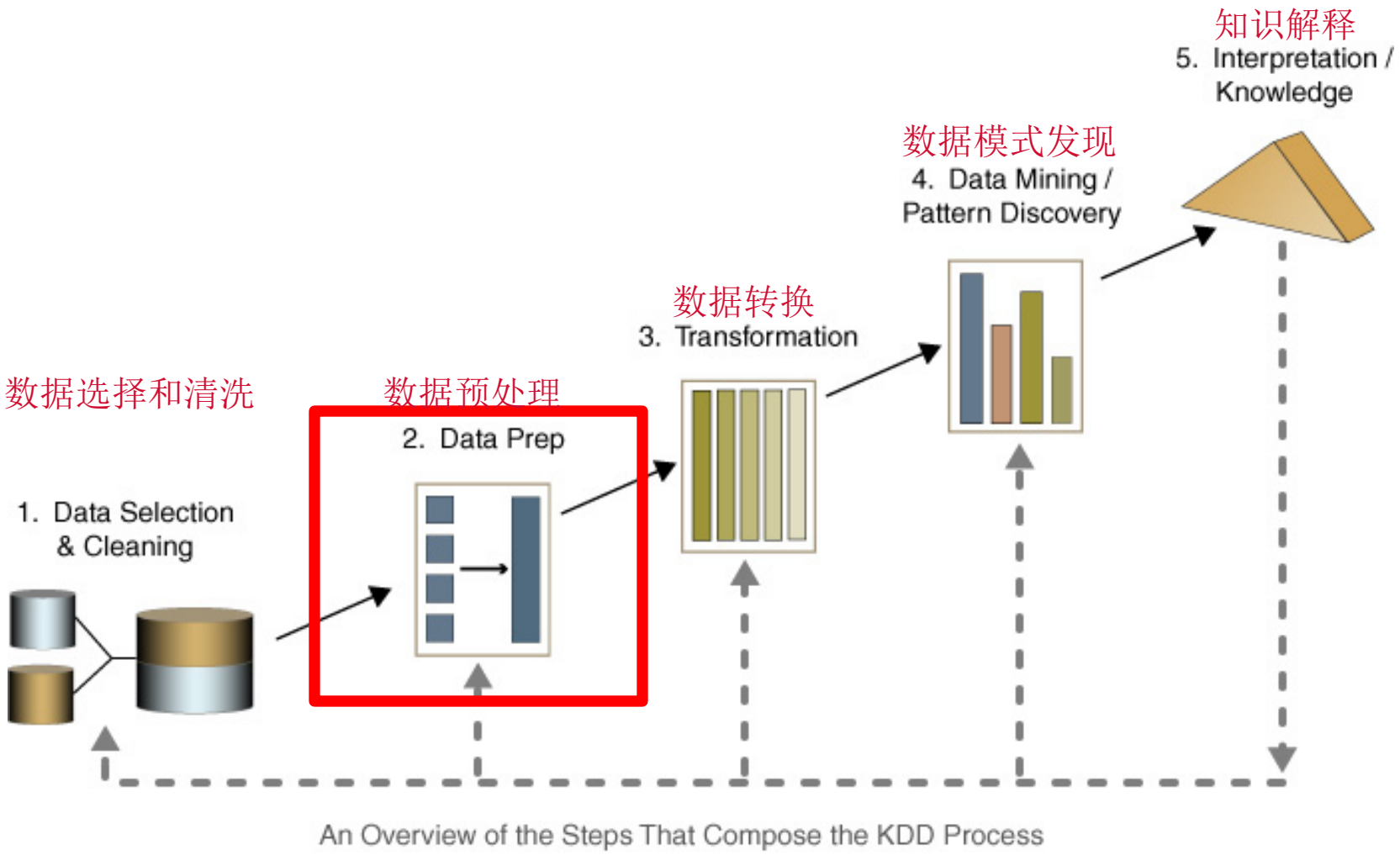
第二章　数据预处理

2014年9月17日

数据挖掘流程



An Overview of the Steps That Compose the KDD Process

第二章 数据预处理

1. Why Data Preprocessing
2. Data Integration
3. Missing Value
4. Noisy Data
5. Data Reduction

2.1 Why Data Preprocessing?

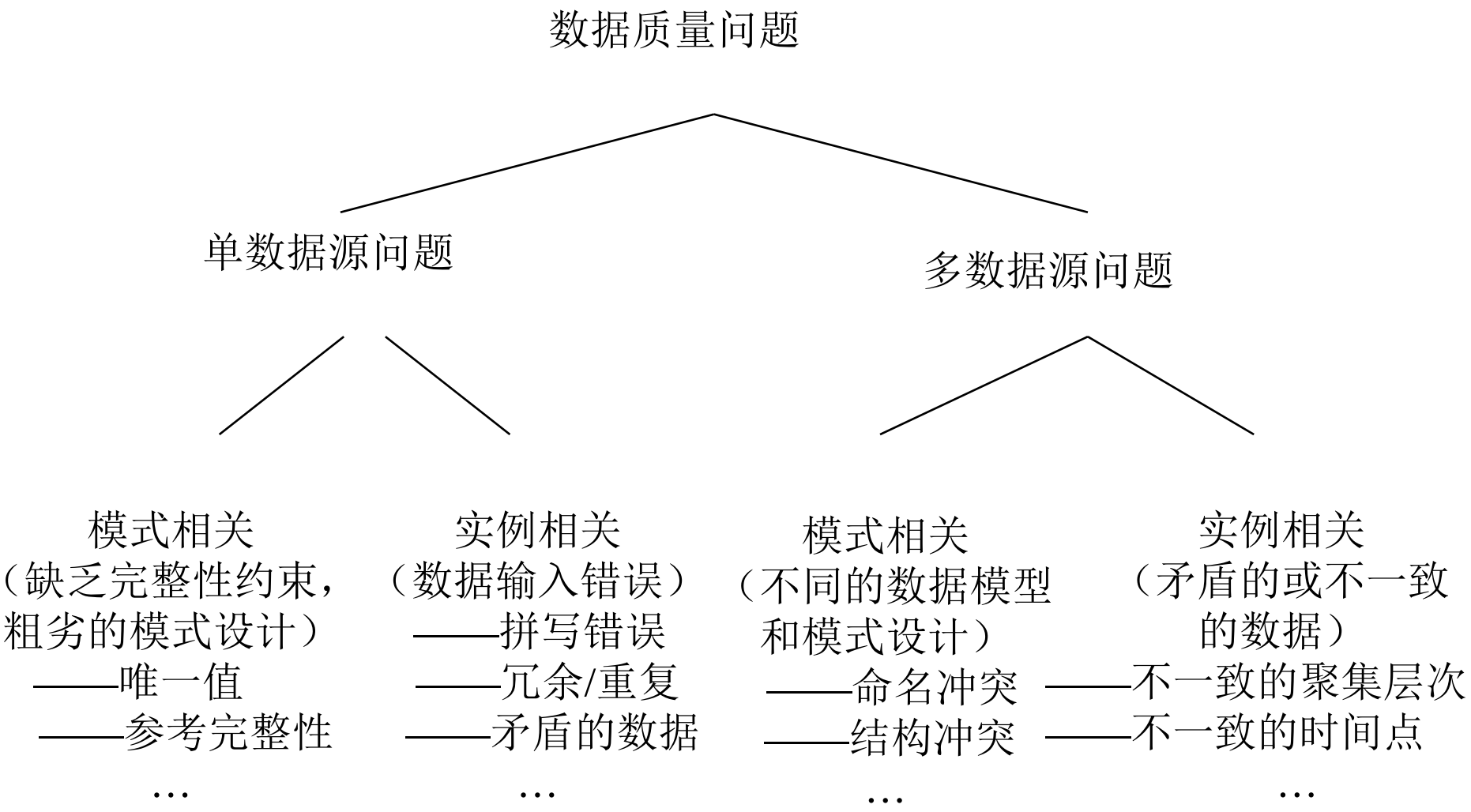
- 高质量决策来自于高质量数据（Quality decisions come from quality data）.
- 数据质量的含义
  - 正确性（Correctness）
  - 一致性（Consistency）
  - 完整性（Completeness）
  - 可靠性（Reliability）

2.1 Why Data Preprocessing?

- 数据错误的不可避免性
  - 数据输入和获得过程数据错误的不可避免性
  - 数据集成所表现出来的错误
  - 数据传输过程所引入的错误
  - 据统计有错误的数据占总数据的5%左右

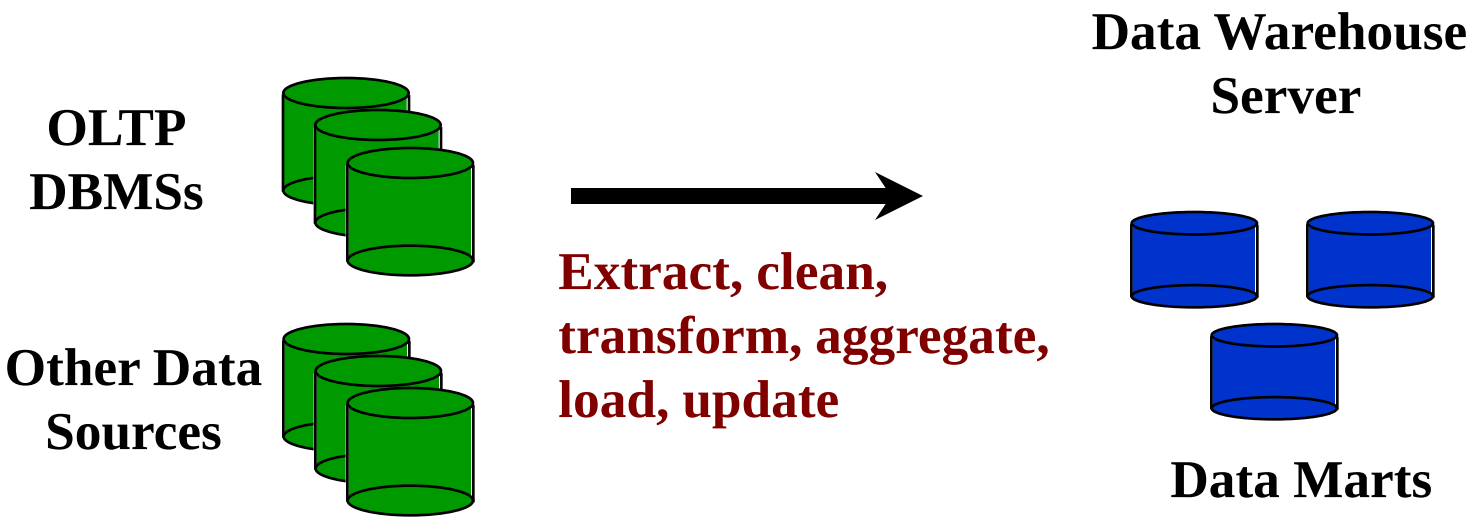
- 其他类型的数据质量问题:
  - Data needs to be integrated from different sources
  - Missing values
  - Noisy and inconsistent values
  - Data is not at the right level of aggregation

数据质量问题的分类


数据质量问题
单数据源问题
多数据源问题
模式相关
（缺乏完整性约束，
粗劣的模式设计）
——唯一值
——参考完整性
…
实例相关
（数据输入错误）
——拼写错误
——冗余/重复
——矛盾的数据
…
模式相关
（不同的数据模型
和模式设计）
——命名冲突
——结构冲突
…
实例相关
（矛盾的或不一致
的数据）
——不一致的聚集层次
——不一致的时间点
…

2.2 Data Integration

- Integrate data from multiple sources into a common format for data mining.
- Note: A good data warehouse has already taken care of this step.



Data Integration (Contd.)

Problem: Heterogeneous schema integration

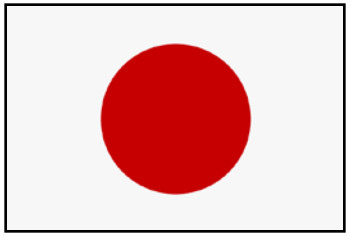
- Different attribute names

cid	name	byear
1	Jones	1960
2	Smith	1974
3	Smith	1950

Customer-ID	state
1	NY
2	CA
3	NY

- Different units: Sales in $, sales in Yen, sales in DM





Data Integration (Contd.)

Problem: Heterogeneous schema integration

- Different scales: Sales in dollars versus sales in pennies

- Derived attributes: Annual salary versus monthly salary

cid	monthlySalary
1	5000
2	2400
3	3000

cid	Salary
6	50,000
7	100,000
8	40,000

Data Integration (Contd.)

Problem: Inconsistency due to redundancy

- Customer with customer-id 150 has three children in relation1 and four children in relation2

cid	numChildren
1	3

cid	numChildren
1	4

- Computation of annual salary from monthly salary in relation1 does not match “annual-salary” attribute in relation2

cid	monthlySalary
1	5000
2	6000

cid	Salary
1	60,000
2	80,000

2.3 Missing Values

常常会有一些记录的某些属性值不知道的情况出现

.出现缺失值的原因:

- **Attribute does not apply (e.g., maiden娘家姓 name)**
- **Inconsistency with other recorded data**
- **Equipment malfunction**
- **Human errors**
- **Attribute introduced recently (e.g., email address)**

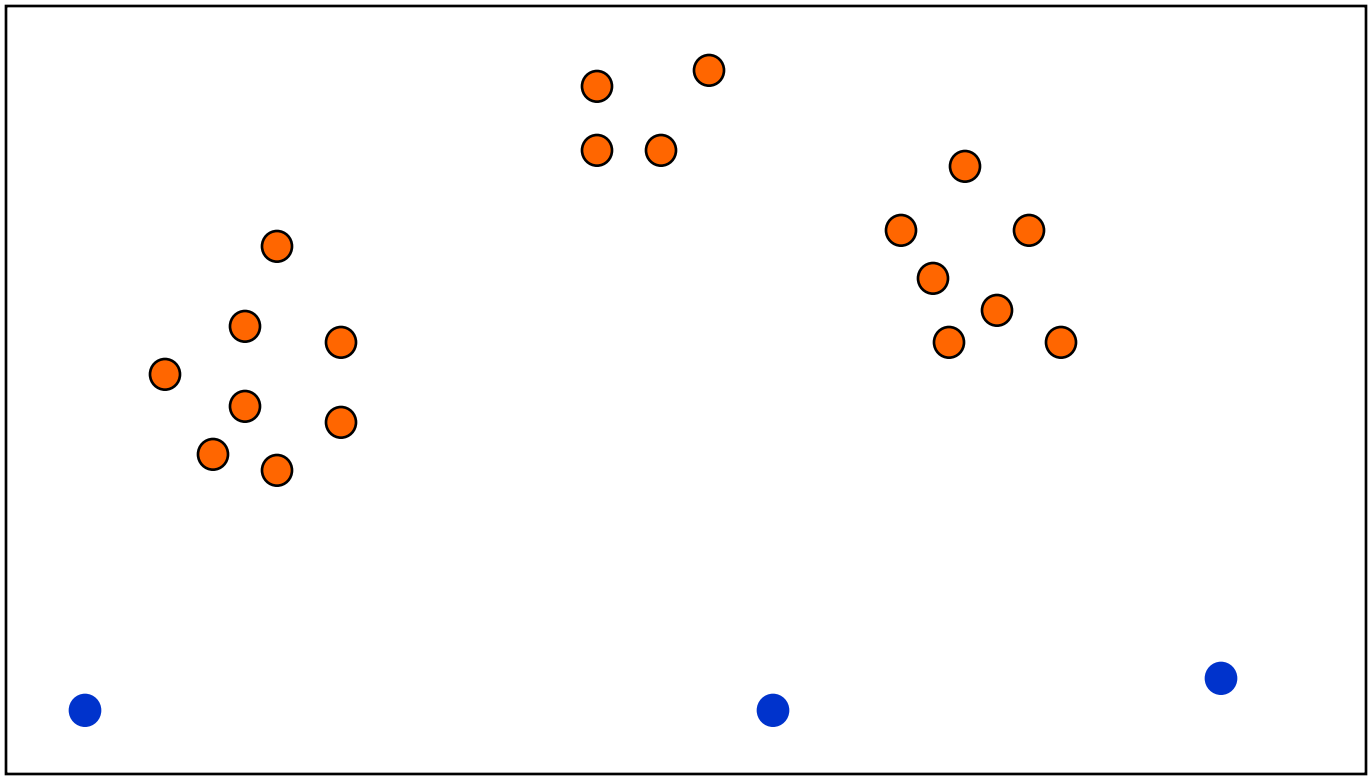
Missing Values: Approaches

- Ignore the record
- Complete the missing value:
  - Manual completion: Tedious and likely to be infeasible
  - Fill in a global constant, e.g., “NULL”, “unknown”
  - Use the attribute mean
  - Construct a data mining model that predicts the missing value

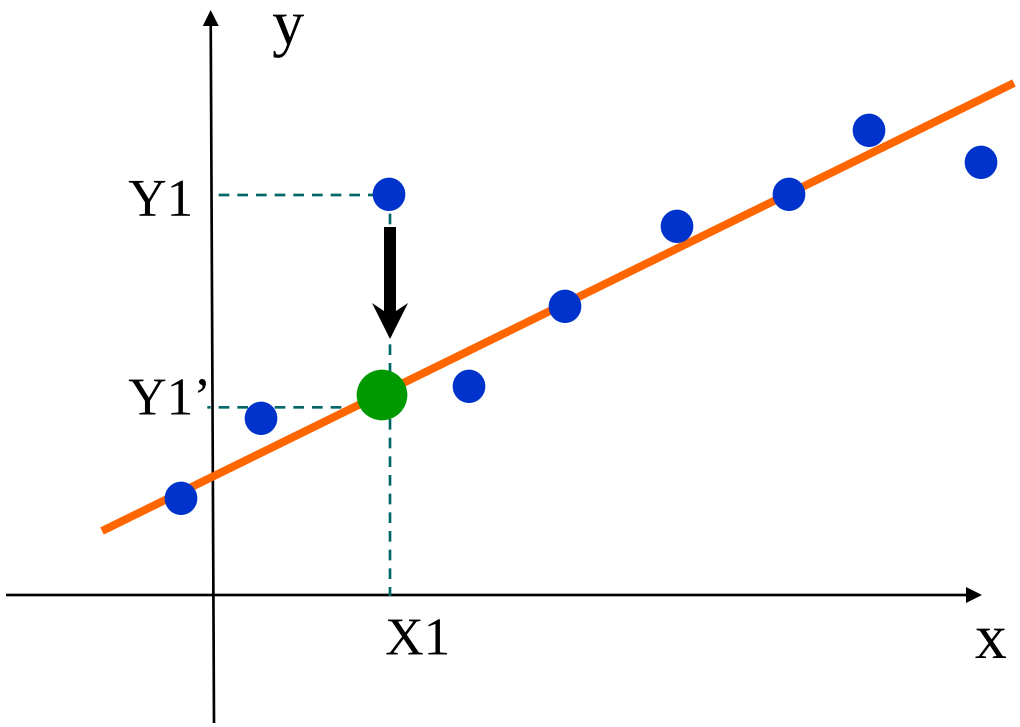
2.3 Noisy Data

- Examples:
  - Faulty data collection instruments
  - Data entry problems, misspellings
  - Data transmission problems
  - Technology limitation
  - Inconsistency in naming conventions（命名习惯）
  - Duplicate records with different values for a common field

Noisy Data: Remove Outliers



Noisy Data: Smoothing



Noisy Data: Normalization

- Scale data to fall within a small, specified range
  - Leave out extreme order statistics
  - Min-max normalization
  - Z-score normalization
  - Normalization by decimal scaling

2.4 Data Reduction

Problem:

- Data might not be at the right scale for analysis.
  Example: Individual phone calls versus monthly phone call usage
- Complex data mining tasks might run a very long time.
  Example: Multi-terabyte data warehouses
  One disk drive: About 20MB/s

Data Reduction: Attribute Selection

- Select the “relevant” attributes for the data mining task
- If there are k attributes, there are 2^k-1 different subsets
  - Example: {salary, children, wyear}, {salary, children}, {salary, wyear}, {children, wyear}, {salary}, {children}, {wyear}
- Choice of the right subset depends on:
  - Data mining task
  - Underlying probability distribution

Attribute Selection (Contd.)

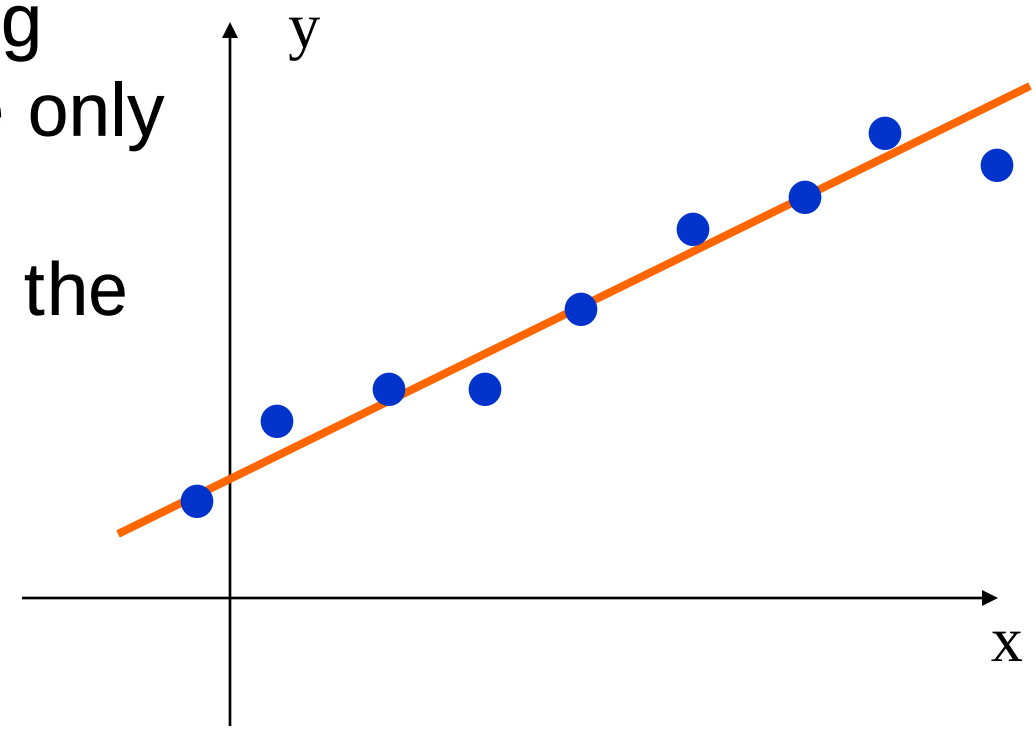
- How to choose relevant attributes:
  - Forward selection: Select greedily one attribute at a time
  - Backward elimination: Start with all attributes, eliminate irrelevant attributes
  - Combination of forward selection and backward elimination

Data Reduction: Parametric Models

- Main idea:
  - Fit a parametric model to the data (e.g., multivariate normal distribution)
  - Store the model parameters, discard the data (except for outliers)

Parametric Models: Example

- Instead of storing (x,y) pairs, store only the x-value.
  Then recompute the y-value using
  $y = ax + b$



Data Reduction: Sampling

- Choose a representative subset of the data

- Simple random sampling may have very poor performance in the presence of skew

Data Reduction: Sampling (Contd.)

Stratified sampling(分层取样): Biased sampling

- Example: Keep population group ratios

- Example: Keep minority population group count

Data Reduction: Histograms(柱状图)

- Divide data into buckets and store average (sum) for each bucket
- Can be constructed “optimally” for one attribute using dynamic programming
- Example:
  Dataset: 1,1,1,1,1,1,1,1,2,2,2,2,3,3,4,4,5,5,6,6,
  7,7,8,8,9,9,10,10,11,11,12,12
  Histogram: (range, count, sum)
  (1-2,12,16), (3-6,8,36), (7-9,6,48), (10-12,6,66)

Histograms (Contd.)

- Equal-width histogram
  - Divides the domain of an attribute into k intervals of equal size
  - Interval width = (Max – Min)/k
  - Computationally easy
  - Problems with data skew and outliers
- Example:
  - Dataset: 1,1,1,1,1,1,1,1,2,2,2,2,3,3,4,4,5,5,6,6, 7,7,8,8,9,9,10,10,11,11,12,12
  - Histogram: (range, count, sum) (1-3,14,22), (4-6,6,30), (7-9,6,48), (10-12,6,66)

Histograms (Contd.)

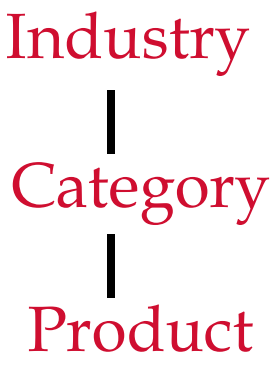
- Equal-depth histogram
  - Divides the domain of an attribute into k intervals, each containing the same number of records
  - Variable interval width
  - Computationally easy
- Example:
  - Dataset: 1,1,1,1,1,1,1,1,2,2,2,2,3,3,4,4,5,5,6,6, 7,7,8,8,9,9,10,10,11,11,12,12
  - Histogram: (range, count, sum)
    (1,8,8), (2-4,8,22), (5-8,8,52), (9-12,8,84)

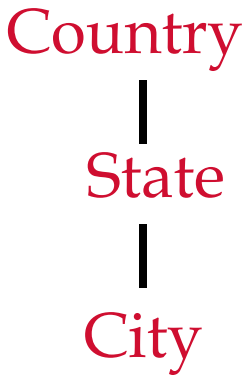
Data Reduction: Discretization

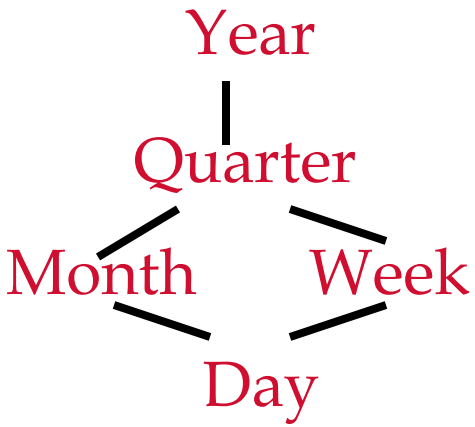
- Same concept as histograms
- Divide domain of a numerical attribute into intervals.
- Replace attribute value with label for interval.
- Example:
  - Dataset (age; salary):
    (25;30,000),(30;80,000),(27;50,000),
    (60;70,000),(50;55,000),(28;25,000)
  - Discretized dataset (age, discretizedSalary):
    (25,low),(30,high),(27,medium),(60,high),
    (50,medium),(28,low)

Data Reduction: Natural Hierarchies

- Natural hierarchies on attributes can be used to aggregate data along the hierarchy.Replace low-level concepts with high-level concepts
- Example replacements:
  - Product Name by category
  - City by state







Aggregation: Example

cid	Date	Amount
1	3/1/2000	$150
1	3/5/2000	$50
1	3/29/2000	$200
1	4/2/2000	$300
1	4/6/2000	$200
2	3/2/2000	$100
2	3/5/2000	$250
2	3/10/2000	$100
2	3/11/2000	$50
2	4/7/2000	$200
3	4/2/2000	$300
4	3/17/2000	$250
4	4/25/2000	$100

cid	Month	Year	Amount	Visits
1	3	2000	$400	3
1	4	2000	$500	2
2	3	2000	$500	4
2	4	2000	$200	1
3	4	2000	$300	1
4	3	2000	$250	1
4	4	2000	$100	1

Data Reduction: Other Methods

- Principal component analysis
- Fourier transformation
- Wavelet transformation

主成分分析(Principal Component Analysis)

- 在许多领域的研究与应用中，往往需要对反映事物的多个变量进行大量的观测，收集大量数据以便进行分析寻找规律。

- 多变量大样本无疑会为研究和应用提供了丰富的信息，但也在一定程度上增加了数据采集的工作量，更重要的是在大多数情况下，许多变量之间可能存在相关性，从而增加了问题分析的复杂性，同时对分析带来不便。

- 如果分别对每个指标进行分析，分析往往是孤立的，而不是综合的。盲目减少指标会损失很多信息，容易产生错误的结论。

主成分分析

- 因此需要找到一个合理的方法，在减少需要分析的指标同时，尽量减少原指标包含信息的损失，以达到对所收集数据进行全面分析的目的。

- 由于各变量间存在一定的相关关系，因此有可能用较少的综合指标分别综合存在于各变量中的各类信息。主成分分析与因子分析就属于这类降维的方法。

- 主成分分析与因子分析是将多个实测变量转换为少数几个不相关的综合指标的多元统计分析方法。

主成分分析

- 例1：生产服装有很多指标，比如袖长、肩宽、身高等十几个指标，服装厂生产时，不可能按照这么多指标来做，怎么办？一般情况，生产者考虑几个综合的指标，象标准体形、特形等。

- 例2：企业经济效益的评价，它涉及到很多指标。例百元固定资产原值实现产值、百元固定资产原值实现利税，百元资金实现利税，百元工业总产值实现利税，百元销售收入实现利税，每吨标准煤实现工业产值，每千瓦时电力实现工业产值，全员劳动生产率，百元流动资金实现产值等，我们要找出综合指标，来评价企业的效益。

主成分分析

- 例3：假定你是一个公司的财务经理，掌握了公司的所有数据，比如固定资产、流动资金、每一笔借贷的数额和期限、各种税费、工资支出、原料消耗、产值、利润、折旧、职工人数、职工的分工和教育程度等等。
- 如果让你向上司介绍公司状况，你能原封不动地介绍所有指标和数字吗？
- 当然不能。
- 你必须要对各个方面作出高度概括，用一两个指标简单明了地说清楚情况。

主成分分析

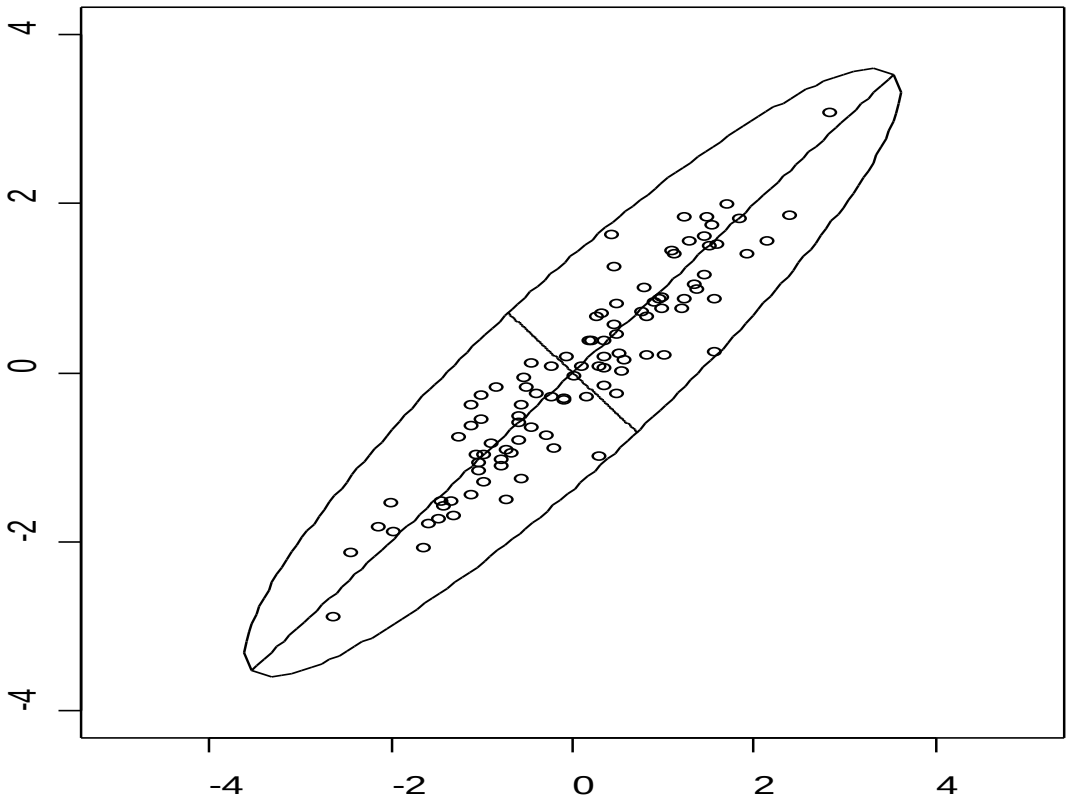
- 很多应用都会遇到有很多变量的数据，这些数据的共同特点是变量很多，在如此多的变量之中，有很多是相关的。
- 因此我们希望从中综合出一些少数主要的指标，这些指标所包含的信息量又很多。这些特点，使我们在研究复杂的问题时，容易抓住主要矛盾。
- 如何才能找出综合指标？

主成分分析

- 由于实测的变量间存在一定的相关关系，因此有可能用较少数的综合指标分别综合存在于各变量中的各类信息，而综合指标之间彼此不相关，即各指标代表的信息不重叠。综合指标称为主成分（提取几个主成分）。

- 若有一些指标 $X_1,\cdots,X_P$，取综合指标即它们的线性组合**F**，当然有很多，我们希望线性组合**F**包含很多的信息，即**var(F)**最大，这样得到**F**记为 F_1，然后再找 F_2，F_1与F_2无关，以此类推，我们找到了一组综合变量 $F_1,F_2,\ldots,F_m$，这组变量基本包含了原来变量的所有信息。

主成分分析



主成分分析

- 在数据挖掘领域，以及图像处理、通讯技术等信息处理领域，主成分分析是很常用的一种方法
- 通过对一组变量的几种线性组合来解释这组变量的方差和协方差结构，以达到数据的降维和数据的解释的目的。

成绩数据

- 学生的数学、物理、化学、语文、历史、英语的成绩如下表（部分）。

学生代码	数学	物理	化学	语文	历史	英语
1	65	61	72	84	81	79
2	77	77	76	64	70	55
3	67	63	49	65	67	57
4	80	69	75	74	74	63
5	74	70	80	84	81	74
6	78	84	75	62	71	64
7	66	71	67	52	65	57
8	77	71	57	72	86	71
9	83	100	79	41	67	50
…	…	…	…	…	…	…	…

问　　题

- 目前的问题是，能否将该数据的6个变量用一两个综合变量来表示？
- 这一两个综合变量包含了多少原来的信息？
- 能不能利用找到的综合变量来对学生排序？这一类数据所涉及的问题可以推广到对其他应用的分析、排序、判别和分类等问题。

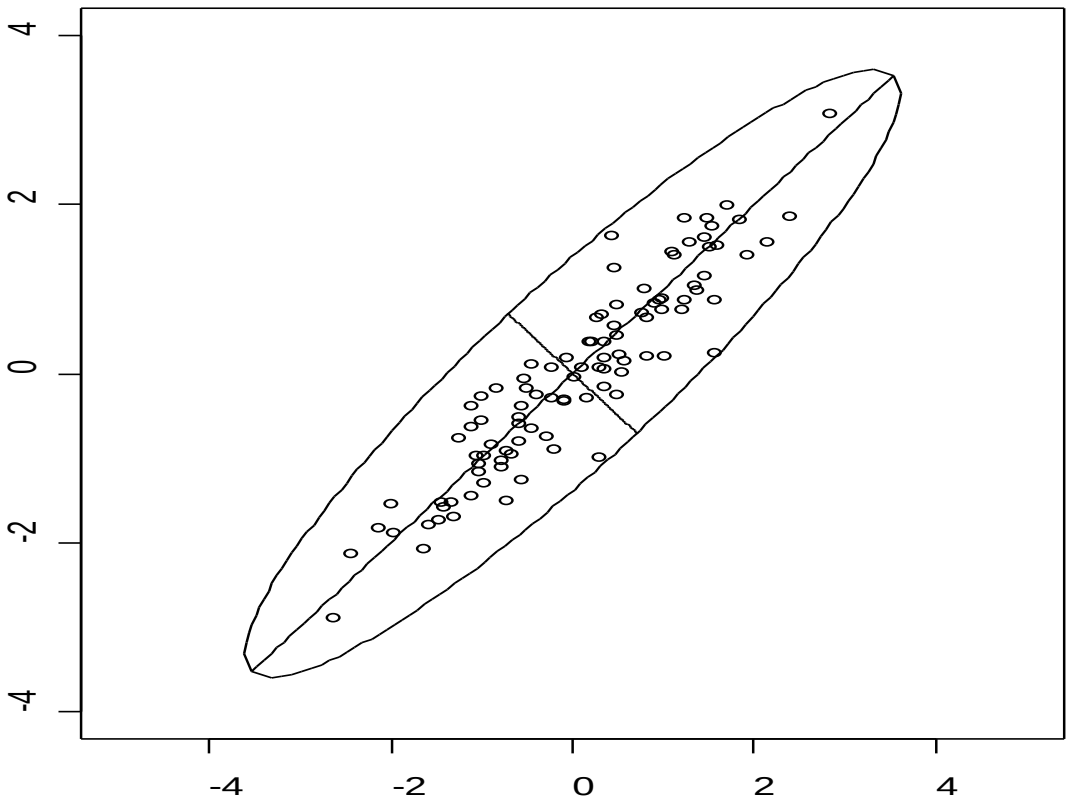
主成分分析

- 例中的数据点是六维的，即每个观测值是6维空间中的一个点。我们希望将6维空间用低维空间表示。

- 先假定只有二维，即只有两个变量，它们由横坐标和纵坐标所代表；因此每个观测值都有相应于这两个坐标轴的两个坐标值；如果这些数据形成一个椭圆形状的点阵，那么这个椭圆有一个长轴和一个短轴。在短轴方向上，数据变化很少；在极端的情况，短轴如果退化成一点，那只有在长轴的方向才能够解释这些点的变化了；这样，由二维到一维的降维就自然完成了。

主成分分析

- 当坐标轴和椭圆的长短轴平行，那么代表长轴的变量就描述了数据的主要变化，而代表短轴的变量就描述了数据的次要变化。
- 但是，坐标轴通常并不和椭圆的长短轴平行。因此，需要寻找椭圆的长短轴，并进行变换，使得新变量和椭圆的长短轴平行。
- 如果长轴变量代表了数据包含的大部分信息，就用该变量代替原先的两个变量（舍去次要的一维），降维就完成了。
- 椭圆（球）的长短轴相差得越大，降维也越有道理。


4
2
0
-2
-4
-4
-2
0
2
4

主成分分析

- 对于多维变量的情况和二维类似，也有高维的椭球，只不过无法直观地看见罢了。
- 首先找出高维椭球的主轴，再用代表大部分数据信息的最长几个轴作为新变量；这样，就基本完成了主成分分析。
- 注意，和二维情况类似，高维椭球的主轴也是互相垂直的。这些互相正交的新变量是原先变量的线性组合，称为主成分**(principal component)**。

主成分分析

- 正如二维椭圆有两个主轴，三维椭球有三个主轴一样，有几个变量，就有几个主成分。

- 选择越少的主成分，降维幅度就越大。什么是降维标准呢？那就是这些被选的主成分所代表的主轴的长度之和占主轴长度总和的大部分。有些研究工作表明，所选的主轴总长度占所有主轴长度之和的大约**85%**即可，其实，这只是一个大体的说法；具体选多少个，要看实际情况而定。

• 对于上面的数据，主成分分析的结果

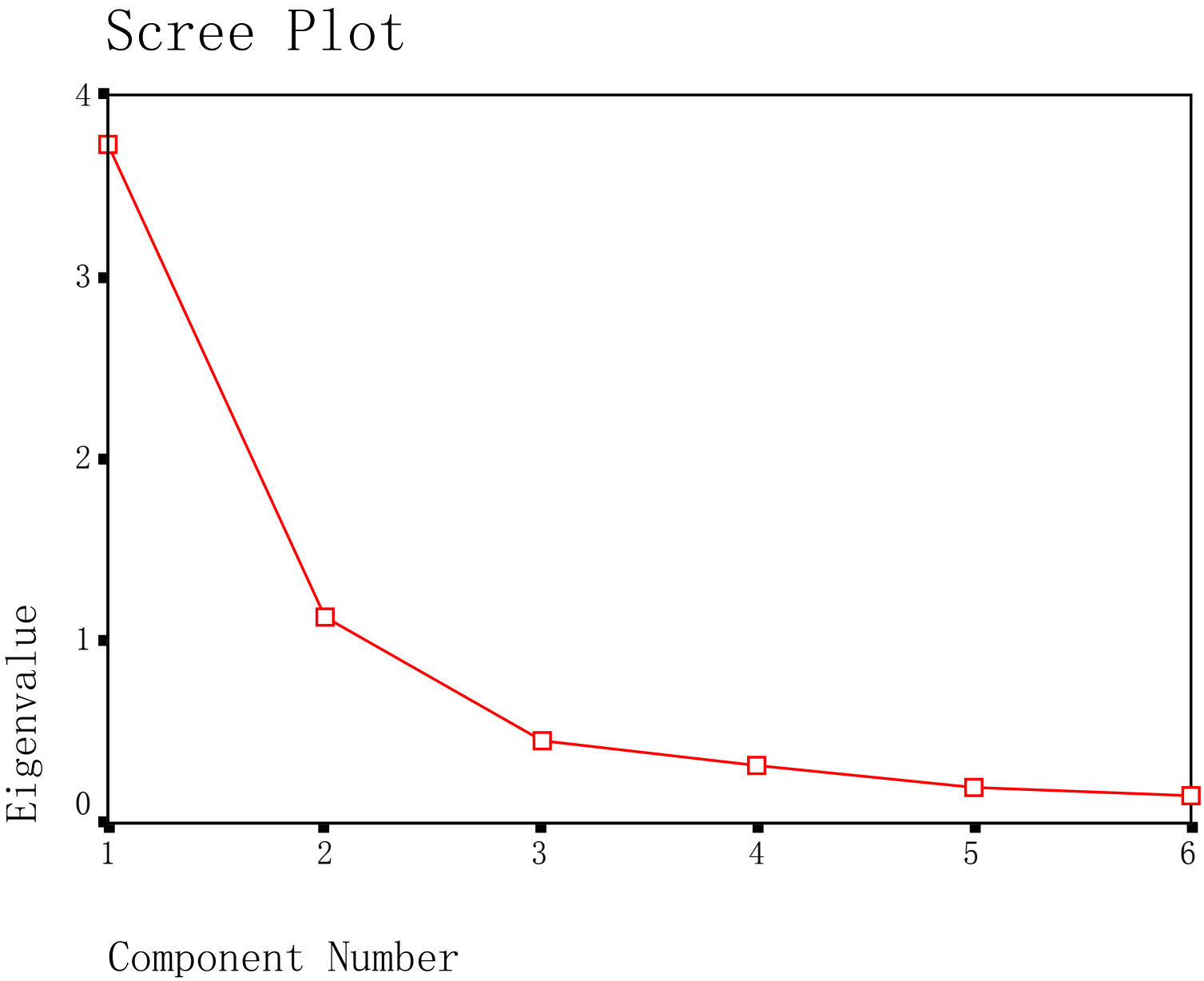
Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.735	62.254	62.254	3.735	62.254	62.254
2	1.133	18.887	81.142	1.133	18.887	81.142
3	.457	7.619	88.761			
4	.323	5.376	94.137			
5	.199	3.320	97.457			
6	.153	2.543	100.000			

Extraction Method: Principal Component Analysis.

- 这里的**Initial Eigenvalues**就是这里的六个主轴长度，又称特征值（数据相关阵的特征值）。头两个成分特征值累积占了总方差的**81.142%**。后面的特征值的贡献越来越少。

- 特征值的贡献还可以从碎石图看出



- 怎么解释这两个主成分。前面说过主成分是原始六个变量的线性组合。是怎么样的组合呢？具体结果见下表。

Component Matrix[a]

	Component					
	1	2	3	4	5	6
MATH	-.806	.353	-.040	.468	.021	.068
PHYS	-.674	.531	-.454	-.240	-.001	-.006
CHEM	-.675	.513	.499	-.181	.002	.003
LITERAT	.893	.306	-.004	-.037	.077	.320
HISTORY	.825	.435	.002	.079	-.342	-.083
ENGLISH	.836	.425	.000	.074	.276	-.197

Extraction Method: Principal Component Analysis.

a. 6 components extracted.

- 这里每一列代表一个主成分作为原来变量线性组合的系数（比例）。比如第一主成分作为数学、物理、化学、语文、历史、英语这六个原先变量的线性组合，系数（比例）为**-0.806, -0.674, -0.675, 0.893, 0.825, 0.836**。

主成分分析

- 如用$x_1, x_2, x_3, x_4, x_5, x_6$分别表示原先的六个变量，而用$y_1, y_2, y_3, y_4, y_5, y_6$表示新的主成分，那么，第一和第二主成分$y_1, y_2$同原来6个变量$x_1, x_2, x_3, x_4, x_5, x_6$与的关系为：

$$y_1 = -0.806x_1 - 0.674x_2 - 0.675x_3 + 0.893x_4 + 0.825x_5 + 0.836x_6$$
$$y_2 = 0.353x_1 + 0.531x_2 + 0.513x_3 + 0.306x_4 + 0.435x_5 + 0.425x_6$$

- 这些系数称为主成分载荷（**loading**），它表示主成分和相应的原先变量的相关系数。

- 比如y_1表示式中x_1的系数为**-0.806**，这就是说第一主成分和数学变量的相关系数为**-0.806**。

- 相关系数(绝对值）越大，主成分对该变量的代表性也越大。可以看出，第一主成分对各个变量解释得都很充分。而最后的几个主成分和原先的变量就不那么相关了。

主成分分析的一些注意事项

- 主成分分析从原理上是寻找椭球的所有主轴。因此，原先有几个变量，就有几个主成分。
- 可以看出，主成分分析都依赖于原始变量，也只能反映原始变量的信息。所以原始变量的选择很重要。
- 另外，如果原始变量都本质上独立，那么降维就可能失败，这是因为很难把很多独立变量用少数综合的变量概括。数据越相关，降维效果就越好。
- 在得到分析的结果时，并不一定会都得到如我们例子那样清楚的结果。这与问题的性质，选取的原始变量以及数据的质量等都有关系

数学模型

- 设有n个数据样本，每个样本观测p个变量（指标）：X_1，X_2，…，X_P，得到原始数据矩阵：

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = (X_1, X_2, \ldots, X_P)$$

其中，$x_i = \begin{bmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{ni} \end{bmatrix}$，$i = 1, \ldots, p$

数学模型

- 用数据矩阵X的p个向量（即p个指标向量）X1，X2，…，XP，作线性组合（即综合指标向量）为：

- 简写成

$$\begin{cases} F_1 = a_{11}X_1 + a_{21}X_2 + \ldots + a_{P1}X_P \\ F_2 = a_{12}X_1 + a_{22}X_2 + \ldots + a_{P2}X_P \\ \vdots \\ F_P = a_{1P}X_1 + a_{2P}X_2 + \ldots + a_{PP}X_P \end{cases}$$

- 上述方程组要求：$F_i = a_{1i}X_1 + a_{2i}X_2 + \ldots + a_{Pi}X_P$， $i = 1,\ldots, p$

$$a_{1i}^2 + a_{2i}^2 + \ldots + a_{pi}^2 = 1$$

数学模型

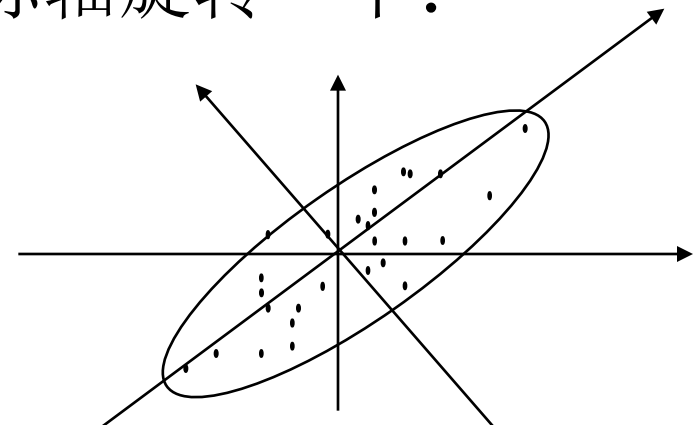
- 且系数a_{ij}由下列原则决定：
  - F_i与F_j（$i \neq j$，i，$j=1$，…，p）不相关；
  - F_1 是X_1，X_2，…，X_P，的一切线性组合（系数满足上述方程组）中方差最大的，F_2 是与 F_1 不相关的X_1，X_2，…，X_P，一切线性组合中方差最大的，……，F_p 是与F_1，F_2，…，F_{P-1}，都不相关的X_1，X_2，…，X_P，一切线性组合中方差最大的。

- 从代数学观点看**主成分就是p个变量X_1，X_2，…，X_P的一些特殊的线性组合**，而在几何上这些线性组合正是把X_1，X_2，…，X_P构成坐标系旋转产生的新坐标系，新坐标轴通过样本方差最大的方向（或说具有最大的样本方差）。

主成分的几何意义

- 设有n个样品，每个样品有两个观测变量 x_1, x_2，二维平面的散点图。n个样本点，无论沿着 x_1 轴方向还是 x_2 轴方向，都有较大的离散性，其离散程度可以用 x_1 或 x_2 的方差表示。

- 当只考虑一个时，原始数据中的信息将会有较大的损失。若将坐标轴旋转一下：



主元分析示意图

数学模型

- 若在椭圆长轴方向取坐标轴F1，在短轴方向取F2，这相当于在平面上作一个坐标变换，即按逆时针方向旋转θ角度，根据旋轴变换公式新老坐标之间有关系：

$$\begin{cases} F_1 = X_1\cos\theta + X_2\sin\theta \\ F_2 = -X_1\sin\theta + X_2\cos\theta \end{cases}$$

- F1，F2 是原变量X1 和X2 的线性组合，用矩阵表示是

$$\begin{pmatrix} F_1 \\ F_2 \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} = U \cdot X$$

- 显然 $U' = U^{-1}$且是正交矩阵，即$U'U = I$

- 则n个样品在F_1 轴的离散程度最大（方差最大），变量 F_1 代表了原始数据的绝大部分信息，即使不考虑F_2，信息损失也不多。而且，F_1F_2不相关。只考虑 F_1 时，二维降为一维。

主成分的推导

- 设$F = a_1X_1 + a_2X_2 + \ldots + a_pX_p = a'X$,其中 $a=(a_1, a_2, \ldots, a_p)'$, $X=(X_1, X_2, \ldots, X_p)'$, 求主成分就是寻求$X$的线性函数$a'X$ 使相应的方差尽可能地大，即

$$\begin{aligned} Var(a,X) &= E\,(a'X - E\,(a'X))\,(a'X - E\,(a'X))' \\ &= a'\,E\,(X - E\,(X))\,(X - E\,(X))'a \\ &= a'\Sigma a \end{aligned}$$

达到最大值，且$a'a=1$。

主成分的推导

- 设协差阵$\sum$的特征根为$\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p > 0$，相应的单位特征向量为$u_1, \ldots, u_p$，
- 令$U = (u_1, \ldots, u_p) =$ $\begin{bmatrix} u_{11} & u_{12} & \ldots & u_{1p} \\ u_{21} & u_{22} & \ldots & u_{2p} \\ \vdots & \vdots & & \vdots \\ u_{n1} & u_{n2} & \ldots & u_{np} \end{bmatrix}$
- 则$UU' = U'U = I$，且 $\sum = U \begin{bmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_p \end{bmatrix} U' = \sum_{i=1}^{p} \lambda_i u_i u_i'$

主成分的推导

- 因此

$$a'\Sigma a = \sum_{i=1}^{p}\lambda_i a'u_i u_i' a = \sum_{i=1}^{p}\lambda_i (a'u_i)(a'u_i)' = \sum_{i=1}^{p}\lambda_i (a'u_i)^2$$

- 所以

$$a'\Sigma a \le \lambda_1 \sum_{i=1}^{p}(a'u_i)^2 = \lambda_1 (a'U)(a'U)' = \lambda_1 a'UU'a$$

$$= \lambda_1 a'a = \lambda_1$$

- 而且，当$a=u_1$时有

$$u_1'\Sigma u_1 = u_1'(\sum_{i=1}^{p}\lambda_i u_i u_i')u_1 = \sum_{i=1}^{p}\lambda_i u_1' u_i u_i' u_1 = \lambda_1 (u_1'u_1)^2 = \lambda_1$$

- 因此$a= u_1$ 使 $Var(a',X)= a'\Sigma a$达到最大值，且$Var(u_1',X)= u_1'\Sigma u_1=\lambda_1$。

主成分的推导

- 同理Var(u_i', X)=λ_i, 而且Cov(u_i'X, u_j'X) = u_i'Σu_j = u_i'$(\sum_{i=1}^{p}\lambda_k u_k u_k')u_j$ = $\sum_{k=1}^{p}\lambda_k(u_i u_k')(u_k' u_j)=0$ ， $i \neq j$

- 上述推导表明：X_1，X_2，…，X_P 的主要成分就是以Σ的特征向量为系数的线性组合，它们互不相关，其方差为Σ的特征根。

- 由于Σ的特征根$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$,所以Var($F_1$) ≥Var($F_2$) ≥…≥Var($F_p$) >0。主要成分的名次是按特征根取值大小的顺序排列的。

主成分的推导

- 在解决实际问题时，一般不是取p个主成分，而是根据累计贡献率的大小取前k个。
- 定义：称第一主要成分的贡献率为$\lambda_1 / \sum_{i=1}^{p} \lambda_i$，由于$Var(F_1) = \lambda_1$，所以$\lambda_i / \sum_{i=1}^{p} \lambda_i = Var(F_1) / \sum_{i=1}^{p} Var(F_i)$。因此第一主成分的贡献率就是第一主成分的方差在全部方差中的比值。这个值越大，表明第一主成分综合X_1，X_2，…，X_P信息的能力越强。
- 前k个主成分的累计贡献率定义为$\sum_{i=1}^{k} \lambda_i / \sum_{i=1}^{p} \lambda_i$。如果前$k$个主成分的贡献率达到85%，表明取前$k$个主成分基本包含了全部测量指标所具有的信息，这样既减少了变量的个数又方便于对实际问题的分析和研究。

主成分的推导

- 值得指出的是：当协差阵∑未知时，可用其估计值S（样本协差阵）来代替。
- 设原始数据矩阵为：

$$X=\begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix}$$

- 则S=（s_{ij}），其中 $s_{ij}=\frac{1}{n}\sum_{k=1}^{n}(x_{ki}-\overline{x_i})(x_{kj}-\overline{x_j})$
- 而相关系数阵　R=（r_{ij}）

  其中 $r_{ij}=\frac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}}$，显然当原始变量$X_1$，$X_2$，…，$X_P$标准化后，则　$S=R=X'X/n$

标准化就是已知随机变量X的期望Q，方差为S的平方的话，令新变量为（X-Q）/S，新变量就是一个标准的方差

主成分的推导

- 实际应用时，往往指标的量纲不同，所以在计算之前先消除量纲的影响，而将原始数据标准化，这样一来S和R相同。因此一般求R的特征根和特征向量，并且不妨取R=X'X。因为此时的R和X'X/n只差一个系数，特征根差*n*倍，但特征向量不变，不影响求主成分。

计算步骤

- 设有n个样品，每个样品观测p个指标，将原始数据写成矩阵

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

1）将原始数据标准化(标准化就是已知随机变量**X**的期望**Q**，方差为**S**的平方的话，令新变量为（**X-Q**）***/S***，新变量就是一个标准的方差)

2）建立变量的相关系数阵R=X’X

3）R的特征根$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p > 0$及相应的单位特征向量：$a_i=[a_{1i}, a_{2i}, \ldots, a_{pi}]'$。

4）写出主成分$F_i = a_{1i}X_1 + a_{2i}X_2 + \ldots + a_{pi}X_p$

Data Preprocessing: Summary

- Problems during data integration:
  - Different attribute names
  - Different units
  - Different scales
  - Derived attributes
  - Redundant data
- Missing values
  - Imputation
  - Prediction

- Noisy data:
  - Outlier removal
  - Smoothing
  - Normalization
- Data Reduction:
  - Attribute selection
  - Fitting parametric models
  - Sampling
  - Histograms
  - Discretization
  - Aggregation