1、为什么想做数据分析的工作？

在大数据时代，我们能够通过数据分析了解更多数据背后的潜在价值，从而提升销售额或者做出更为理性的决策。

2、你理解中的数据分析是什么内容？

当我们最开始得到数据时，通过作图、造表、公式拟合等方法去探索如何分析数据；在这个基础之上，提出一类或者几类可能的模型再通过进一步的分析挑选相应的模型；最后，应用数理统计的方法去对所定的模型或者估计的可靠程度和精确程度作出推断。

3、你能带给公司什么价值？

慌里慌张的，自己都觉得自己说得不好，逻辑混乱。最重要的是毫无准备。对数据分析了解得太少。其实这种公司的实习感觉不难，挺好进的，但是自己就这么眼睁睁地错失了一个大好机会。投出简历那一刻，就要做好面试的准备。

4、谓假设检验,就是根据研究目的,对样本所属总体特征提出一个假设,然后用适当方法根据样本所提供的信息,对所提出的假设作出拒绝或不拒绝的结论的过程.假设检验一般分为五个步骤:① 建立假设：包括:H0,称无效假设；H1:称备择假设；② 确定检验水准：检验水准用α表示,α一般取0.05；③ 计算检验统计量：根据不同的检验方法,使用特定的公式计算；④确定P值：通过统计量及相应的界值表来确定P值；⑤推断结论：如P＞α,则接受H0,差别无统计学意义；如P≤α,则拒绝H0,差别有统计学意义.

5、假设检验中的两类错误

|  |  |  |  |
| --- | --- | --- | --- |
|  | | 实际情况 | |
| H0正确 | H0错误 |
| 研究结论 | 拒绝H0 | I类错误 | 正确 |
| 接受H0 | 正确 | II类错误 |

[**第一类错误**](https://baike.baidu.com/item/%E7%AC%AC%E4%B8%80%E7%B1%BB%E9%94%99%E8%AF%AF)（**Ⅰ类错误**）也称为 α错误，是指当[虚无假设](https://baike.baidu.com/item/%E8%99%9A%E6%97%A0%E5%81%87%E8%AE%BE/10842837" \t "_blank)(H0)正确时，而拒绝H0所犯的错误。这意味着研究者的结论并不正确，即观察到了实际上并不存在的处理效应。

可能产生原因：

1、样本中极端数值。

2、采用[决策标准](https://baike.baidu.com/item/%E5%86%B3%E7%AD%96%E6%A0%87%E5%87%86/2684523" \t "_blank)较宽松。

[**第二类错误**](https://baike.baidu.com/item/%E7%AC%AC%E4%BA%8C%E7%B1%BB%E9%94%99%E8%AF%AF)（**Ⅱ类错误**）也称为β错误，是指虚无假设错误时，反而接受虚无假设的情况，即没有观察到存在的处理效应。

可能产生的原因：

1、实验设计不灵敏。

2、样本数据变异性过大。

3、处理效应本身比较小。

**两类错误的关系**：

1、 α+β不一定等于1。

2、在[样本容量](https://baike.baidu.com/item/%E6%A0%B7%E6%9C%AC%E5%AE%B9%E9%87%8F/155153" \t "_blank)确定的情况下，α与β不能同时增加或减少。

3、[统计检验](https://baike.baidu.com/item/%E7%BB%9F%E8%AE%A1%E6%A3%80%E9%AA%8C" \t "_blank)力。（1-β）

危害

[编辑](javascript:;)

犯Ⅰ类错误得危害较大，由于报告了本来不存在的现象，则因此现象而衍生出的后续研究、应用的危害将是不可估量的。相对而言，Ⅱ类错误的危害则相对较小，因为研究者如果对自己的假设很有信心，可能会重新设计实验，再次来过，直到得到自己满意的结果（但是如果对本就错误的观点坚持的话，可能会演变成Ⅰ类错误）。

6、回归

Regression Analysis 是一种用来估算两个或者多个变量之间关系.

它有两个明显的好处:

1.能显示自变量和因变量之间的关系

2.能显示自变量和因变量之间的关系的强弱

有多少种:

1. Linear Regression (线性回归)

最常用的回归方法.

因变量是连续的; 自变量可以是连续的,也可以是离散的; 线性的.

注意:

1. 自变量和因变量的关系必须是线性的
2. 噪点对 线性回归 的影响是致命 ( 去除噪点)
3. 自变量可以为多个( multiple linear regression)

2. Logistic Regression

用来估算因变量发生的概率,前提是因变量是binary 结果.

注意:

1. 通常被用在classification 问题上
2. 不要求自变量和因变量的关系是线性的
3. 要求更大的样本数据 - 使用maximum likelihood 方法在大样本 下更加精确
4. 因变量之间不应该相互影响
5. 如果因变量是ordinal ,那么称为 ordinal logistic regression  [了解ordinal](http://blog.csdn.net/ytdxyhz/article/details/51912649)

7、主成分分析

主成分分析也称[主分量分析](https://baike.baidu.com/item/%E4%B8%BB%E5%88%86%E9%87%8F%E5%88%86%E6%9E%90)，旨在利用降维的思想，把多指标转化为少数几个[综合指标](https://baike.baidu.com/item/%E7%BB%BC%E5%90%88%E6%8C%87%E6%A0%87)（即主成分），其中每个主成分都能够反映原始变量的大部分信息，且所含信息互不重复。这种方法在引进多方面变量的同时将复杂因素归结为几个主成分，使问题简单化，同时得到的结果更加科学有效的数据信息。在实际问题研究中，为了全面、系统地分析问题，我们必须考虑众多影响因素。这些涉及的因素一般称为指标，在多元统计分析中也称为变量。因为每个变量都在不同程度上反映了所研究问题的某些信息，并且指标之间彼此有一定的相关性，因而所得的统计数据反映的信息在一定程度上有重叠。主要方法有特征值分解，SVD，NMF等。

principal component analysis(PCA)　主成分分析法是一种数学变换的方法, 它把给定的一组相关变量通过线性变换转成另一组不相关的变量，这些新的变量按照方差依次递减的顺序排列。在数学变换中保持变量的总[方差](https://baike.baidu.com/item/%E6%96%B9%E5%B7%AE" \t "_blank)不变，使第一变量具有最大的方差，称为第一主成分，第二变量的方差次大，并且和第一变量不相关，称为第二主成分。依次类推，I个变量就有I个主成分。

其中Li为p维[正交化](https://baike.baidu.com/item/%E6%AD%A3%E4%BA%A4%E5%8C%96" \t "_blank)[向量](https://baike.baidu.com/item/%E5%90%91%E9%87%8F)（Li\*Li=1），Zi之间互不相关且按照方差由大到小排列，则称Zi为X的第I个主成分。设X的[协方差矩阵](https://baike.baidu.com/item/%E5%8D%8F%E6%96%B9%E5%B7%AE%E7%9F%A9%E9%98%B5" \t "_blank)为Σ，则Σ必为半正定[对称矩阵](https://baike.baidu.com/item/%E5%AF%B9%E7%A7%B0%E7%9F%A9%E9%98%B5" \t "_blank)，求[特征值](https://baike.baidu.com/item/%E7%89%B9%E5%BE%81%E5%80%BC)λi（按从大到小排序）及其[特征向量](https://baike.baidu.com/item/%E7%89%B9%E5%BE%81%E5%90%91%E9%87%8F" \t "_blank)，可以证明，λi所对应的正交化特征向量，即为第I个主成分Zi所对应的系数向量Li，而Zi的方差贡献率定义为λi/Σλj,通常要求提取的主成分的数量k满足Σλk/Σλj>0.85。

进行主成分分析后，还可以根据需要进一步利用[K-L变换](https://baike.baidu.com/item/K-L%E5%8F%98%E6%8D%A2)（霍特林变换）对原数据进行投影变换，达到降维的目的。

## 基本思想

[编辑](javascript:;)

PCA的基本原理就是将一个矩阵中的样本数据投影到一个新的空间中去。对于一个矩阵来说，将其对角化即产生特征根及特征向量的过程，也是将其在标准正交基上投影的过程，而特征值对应的即为该特征向量方向上的投影长度，因此该方向上携带的原有数据的信息越多。

## 主要目的

[编辑](javascript:;)

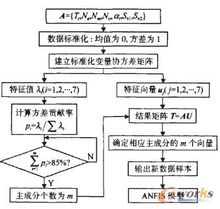
是希望用较少的变量去解释原来资料中的大部分变量，将我们手中许多相关性很高的变量转化成彼此相互独立或不相关的变量。通常是选出比原始变量个数少，能解释大部分资料中变量的几个新变量，即所谓主成分，并用以解释资料的综合性指标。由此可见，主成分分析实际上是一种降维方法。

## 分析步骤

[编辑](javascript:;)

1. 将原始数据按行排列组成矩阵X
2. 对X进行[数据标准化](https://baike.baidu.com/item/%E6%95%B0%E6%8D%AE%E6%A0%87%E5%87%86%E5%8C%96" \t "_blank)，使其均值变为零
3. 求X的协方差矩阵C
4. 将特征向量按特征值由大到小排列，取前k个按行组成矩阵P
5. 通过计算Y = PX，得到降维后数据Y
6. 用下式计算每个特征根的贡献率Vi;Vi=xi/(x1+x2+........)

根据特征根及其特征向量解释主成分[物理意义](https://baike.baidu.com/item/%E7%89%A9%E7%90%86%E6%84%8F%E4%B9%89" \t "_blank)。

[](https://baike.baidu.com/pic/%E4%B8%BB%E6%88%90%E5%88%86%E5%88%86%E6%9E%90%E6%B3%95/2652206/0/908fa0ec08fa513d042055d93e6d55fbb2fbd92a?fr=lemma&ct=single)主成分分析法步骤

举例来说，二维平面有5个点，可以用2\*5的矩阵X来表示：

https://gss2.bdstatic.com/-fo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D156/sign=bb17d0299e0a304e5622a4ffe7c8a7c3/d1160924ab18972b920fa4e7efcd7b899e510ae1.jpg

对X进行归一化，使X每一行减去其对应的均值，得到：

https://gss2.bdstatic.com/-fo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D186/sign=cbe4e6539b58d109c0e3adbae759ccd0/fd039245d688d43f912414df741ed21b0ef43b09.jpg

求X的协方差矩阵：

https://gss1.bdstatic.com/9vo3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D171/sign=ad07ee7d3da85edffe8cfa24785409d8/c9fcc3cec3fdfc0383736495dd3f8794a4c22697.jpg

求解C的特征值，利用线性代数知识或是MATLAB中eig函数可以得到：

https://gss3.bdstatic.com/7Po3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D58/sign=b6bb09bcc411728b342d8c2ac9fc648d/1c950a7b02087bf46d6ad1c4fbd3572c11dfcf1d.jpg

https://gss0.bdstatic.com/-4o3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D58/sign=c2aca3255a43fbf2c12ca62bb17ec740/6609c93d70cf3bc79bcd4bc1d800baa1cc112ac9.jpg

对应的特征向量分别是：

https://gss3.bdstatic.com/7Po3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D101/sign=9ac1e3e6bbfb43161e1f7e7a11a54642/faf2b2119313b07e6d2cd32505d7912397dd8c0e.jpg

https://gss0.bdstatic.com/94o3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D111/sign=91c56fe4de2a60595610e51b1935342d/3801213fb80e7bec942d80cc262eb9389b506b2c.jpg

将原数据降为一维，选择最大的特征值对应的特征向量，因此P为：

https://gss0.bdstatic.com/-4o3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D155/sign=7b62c1d60346f21fcd345a56c3256b31/a044ad345982b2b77c0ab6e938adcbef76099b02.jpg

降维后的数据：

https://gss0.bdstatic.com/-4o3dSag_xI4khGkpoWK1HF6hhy/baike/s%3D425/sign=e22868228e94a4c20e23e6293bf51bac/79f0f736afc379313ef50f04e2c4b74543a91179.jpg

8、因子分析

**因子分析法**是指从研究指标相关矩阵内部的依赖关系出发，把一些信息重叠、具有错综复杂关系的变量归结为少数几个不相关的综合因子的一种多元统计分析方法。基本思想是：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，但不同组的变量不相关或相关性较低，每组变量代表一个基本结构一即公共因子。

**因子分析法的步骤**

　　应用因子分析法的主要步骤如下：

　　(1)对数据样本进行标准化处理。

　　(2)计算样本的相关矩阵R。

　　(3)求相关矩阵R的特征根和特征向量。

　　(4)根据系统要求的累积[贡献率](http://wiki.mbalib.com/wiki/%E8%B4%A1%E7%8C%AE%E7%8E%87" \o "贡献率)确定主因子的个数。

　　(5)计算因子载荷矩阵A。

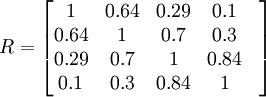
　　(6)确定因子模型。

　　(7)根据上述计算结果，对系统进行分析。

**因子分析法的实例**[**[1]**](http://wiki.mbalib.com/wiki/%E5%9B%A0%E5%AD%90%E5%88%86%E6%9E%90%E6%B3%95#_note-N)

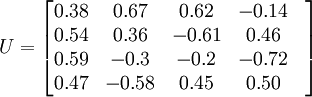
【例：1】

　　假设某一社会经济系统问题，其主要特性可用4个指标表示，它们分别是生产、 技术、交通和环境。其[相关矩阵](http://wiki.mbalib.com/wiki/%E7%9B%B8%E5%85%B3%E7%9F%A9%E9%98%B5)为：



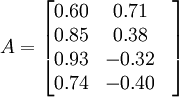
　　相应的特征值、占总体百分比和累计百分比如下表：

[](http://wiki.mbalib.com/wiki/Image:%E5%9B%A0%E5%AD%90%E5%88%86%E6%9E%90%E6%B3%95_.jpg)

　　对应特征值的特征向量矩阵为： 

　　假如要求所取特征值反映的信息量占总体信息量的90％以上，则从累计特征值所占百分 比看，只需取前两项即可。也就是说，只需取两个主要因子。对应于前两列特征值的特征向量，

　　可求得其因子载荷矩阵A为：



　　于是，该问题的因子模型为：

*Xl* = 0.60*f*1 + 0.71*f*2

*X*2 = 0.85*f*1 + 0.38*f*2

*X*3 = 0.93*f*1 − 0.32*f*2

*X*4 = 0.74*f*1 − 0.40*f*2

由以上可以看出，两个因子中，*f*1是全面反映生产、技术、交通和环境的因子，而*f*2却不同，它反映了对生产和技术这两项增长有利，而对交通和环境增长不利的因子。也就是说，按照原有统计资料得出的相关矩阵分析的结果是如果生产和技术都随*f*2增长了，将有可能出现交通紧张和环境恶化的问题，*f*2反映了这两方面的相互制约状况。

[[编辑](http://wiki.mbalib.com/w/index.php?title=%E5%9B%A0%E5%AD%90%E5%88%86%E6%9E%90%E6%B3%95&action=edit&section=4)]

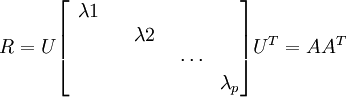
**因子分析与主成分分析的区别**[**[2]**](http://wiki.mbalib.com/wiki/%E5%9B%A0%E5%AD%90%E5%88%86%E6%9E%90%E6%B3%95#_note-A)

　　因子分析法与主成分分析法都属于[因素分析法](http://wiki.mbalib.com/wiki/%E5%9B%A0%E7%B4%A0%E5%88%86%E6%9E%90%E6%B3%95" \o "因素分析法)，都基于[统计分析方法](http://wiki.mbalib.com/wiki/%E7%BB%9F%E8%AE%A1%E5%88%86%E6%9E%90%E6%96%B9%E6%B3%95)，但两者有较大的区别：[主成分分析](http://wiki.mbalib.com/wiki/%E4%B8%BB%E6%88%90%E5%88%86%E5%88%86%E6%9E%90)是通过坐标变换提取主成分，也就是将一组具有相关性的变量变换为一组独立的变量，将主成分表示为原始观察变量的线性组合；而因子分析法是要构造因子模型，将原始观察变量分解为因子的线性组合。通过对上述内容的学习，可以看出因子分析法和主成分分析法的主要区别为：

　　(1)主成分分析是将主要成分表示为原始观察变量的线性组合，而因子分析是将原始观察变量表示为新因子的线性组合，原始观察变量在两种情况下所处的位置不同。

　　(2)主成分分析中，新变量Z的坐标维数j(或主成分的维数)与原始变量维数相同，它只是将一组具有相关性的变量通过正交变换转换成一组维数相同的独立变量，再按总方差误差的允许值大小，来选定q个(q<p)主成分；而因子分析法是要构造一个模型，将问题的为数众多的变量减少为几个新因子，新因子变量数m小于原始变量数P，从而构造成一个结构简单的模型。可以认为，因子分析法是主成分分析法的发展。

　　(3)主成分分析中，经正交变换的变量系数是相关矩阵R的特征向量的相应元素；而因子分析模型的变量系数取自因子负荷量，即a_{\ddot{y}}=u_{\ddot{y}}\sqrt{\lambda}。因子负荷量矩阵A与相关矩阵R满足以下关系：



　　其中，U为R的特征向量。

　　在考虑有残余项ε时，可设包含ε*i*的矩阵ρ为误差项，则有*R* − *AAT* = ρ。

　　在因子分析中，残余项应只在ρ的对角元素项中，因特殊项只属于原变量项，因此，a_\ddot{y}的选择应以ρ的非对角元素的方差最小为原则。而在主成分分析中，选择原则是使舍弃成分所对应的方差项累积值不超过规定值，或者说被舍弃项各对角要素的自乘和为最小，这两者是不同的。

9、分类聚类

### 什么是分类

分类任务就是明确对象属于哪个预定义的目标类。其中预定义的目标类是离散时为分类，连续时为回归。

### 有哪些分类方法

常用的分类算法有决策树，基于规则的分类算法，神经网络，支持向量机和朴素贝叶斯分类法等。（参考[常用分类算法总结](http://blog.csdn.net/shouwangcc/article/details/48101761)）

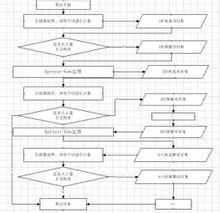
* 决策树
* 基于规则的分类算法
* 朴素贝叶斯 [杂货铺](http://www.cnblogs.com/leoo2sk/archive/2010/09/17/naive-bayesian-classifier.html)
* 最近邻分类器
* 贝叶斯信念网络（BBN）
* 人工神经网络
* 支持向量机的特征（SVM）

### 什么是聚类

聚类就是按照某个特定标准(如距离准则，即数据点之间的距离)把一个数据集分割成不同的类或簇，使得同一个簇内的数据对象的相似性尽可能大，同时不在同一个簇中的数据对象的差异性也尽可能地大。

10、关联分析

Apriori算法

[](https://baike.baidu.com/pic/%E5%85%B3%E8%81%94%E5%88%86%E6%9E%90/1198018/0/8718367adab44aedaf0000bdb01c8701a18bfb19?fr=lemma&ct=single)

[Apriori](https://baike.baidu.com/item/Apriori)算法是挖掘产生布尔关联规则所需频繁项集的基本算法，也是最著名的关联规则挖掘算法之一。Apriori算法就是根据有关频繁项集特性的先验知识而命名的。它使用一种称作逐层搜索的迭代方法，k—项集用于探索（k+1）—项集。首先，找出频繁1—项集的集合．记做L1，L1用于找出频繁2—项集的集合L2，再用于找出L3，如此下去，直到不能找到频繁k—项集。找每个Lk需要扫描一次数据库。

11、时间序列模型

[时间序列分析](https://baike.baidu.com/item/%E6%97%B6%E9%97%B4%E5%BA%8F%E5%88%97%E5%88%86%E6%9E%90)是根据系统观测得到的[时间序列数据](https://baike.baidu.com/item/%E6%97%B6%E9%97%B4%E5%BA%8F%E5%88%97%E6%95%B0%E6%8D%AE)，通过[曲线拟合](https://baike.baidu.com/item/%E6%9B%B2%E7%BA%BF%E6%8B%9F%E5%90%88)和参数估计来建立数学模型的理论和方法。它一般采用曲线拟合和[参数估计](https://baike.baidu.com/item/%E5%8F%82%E6%95%B0%E4%BC%B0%E8%AE%A1)方法（如[非线性最小二乘法](https://baike.baidu.com/item/%E9%9D%9E%E7%BA%BF%E6%80%A7%E6%9C%80%E5%B0%8F%E4%BA%8C%E4%B9%98%E6%B3%95)）进行。时间序列分析常用在国民经济宏观控制、区域综合发展规划、[企业经营管理](https://baike.baidu.com/item/%E4%BC%81%E4%B8%9A%E7%BB%8F%E8%90%A5%E7%AE%A1%E7%90%86)、[市场潜量预测](https://baike.baidu.com/item/%E5%B8%82%E5%9C%BA%E6%BD%9C%E9%87%8F%E9%A2%84%E6%B5%8B)、气象预报、[水文预报](https://baike.baidu.com/item/%E6%B0%B4%E6%96%87%E9%A2%84%E6%8A%A5)、[地震前兆](https://baike.baidu.com/item/%E5%9C%B0%E9%9C%87%E5%89%8D%E5%85%86)预报、农作物病虫灾害预报、[环境污染控制](https://baike.baidu.com/item/%E7%8E%AF%E5%A2%83%E6%B1%A1%E6%9F%93%E6%8E%A7%E5%88%B6)、[生态平衡](https://baike.baidu.com/item/%E7%94%9F%E6%80%81%E5%B9%B3%E8%A1%A1)、天文学和[海洋学](https://baike.baidu.com/item/%E6%B5%B7%E6%B4%8B%E5%AD%A6)等方面。

### ARMA模型

ARMA模型的全称是自回归移动平均(auto regression moving average)模型，它是目前最常用的[拟合](https://baike.baidu.com/item/%E6%8B%9F%E5%90%88)平稳序列的模型，它又可细分为**AR模型**(auto regression model)、**MA模型**(moving average model)和**ARMA模型**(auto regression moving average model)三大类。

### ARIMA模型

ARIMA模型又称自回归求和移动平均模型，当时间序列本身不是平稳的时候，如果它的增量，即的一次差分，稳定在零点附近，可以将看成是平稳序列。在实际的问题中，所遇到的多数[非平稳序列](https://baike.baidu.com/item/%E9%9D%9E%E5%B9%B3%E7%A8%B3%E5%BA%8F%E5%88%97)可以通过一次或多次差分后成为平稳时间序列，则可以建立模型：

## 步骤

[编辑](javascript:;)

### 抽样

用观测、调查、统计、抽样等方法取得被观测系统时间序列动态数据。

### 作图

根据动态数据作[相关图](https://baike.baidu.com/item/%E7%9B%B8%E5%85%B3%E5%9B%BE)，进行[相关分析](https://baike.baidu.com/item/%E7%9B%B8%E5%85%B3%E5%88%86%E6%9E%90)，求[自相关函数](https://baike.baidu.com/item/%E8%87%AA%E7%9B%B8%E5%85%B3%E5%87%BD%E6%95%B0)。相关图能显示出变化的趋势和周期，并能发现跳点和拐点。跳点是指与其他数据不一致的[观测值](https://baike.baidu.com/item/%E8%A7%82%E6%B5%8B%E5%80%BC)。如果跳点是正确的观测值,在建模时应考虑进去,如果是反常现象，则应把跳点调整到期望值。拐点则是指时间序列从上升趋势突然变为下降趋势的点。如果存在拐点，则在建模时必须用不同的模型去分段[拟合](https://baike.baidu.com/item/%E6%8B%9F%E5%90%88)该时间序列，例如采用门限[回归模型](https://baike.baidu.com/item/%E5%9B%9E%E5%BD%92%E6%A8%A1%E5%9E%8B)。

### 拟合

辨识合适的[随机模型](https://baike.baidu.com/item/%E9%9A%8F%E6%9C%BA%E6%A8%A1%E5%9E%8B),进行[曲线拟合](https://baike.baidu.com/item/%E6%9B%B2%E7%BA%BF%E6%8B%9F%E5%90%88),即用通用随机模型去拟合时间序列的观测数据。对于短的或简单的时间序列，可用趋势模型和季节模型加上误差来进行[拟合](https://baike.baidu.com/item/%E6%8B%9F%E5%90%88)。对于平稳时间序列，可用通用ARIMA模型（[自回归滑动平均模型](https://baike.baidu.com/item/%E8%87%AA%E5%9B%9E%E5%BD%92%E6%BB%91%E5%8A%A8%E5%B9%B3%E5%9D%87%E6%A8%A1%E5%9E%8B)）及其特殊情况的[自回归模型](https://baike.baidu.com/item/%E8%87%AA%E5%9B%9E%E5%BD%92%E6%A8%A1%E5%9E%8B)、滑动平均模型或组合-ARIMA模型等来进行拟合。当观测值多于50个时一般都采用ARIMA模型。对于非平稳时间序列则要先将观测到的时间序列进行[差分](https://baike.baidu.com/item/%E5%B7%AE%E5%88%86)运算，化为平稳时间序列，再用适当模型去拟合这个差分序列。

时间序列是一种特殊的随机过程，当中的取非负整数时，就可以代表各个时刻，就可以看作是时间序列（time series),因此，当一个随机过程可以看作时间序列时，我们就可以利用现有的时间序列模型建模分析该随机过程的特性。

## 用途

[编辑](javascript:;)

### 描述

根据对系统进行观测得到的时间序列数据，用[曲线拟合](https://baike.baidu.com/item/%E6%9B%B2%E7%BA%BF%E6%8B%9F%E5%90%88)方法对系统进行客观的描述。

### 分析

当观测值取自两个以上变量时，可用一个时间序列中的变化去说明另一个时间序列中的变化，从而深入了解给定时间序列产生的机理。

### 预测

一般用ARMA模型拟合时间序列，预测该时间序列未来值。

### 决策

根据时间序列模型可调整输入变量使系统发展过程保持在目标值上，即预测到过程要偏离目标时便可进行必要的控制。

12、方差分解

偏差：度量了模型的期望预测和真实结果的偏离程度，刻画了模型本身的拟合能力。

方差：度量了同样大小的训练集的变动所导致的学习性能的变化，即刻画了数据扰动所造成的影响。

噪声：表达了当前任务上任何模型所能达到的期望泛化误差的下界，刻画了学习问题本身的难度。

　方差分解（Variance Decomposition）是通过分析每一个结构的冲击对各变量方差的贡献值。显示出每个有着随机扰动影响VAR模型中的干扰变量信息。

VAR的方差分解目的在于将一个变量的方差归因，比如如果一个VAR系统中有x，y两个变量，那么对其进行方差分解就可以明白，在某一个时间t，x的方差有多少是自己引起的，有多少是y引起的；同样，y的方差有多少是y引起的，多少是x引起的