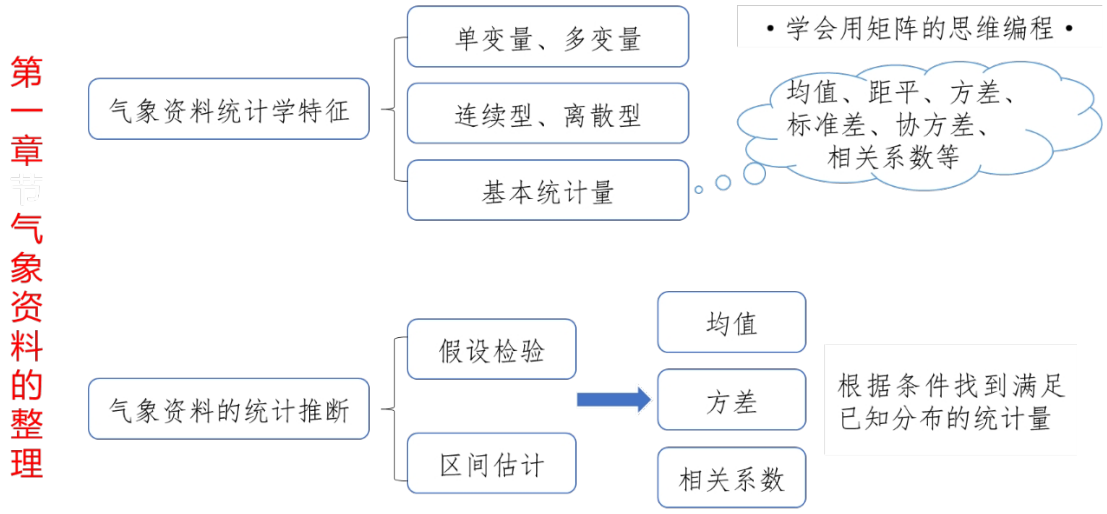
《气象统计方法》重要知识点整理

整理：19 大气 2 班蒋斌

第一章 气象资料的整理

➢ *章节思维导图*



➢ *重要知识点*

一、 **单变量资料的基本统计量表示（向量）**

※某一气象要素的 n 次观测，记作一组向量 $\boldsymbol{x}=[x_1,x_2,x_3,...,x_n]$（以下公式推导均以该向量为例）

1. （算数）平均值

$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，反映要素在时间上的平均状况。实际应用中常常对一年的 12 个月份或者某几个月求平均，来反映季节循环，这也方便资料的进一步处理（如去季节化处理）。

2. 距平(deviation；anomaly) ★

指单个数据与均值之差，即 $x_{di}=x_i-\overline{x}\quad(i=1,2,...,n)$，它反映了单个数据偏离平均值的大小。把原始资料减去均值转化为距平资料的过程，称为“中心化”。从原始资料序列中减去季节循环，得到年际异常（距平）序列。

3. 平均差与相对平均差（应用较少）

$v_a=\frac{1}{n}\sum_{i=1}^{n}|x_i-\bar{x}|$，表示样本各数据相对于均值的平均差异。平均差除以该要素的平均值，称为相对平均差：$v_r=\frac{v_a}{\bar{x}}$。

4. 方差(variance)与标准差(standard deviation) ★

方差表示为：$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2=\frac{1}{n-1}\sum_{i=1}^{n}x_{di}^2=\frac{(\boldsymbol{x}-x)(\boldsymbol{x}-x)^{\mathrm{T}}}{n-1}$；标准差为：$s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$（均使用无偏估计）二者均反映了数据整体的离散程度。

（注意用向量表示的方法，便于编程）

◆ **数据的标准化((Normalization，Standardization)**

消除由于量纲不同造成无法比较的情况。数据的标准化处理，即用距平资料除以标准差，表达式为：$x_{si}=\frac{x_i-\bar{x}}{s}\left(\boldsymbol{x}_s=\frac{\boldsymbol{x}-\bar{x}}{s}\right)$（该关系要牢记：$x_{si}=x_{di}s$）。

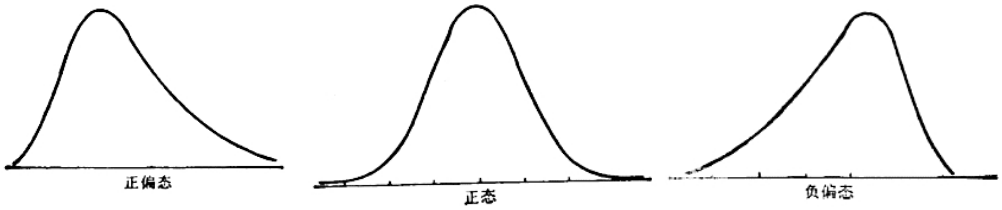
标准化之后，不同气象要素之间、或同一要素不同地点之间的异常程度即可互相比较。标准化数据有两个重要的性质：**(1) 平均值= 0；(2)标准差=1.**

5. 偏度(Skewness)与峰度(Kurtosis)

(1) x 的 k 阶中心矩 $\mu_k=E\left\{\left[x-E(x)\right]^k\right\}$，用样本估计为：$\mu_k=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^k$。

显然 x 的 2 阶矩 μ_2 表示方差，$\sqrt{\mu_2}$ 称为标准差。

(2) 偏度：反映统计分布的偏斜方向和程度，即非对称程度，表达式为 $skew(x)=\frac{\mu_3}{\mu_2^{3/2}}=\frac{\mu_3}{s^3}$。偏度>0（正偏态），此时数据位于均值左边的多，因为有少数变量值很大，使曲线右侧尾部拖得很长，在图上表现为**长尾在右，高峰在左，整体呈现分布右偏**，负偏态则反之。偏度为零，表示数值相对均匀地分布在平均值的两侧，但不一定是对称分布。



(3) 峰度：概率密度分布曲线的陡峭程度，表达式为 $Kurtosis(x)=\frac{\mu_4}{\mu_2^2}$。峰度越大，概率密度分布曲线更陡峭。**正态分布的峰度恒等于 3**。

6. 协方差(Covariance) ★

现在对两个气象要素 x 和 y 同时观测 n 次，得到它们的观测序列：$\boldsymbol{x}=[x_1,x_2,x_3,...,x_n],\boldsymbol{y}=[y_1,y_2,y_3,...,y_n]$。则 x 和 y 总体的协方差的无偏估计为：

$$s_{xy}=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\frac{(\boldsymbol{x}-\bar{x})(\boldsymbol{y}-\bar{y})^{\mathrm{T}}}{n-1}$$

协方差表征两变量的协同变化或密切程度，具有对称性：$s_{xy}=s_{yx}$，由此也可以知道，方差是协方差的特例。

7. Pearson 相关系数(Correlation coefficient) ★

用来度量两个变量之间的**线性相关**程度,，表达式为：

$$r_{xy}=\frac{Cov(x,y)}{s_x s_y}=\frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2}}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$
$$=\frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{s_x}\frac{(y_i-\bar{y})}{s_y}=Cov(x_s,y_s)$$

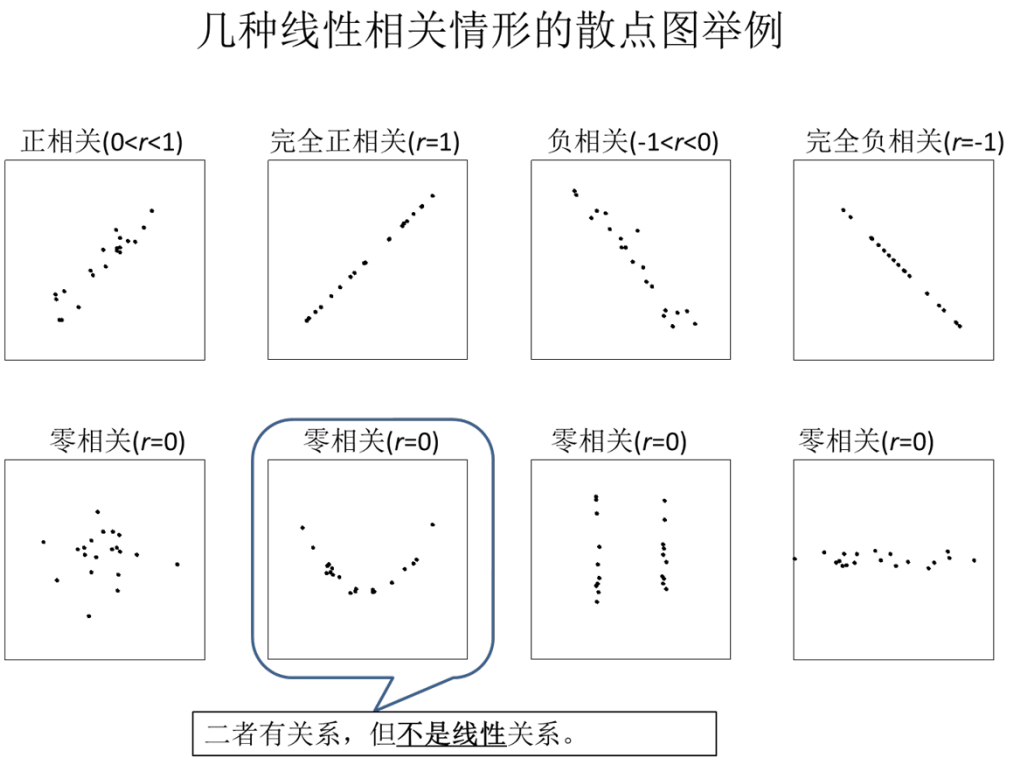
上式表明，当数据经过**标准化处理**后，其**相关系数等同于协方差**。同样的，相关系数具有对称性即 $r_{xy}=r_{yx}$，取值范围为 $[-1,1]$。相关系数的一些性质，可以通过一些数学上的推到得出。比如：

概率论公式回顾：

$$D(cX)=c^2D(X);Cov(aX,bY)=abCov(X,Y);Cov(X_1+X_2,Y)=Cov(X_1,Y)+Cov(X_2,Y)$$

(1) $r_{-x,-y}=\dfrac{Cov(-x,-y)}{\sqrt{D(-x)}\sqrt{D(-y)}}=\dfrac{Cov(x,y)}{\sqrt{D(x)}\sqrt{D(y)}}=r_{xy}$；

(2) $r_{-ax+b,-cy+d}=\dfrac{Cov(-ax+b,-cy+d)}{\sqrt{D(-ax+b)}\sqrt{D(-cy+d)}}=\dfrac{Cov(-ax,-cy)}{\sqrt{D(-ax)}\sqrt{D(-cy)}}=r_{xy}$



◆ 等级相关系数(Spearman 秩相关系数)

将原始数据 x 和 y 分别从小到大（或从大到小）排列，把排列后各数据的位置序号(称为“秩次”)作为新的秩数据 x_2,y_2，然后对 x_2,y_2 计算简单相关系数。

二、 多变量资料的统计量表示（矩阵）

假设有 m 个因子，对每个因子均同时进行 n 次观测，那么此时便可用矩阵来表示这些数据，我们以摆放为 m 行 n 列为例，有原始资料矩阵为：

$$\boldsymbol{X}_{(m\times n)}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

1. 均值向量、方差向量以及标准差向量

原始资料矩阵 $\boldsymbol{X}$ 每一行代表一个因子，对每一行均可以进行求均值、方差和标准差，m 个因子便构成了一个列向量：

$$\boldsymbol{x}=\begin{bmatrix}\overline{x}_1\\ \overline{x}_2\\ \vdots\\ \overline{x}_n\end{bmatrix},\overline{x}_i=\frac{1}{n}\sum_{j=1}^{n}x_{ij}\ ;\quad \boldsymbol{s}^2=\begin{bmatrix}s_1^2\\ s_2^2\\ \vdots\\ s_n^2\end{bmatrix},\boldsymbol{s}=\begin{bmatrix}s_1\\ s_2\\ \vdots\\ s_n\end{bmatrix},s_i^2=\frac{1}{n-1}\sum_{j=1}^{n}\left(x_{ij}-\overline{x}_i\right)^2$$

2. 协方差矩阵（注意对角线上的元素意义）

协方差矩阵可以表示 m 个变量之间的**两两**关系。把原始资料矩阵的每个因子的 n 次观测记为一个随机变量向量：$\boldsymbol{X}_i=\left[x_{i1},x_{i2},...,x_{in}\right]$，原始资料阵为：

$$\boldsymbol{X}_{(m\times n)}=\begin{bmatrix}\boldsymbol{X}_1\\ \boldsymbol{X}_2\\ \vdots\\ \boldsymbol{X}_m\end{bmatrix}$$ 于是协方差矩阵可以表示为：

$$Cov\_matrix\left(\boldsymbol{X}\right)=\begin{bmatrix}Cov\left(\boldsymbol{X}_1,\boldsymbol{X}_1\right) & Cov\left(\boldsymbol{X}_1,\boldsymbol{X}_2\right) & \cdots & Cov\left(\boldsymbol{X}_1,\boldsymbol{X}_n\right)\\ Cov\left(\boldsymbol{X}_2,\boldsymbol{X}_1\right) & Cov\left(\boldsymbol{X}_2,\boldsymbol{X}_2\right) & \cdots & Cov\left(\boldsymbol{X}_2,\boldsymbol{X}_n\right)\\ \vdots & \vdots & \ddots & \vdots\\ Cov\left(\boldsymbol{X}_m,\boldsymbol{X}_1\right) & Cov\left(\boldsymbol{X}_m,\boldsymbol{X}_2\right) & \cdots & Cov\left(\boldsymbol{X}_m,\boldsymbol{X}_m\right)\end{bmatrix}$$

由协方差的性质可知，**协方差矩阵是实对称矩阵，且对角线上的元素表示每个因子的方差**。

对于双变量，协方差为 $\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)$，也就是距平乘积的均值，于是我们可以先计算得到原始资料的距平阵，记为 $\boldsymbol{X}_d$，那么有：

$$\boldsymbol{X}_d=\begin{bmatrix}x_{11} & x_{12} & \cdots & x_{1n}\\ x_{21} & x_{22} & \cdots & x_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ x_{m1} & x_{m2} & \cdots & x_{mn}\end{bmatrix}-\begin{bmatrix}\overline{x}_1 & \overline{x}_1 & \cdots & \overline{x}_1\\ \overline{x}_2 & \overline{x}_2 & \cdots & \overline{x}_2\\ \vdots & \vdots & \ddots & \vdots\\ \overline{x}_n & \overline{x}_n & \cdots & \overline{x}_n\end{bmatrix}=\begin{bmatrix}x_{d11} & x_{d12} & \cdots & x_{d1n}\\ x_{d21} & x_{d22} & \cdots & x_{d2n}\\ \vdots & \vdots & \ddots & \vdots\\ x_{dm1} & x_{dm2} & \cdots & x_{dmn}\end{bmatrix}$$

由此可以计算得到协方差矩阵为：

$$\boldsymbol{S}=\frac{1}{n-1}\boldsymbol{X}_d\boldsymbol{X}_d^{\mathrm{T}}=\begin{bmatrix}s_{11} & s_{12} & \cdots & s_{1n}\\ s_{21} & s_{22} & \cdots & s_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ s_{m1} & s_{m2} & \cdots & s_{mn}\end{bmatrix},s_{ij}=\frac{1}{n-1}\boldsymbol{x}_{di}\boldsymbol{x}_{dj}^{\mathrm{T}}\left(i.j=1,2,...,m\right)$$

3. 相关系数阵

同上述协方差矩阵一样，写出相关系数矩阵的形式为：

$$R\_matrix(\boldsymbol{X})=\begin{bmatrix} r_{X_1X_1} & r_{X_1X_2} & \cdots & r_{X_1X_n} \\ r_{X_2X_1} & r_{X_2X_2} & \cdots & r_{X_2X_n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{X_mX_1} & r_{X_mX_2} & \cdots & r_{X_mX_m} \end{bmatrix}$$

根据相关系数的性质，可知**相关系数矩阵也是实对称矩阵，且对角线上的元素得值均为 1**。

对于双变量，相关系数为 $\frac{1}{n-1}\sum_{i=1}^{n}\frac{(x_i-\bar{x})}{s_x}\frac{(y_i-\bar{y})}{s_y}$，因此先根据原始数据制作标准化资料阵：

$$\boldsymbol{X}^*=\begin{bmatrix} x_{11}^* & x_{12}^* & \cdots & x_{1n}^* \\ x_{21}^* & x_{22}^* & \cdots & x_{2n}^* \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1}^* & x_{m2}^* & \cdots & x_{mn}^* \end{bmatrix}\left(x_{ij}^*=\frac{x_{ij}-\bar{x}_i}{s_i}, i=1,2,...,m; j=1,2,...,\mathrm{n}\right)$$

则相关系数阵为：

$$\boldsymbol{R}=\frac{1}{n-1}\boldsymbol{X}^*\left(\boldsymbol{X}^*\right)^{\mathrm{T}}=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{bmatrix}, r_{ij}=\frac{1}{n-1}\boldsymbol{x}_i^*\left(\boldsymbol{x}_j^*\right)^{\mathrm{T}}(i,j=1,2,...,m)$$

◆ **用协方差矩阵写出相关系数矩阵**

已知协方差阵 $\boldsymbol{S}=\begin{bmatrix} 4 & -4 & 3 \\ -4 & 9 & -2 \\ 3 & -2 & 16 \end{bmatrix}$，求相关系数阵 $\boldsymbol{R}$.

[解析]：根据协方差的性质，主对角线的元素为每个变量因子的方差，其余位置的表示不同因子之间的协方差，所以有

$$s_1=2, s_2=3, s_3=4 \qquad s_{12}=-4, s_{13}=3, s_{23}=-2$$

根据公式 $r_{xy}=\frac{s_{xy}}{s_x s_y}$，得到 $r_{12}=-\frac{2}{3}, r_{13}=\frac{3}{8}, r_{23}=-\frac{1}{6}$，故相关系数阵为：

$$\boldsymbol{R}=\begin{bmatrix} 1 & -\frac{2}{3} & \frac{3}{8} \\ -\frac{2}{3} & 1 & -\frac{1}{6} \\ \frac{3}{8} & -\frac{1}{6} & 1 \end{bmatrix}$$

三、 区间估计与假设检验

※ 此部分不需要背诵很长的公式，重在理解，知道如何编程实现。

1. 主要统计量与服从的分布

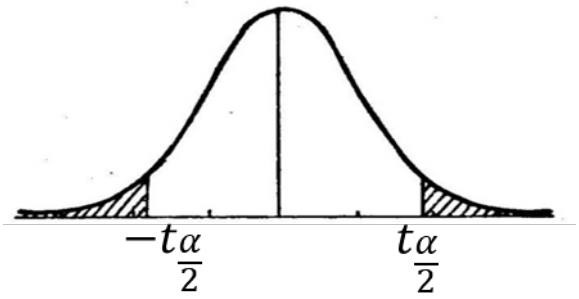
	单个均值	单个方差	两个均值（方差未知）	两个方差	相关系数	两个相关系数差异
统计量	$\frac{\bar{x}-\mu}{s/\sqrt{n}}$	$\frac{(n-1)s^2}{\sigma^2}$	$t=\frac{\bar{x}_1-\bar{x}_2-(\mu_1-\mu_2)}{s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$	$\frac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2}$	$\frac{r}{\sqrt{1-r^2}}\sqrt{n-2}$	$\frac{z_{r_1}-z_{r_2}}{\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}}$
服从的分布	$t(n-1)$	$\chi^2(n-1)$	$t(n_1+n_2-2)$	$F(n_1-1,n_2-1)$	$t(n-2)$	$N(0,1)$

注：两个均值假设检验与区间估计时，统计量s的值为 $\sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$

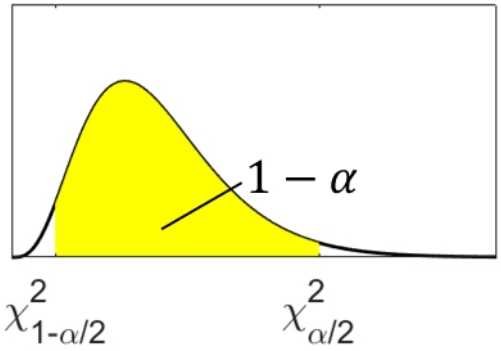
两个相关系数的差异检验用到的Fisher变换为 $z_r=\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)\sim N\left(\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right),\frac{1}{n-3}\right)$

2. 三种分布的表现形式

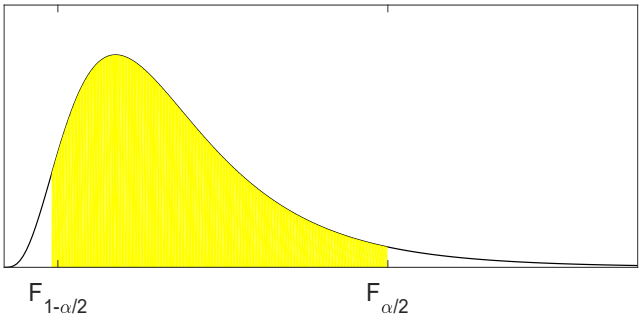
(1) *t* 分布（下图阴影部分为拒绝域）



(2) 卡方分布（下图黄色部分之外为拒绝域）



(3) *F* 分布（下图黄色部分之外为拒绝域）



三种分布中，只有 *t* 分布是对称的。正确记住不同分布的形式才能写出正确的拒绝域以及求出临界值。

3. 术语区分

(1) 显著性水平与置信水平

- 显著性水平，又叫信度，英文为 **significance level**，用***α***表示，是一个小概率，基于假设检验，正确的表述为：“信度为 0.05” “0.05 的显著性水平”。
- 置信水平，又叫置信概率，英文为 **confidence level**，用 **1- *α***表示，是一个大概率，基于参数估计，正确的表述为：“置信水平为 95%”“通过了 95%的置信水平”。

(2) “相关系数很显著”与“相关很高(大、密切)”

- 相关系数很高，是指相关系数的数值(*r* 或 ρ 的绝对值)很大；
- 相关系数的显著性程度则对应于显著性水平 *α*值(或 *p* 值)：相关很显著是指 *α*值很小，两总体存在相关的概率很高。

(3) 当对样本相关系数 *r* 进行假设检验的结论为“拒绝原假设 $H_0(\rho=0)$”时，常描述为：**“相关系数通过了显著性水平为*α*的显著性检验”**，其等价描述有：

- 从 $\rho=0$ 的一对总体中随机抽样得到相关为 *r* 的一对样本，是个小概率事件；

- 两总体存在相关($\rho \neq 0$)的可能性很大 ，相互独立($\rho = 0$)的可能性很小；
- “H_0为真($\rho = 0$)的情况下,却拒绝H_0”(犯第一类错误,弃真)的概率为α。

4. 应用

(1) 两均值检验与合成分析法

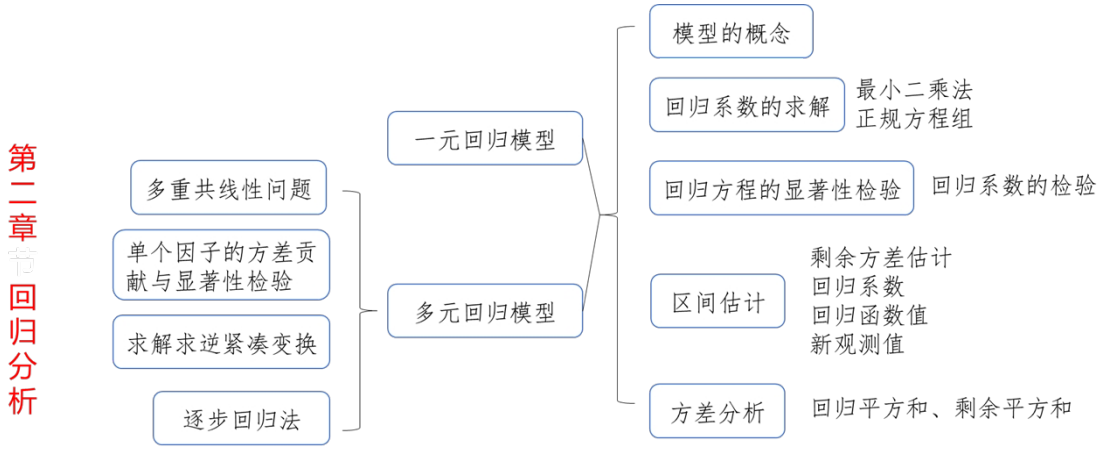
对有特殊气候现象的年份的某一气象要素求均值，将其与正常年份的均值进行比较。

(2) 两均值差异检验与检测极端气候事件

检验样本 x 中的某个数据 x_i 与其他数据是否有显著差别，相当于“两均值差异检验”的特殊情况：一个样本的容量为 1（即 x_i），另一样本容量设为 n（ $x(t)$ ，表示除 x_i 以外的其他所有数据）。则可以假设 H_0： x_i 与其他数据的均值无显著差别，原统计量 $t = \dfrac{\bar{x}_1 - \bar{x}_2}{s\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$ 调整为 $t = \dfrac{x_i - \bar{x}(t)}{s}\sqrt{\dfrac{n}{n+1}} \sim t(n-1)$

第二章 回归分析

➢ ***章节思维导图***



一、一元回归模型

1. 一元回归模型概述

随机变量 y(预报量)的取值与一个预报因子 x 之间存在关系：

$$y=\beta_0+\beta x+\varepsilon \quad \varepsilon\sim N\left(0,\sigma^2\right)$$

称这是**以 x 为自变量，以 y 为因变量建立的一元回归模型**（或者“对 x 建立关于 y 的一元回归模型”“建立 y 对 x 的一元回归模型”）。x 称为自变量、回归变量，y 称为因变量、响应变量，回归模型中 y 的取值由两部分组成：

(1) 关于 x 的线性函数 $\beta_0+\beta x$ ，β_0 和 β 都是不依赖于 x 的常数，称为回归系数，但它们的取值是**未知**的；

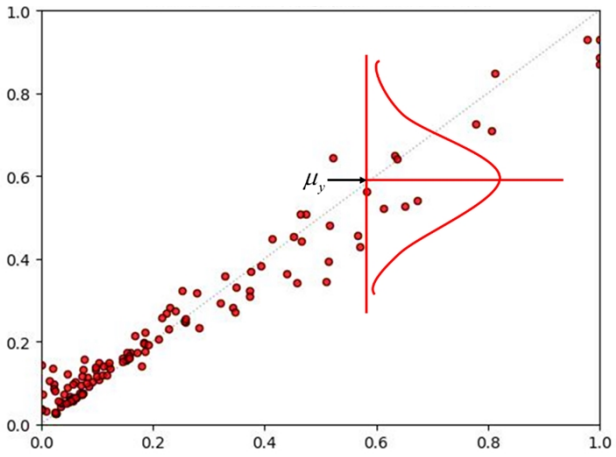
(2) **随机误差ε**，服从正态分布 $N(0,\sigma^2)$， 是 x 以外的其他各种因素导致的 y 的随机误差，属于 y 中“不可控部分”。

▲ 进一步理解回归模型

(1) ***x* 不是随机变量**，是可以精确控制或观察的变量；

(2) 对于一个确定的 x， y 的取值具有随机性，但 y 的数学期望 $\beta_0+\beta x$ 是确定的，y 值围绕期望上下波动。

$$E\left(y\right)=\beta_0+\beta E\left(x\right)\Leftrightarrow \mu_y=\beta_0+\beta x$$



2. 回归模型的求解（最小二乘法）

我们的目的是已知 x_0 的值，估计出它对应的 y_0，依据公式 $y_0=\beta_0+\beta x_0+\varepsilon$，需要：(1)把 x_0 代入得 y_0 的期望 $\mu_{y0}=\beta_0+\beta x_0$；(2)根据$\varepsilon$的方差，得到的 y_0 的置信区间。因此，问题的求解便转化为：**(1) 估即回归系数 β_0 和 β 的值；(2) 估计 ε的方差。**

方法：根据观测得到 x 和 y 数据进行线性拟合，得到回归方程 $\hat{y}=b_0+bx$。b,b_0 分别是对 β,β_0 的估计。实际的 y 与估计的 $\hat{y}$ 总是存在差异，称 $e=y-\hat{y}$ 为残差，故又有 $y=\hat{y}+e=b_0+bx+e$。采用最小二乘法求解拟合的回归系数：

$$b=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2};\quad b_0=\bar{y}-b\bar{x}$$

▲ 回归方程的性质

(1) 样本中心点 $(\bar{x},\bar{y})$ 必过回归直线；

(2) $\bar{\hat{y}}=b_0+b\bar{x}=(\bar{y}-b\bar{x})+b\bar{x}=\bar{y}$；

(3) 回归系数的符号(正负)取决于 $Cov(x,y)$（x,y 的离差乘积）的符号；

(4) 当 y 和 x 都是**距平资料**时(即 $\bar{x}=\bar{y}=0$)，有 b_0=0，于是回归方程变为：

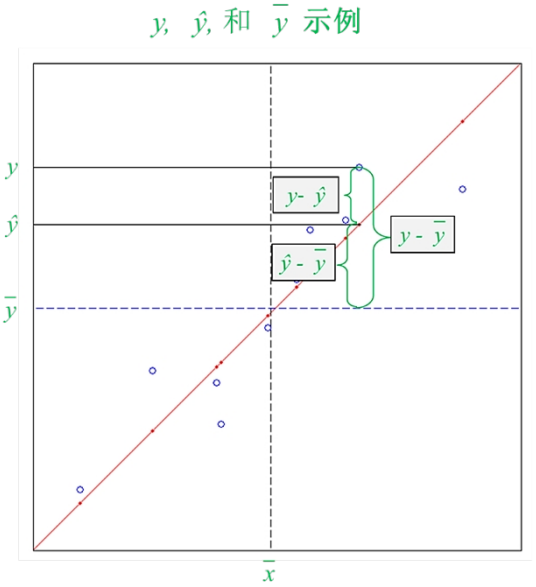
$\hat{y}=bx$，回归系数的含义：**当 x 的变化(距平)为 1 个单位时，y 距平的估计值**

(5) 回归系数与相关系数的关系：

$$b=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{S_{xy}}{S_{xx}}=\frac{s_{xy}}{s_x^2} \qquad r_{xy}=\frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}}=\frac{s_{xy}}{s_x s_y}$$

对比上述两个表达式，可以得到回归系数与相关系数之间的关系：$b=r_{xy}\frac{s_y}{s_x}$。可以看到b,r同号，其正负都取决于$Cov(x,y)$（或者x,y的离差乘积）的符号。（考虑x,y是否为标准化数据，关系又会有变化）

3. 回归问题的方差分析——衡量回归效果



$$s_{yy}=\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2=\frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i+e-\bar{y})^2=\frac{1}{n-1}\sum_{i=1}^{n}\left[(\hat{y}_i-\bar{y})+(y_i-\hat{y}_i)\right]^2$$
$$=\frac{1}{n-1}\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$$

s_{yy}称为总的方差，记$U=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2, Q=\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$，其中$U$被称为**“回归平方和”**，反映因子$x$的变化对$y$的贡献；$Q$被称为**“残差平方和”**(或**“剩余平方和”**)，反映除x以外的随机因素$e(e=y-\hat{y})$的影响。对于固定的样本容量n，U越大Q越小，或者说U占总方差的比重越大，表明回归效果越好。下面讨论一

下 $\frac{U}{Q+U}$,Q,U 的有关性质。

根据 U,Q 的表达式写出 $\frac{U}{Q+U}$，并做一定的变换如下：

$$\frac{U}{U+Q}=\frac{s_{\hat{y}}^2}{s_y^2}=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=\frac{\sum_{i=1}^{n}(b_0+bx_i-b_0-b\bar{x})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=b^2\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=b^2\frac{s_x^2}{s_y^2}$$

又根据回归系数与相关系数的关系：$b=r_{xy}\frac{s_y}{s_x}$，代入上式，可得关系式：

$$\frac{U}{U+Q}=r_{xy}^2,\frac{Q}{U+Q}=1-r_{xy}^2$$

基于上述关系，可以确定 U,Q 的计算方法，如下：

$$\begin{cases}U=S_{yy}r^2=S_{yy}\dfrac{S_{xy}^2}{S_{xx}S_{yy}}=\dfrac{S_{xy}}{S_{xx}}S_{xy}=bS_{xy}\\ Q=S_{yy}\left(1-r^2\right)=S_{yy}-bS_{xy}\end{cases}$$（大写的 S_{xy} 表示离差乘积和）

4. 回归模型的假设检验与区间估计

(1) 剩余方差的估计

$\varepsilon\sim N\left(0,\sigma^2\right)$，$\sigma^2$ 称为剩余方差，$\frac{Q}{\sigma^2}\sim\chi^2(n-2)\Rightarrow E\left(\frac{Q}{\sigma^2}\right)=n-2$，剩余方差 σ^2 的无偏估计量为：$\hat{\sigma}^2=\frac{Q}{n-2}$

(2) 回归系数 β 的区间估计和检验(*t* 检验)

$$\hat{\beta}=b=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}=\frac{\sum_{i=1}^{n}x_iy_i-\bar{x}\sum_{i=1}^{n}y_i}{S_{xx}}=\sum_{i=1}^{n}\frac{x_i-\bar{x}}{S_{xx}}y_i$$

$Var\left(\hat{\beta}\right)=Var\left(\sum_{i=1}^{n}\frac{x_i-\bar{x}}{S_{xx}}y_i\right)=\sum_{i=1}^{n}Var\left(\frac{x_i-\bar{x}}{S_{xx}}y_i\right)=\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{S_{xx}}\right)^2Var\left(y_i\right)$，此时 y_i

是一个随机变量，它的方差为$Var\left(y_i\right)=Var\left(y_i-\bar{y}\right)=Var\left(e\right)=\sigma^2$，从而上式可以继续化简为：

$$Var\left(\hat{\beta}\right)=\sigma^2\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{S_{xx}}\right)^2=\frac{\sigma^2}{S_{xx}}$$

因此有$\hat{\beta}\sim N\left(\beta,\frac{\sigma^2}{S_{xx}}\right)$，也即$\frac{\hat{\beta}-\beta}{\sqrt{\sigma^2/S_{xx}}}\sim N\left(0,1\right)$，根据(1)，显然有 t 分布：

$$\frac{\hat{\beta}-\beta}{\sqrt{\hat{\sigma}^2/S_{xx}}}\sim t\left(n-2\right)$$

据此可以解出β的置信水平为$1-\alpha$的区间估计为：$\left(\hat{\beta}\pm t_{\frac{\alpha}{2}}\frac{\hat{\sigma}}{\sqrt{S_{xx}}}\right)$

(3) 回归系数的显著性检验(*F* 检验)

回归方程的显著性检验，即检验两“总体”之间是否存在回归关系(*β*是否=0)

假设 H0： 总体的回归系数 β=0，使用的统计量：$F=\frac{U/1}{Q/n-2}\sim F\left(1,n-2\right)$

$$F=\frac{S_{yy}r^2}{S_{yy}\left(1-r^2\right)/\left(n-2\right)}=\frac{r^2}{1-r^2}n-2$$

相关系数检验时，有统计量$t=\frac{r}{\sqrt{1-r^2}}\sqrt{n-2}\sim t\left(n-2\right)$，因此可以说，**一元回归方程的检验等价于相关系数的检验**。

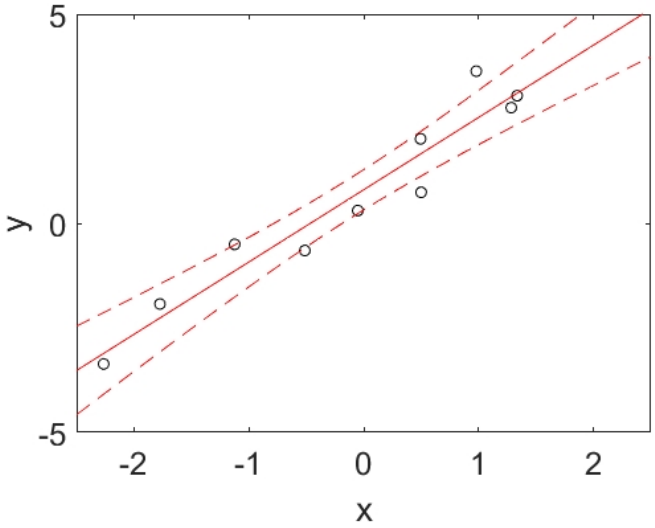
(4) 回归函数 $\beta_0+\beta x$ 函数值（ μ_y ）的估计

对于某一x_0，对应的y_0**的期望**的估计 $\widehat{\mu_{y_0}}$为： $\hat{\mu}_{y_0}=\hat{\beta}_0+\hat{\beta}x=\bar{y}+\hat{\beta}\left(x_0-\bar{x}\right)$，

采取和(3)一样的推导过程（$Var\left(\bar{y}\right)=Var\left(\bar{\mu}+\bar{\varepsilon}\right)=Var\left(\bar{\varepsilon}\right)=\frac{\sigma^2}{n}$），可以得到：

$$\frac{\hat{\mu}_{y_0}-\mu_{y_0}}{\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{\left(x_0-\bar{x}\right)^2}{S_{xx}}\right]}}\sim t\left(n-2\right)$$

由此解得置信区间为：$\left(\hat{\mu}_{y_0} \pm t_{\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2\left[\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]}\right)$，上式可知，区间宽度是$x_0$的函数，$x_0=\bar{x}$时最窄。即:**估计的精确性随着与$\bar{x}$距离的增大而变差**(如下图)。



(5) 新观测值 y_0 的估计

对于某一x_0，点估计值$\hat{y}_0=\hat{\beta}_0+\hat{\beta}x_0$，预测$y_0$应在以$\hat{y}_0$为中心的范围浮动，考虑其方差为$Var(y_0-\hat{y}_0)=Var(y_0)+Var(\hat{y}_0)$经过化简计算，得到：

$$\frac{y_0-\hat{y}_0}{\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]}}\sim t(n-2)$$

求得置信区间为$\left(b_0+bx_0 \pm t_{\frac{\alpha}{2}}\sqrt{\hat{\sigma}^2\left[1+\frac{1}{n}+\frac{(x_0-\bar{x})^2}{S_{xx}}\right]}\right)$

▲ 注意(4)和(5)的区别：虽然有$\hat{y}_0=\hat{\mu}_{y_0}=\hat{\beta}_0+\hat{\beta}x=b_0+bx_0$——这表达了两种含义：一种是对$y_0$的无偏点估计；一种是对$y_0$期望$\mu_{y_0}$的无偏点估计。对回归函数检验，是因为回归系数存在置信区间（此处是不包括ε的影响的）；而对新值的估计是因为新值是在回归函数值附近上下波动的，这个波动的产生就来自ε的影响。

二、多元回归模型

1. 多元回归模型概述

设随机变量 y 与 m 个变量 $x_1,x_2,\ldots,x_m$ 之间存在如下线性关系：

$$y=\beta_1x_1+\beta_2x_2+\cdots+\beta_mx_m+\beta_0+\varepsilon,\quad \varepsilon\sim N\left(0,\sigma^2\right)$$

该模型称为：多元线性回归模型，回归系数 β_i 也称为“偏回归系数”。

若对 y 进行了 n 次观测，得到一组观测值向量 $y=\left[y_1,y_2,\ldots,y_n\right]^{\mathrm{T}}$，根据上述的模型，可以观测值向量可以具体写为：

$$\begin{pmatrix}y_1\\y_2\\\vdots\\y_n\end{pmatrix}=\begin{pmatrix}\beta_1x_{11}+\beta_2x_{12}+\cdots+\beta_mx_{1m}+\beta_0+\varepsilon_1\\\beta_1x_{21}+\beta_2x_{22}+\cdots+\beta_mx_{2m}+\beta_0+\varepsilon_2\\\vdots\\\beta_1x_{n1}+\beta_2x_{n2}+\cdots+\beta_mx_{nm}+\beta_0+\varepsilon_n\end{pmatrix}\begin{pmatrix}x_{11}&x_{12}&\cdots&x_{1m}&1\\x_{21}&x_{22}&\cdots&x_{2m}&1\\\vdots&\vdots&\ddots&\vdots&1\\x_{n1}&x_{n2}&\cdots&x_{nm}&1\end{pmatrix}\begin{pmatrix}\beta_1\\\beta_2\\\vdots\\\beta_m\\\beta_0\end{pmatrix}+\begin{pmatrix}\varepsilon_1\\\varepsilon_2\\\vdots\\\varepsilon_n\end{pmatrix}$$

写成矩阵的形式为：$\boldsymbol{y}=\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$，在矩阵 $\boldsymbol{X}$ 中，**每一列代表一个变量的 n 次观测，每一行代表第 n 次观测所有变量的值**（这与第一章中提到的每一行代表一个变量的提法不同）。类比一元回归，对样本进行经验回归计算，得到回归模型为：

$$\hat{\boldsymbol{y}}=\boldsymbol{X}\hat{\boldsymbol{\beta}}\left(y=b_1x_1+b_2x_2+\cdots+b_mx_m+b_0\right)$$

（$\hat{\boldsymbol{\beta}}=\boldsymbol{b}=\left[b_1,b_2,\cdots,b_m,b_0\right]^{\mathrm{T}}$ 是对 $\boldsymbol{\beta}$ 的估计）

2. 多元回归模型的求解

根据最小二乘法原理，$\boldsymbol{b}$ 应使残差平方和 Q 达到最小，其中

$$Q=\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2=\left(\boldsymbol{y}-\boldsymbol{Xb}\right)^{\mathrm{T}}\left(\boldsymbol{y}-\boldsymbol{Xb}\right)=\boldsymbol{y}^{\mathrm{T}}\boldsymbol{y}-2\boldsymbol{b}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}+\boldsymbol{b}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Xb}$$

于是 $\dfrac{\partial Q}{\partial \boldsymbol{b}}=-2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}+2\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Xb}=0$，解得 $\boldsymbol{b}=\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$。需要指出的是，推导过程中用的是原始资料阵 $\boldsymbol{X}$，所以上述解向量中，$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}_{(m\times m)}$ 表示的是 $m+1$ 个因子的交叉乘积阵，$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}_{(m\times 1)}$ 表示的是 $m+1$ 个因子与 y 的交叉乘积向量。为了能与协方差建立联系，常常使用距平资料阵，这样，$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$ 就可以表示离差乘积阵，$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y}$ 为离差乘积向量。下面在距平资料的条件下，给出相对应的分量形式如下：

$$\begin{cases} S_{11}b_1+S_{12}b_2+\cdots+S_{1m}b_m=S_{1y} \\ S_{21}b_1+S_{22}b_2+\cdots+S_{2m}b_m=S_{2y} \\ \cdots\cdots \\ S_{m1}b_1+S_{m2}b_2+\cdots+S_{mm}b_m=S_{my} \end{cases} \xrightarrow{\text{除以}\frac{1}{n-1}} \begin{cases} s_{11}b_1+s_{12}b_2+\cdots+s_{1m}b_m=s_{1y} \\ s_{21}b_1+s_{22}b_2+\cdots+s_{2m}b_m=s_{2y} \\ \cdots\cdots \\ s_{m1}b_1+s_{m2}b_2+\cdots+s_{mm}b_m=s_{my} \end{cases}$$

上述方程称为求解 $b_i(i=1,2,\ldots,m)$的**正规方程组**，因此在不使用计算机自带回归函数求解时，可以先求出各因子与预报量之间的协方差以及各因子之间的协方差阵，解正规方程组即可。

3. 不同观测资料的回归方程

(1) 原始观测资料 $\hat{y}=b_0+b_1x_1+b_2x_2+\cdots+b_mx_m$

(2) 距平观测资料 $\hat{y}=b_1x_1+b_2x_2+\cdots+b_mx_m$

▲ 原始资料与距平资料的回归系数都一样，区别在于有无常数 b_0

(3) 标准化变量 $\hat{y}=b_1^*x_1+b_2^*x_2+\cdots+b_m^*x_m$

对正规方程组的第 i 个方程 $s_{i1}b_1+s_{i2}b_2+\cdots+s_{im}b_m=s_{iy}$ 变形如下：

$$\frac{s_{i1}}{\sqrt{s_{ii}}\sqrt{s_{11}}}\frac{\sqrt{s_{11}}}{\sqrt{s_{yy}}}b_1+\frac{s_{i2}}{\sqrt{s_{ii}}\sqrt{s_{22}}}\frac{\sqrt{s_{22}}}{\sqrt{s_{yy}}}b_2+\cdots+\frac{s_{im}}{\sqrt{s_{ii}}\sqrt{s_{mm}}}\frac{\sqrt{s_{mm}}}{\sqrt{s_{yy}}}b_m=\frac{s_{iy}}{\sqrt{s_{ii}}\sqrt{s_{yy}}}$$

而标准化变量的协方差就是相关系数，因此上述方程又可写为：

$$r_{i1}\frac{\sqrt{s_{11}}}{\sqrt{s_{yy}}}b_1+r_{i2}\frac{\sqrt{s_{22}}}{\sqrt{s_{yy}}}b_2+\cdots+r_{im}\frac{\sqrt{s_{mm}}}{\sqrt{s_{yy}}}b_m=r_{iy}$$

类比可以得到，标准化变量的回归系数为：$b_j^*=\frac{\sqrt{s_{jj}}}{\sqrt{s_{yy}}}b_j\ (j=1,2,\ldots,m)$（这一表达式也给出了标准化数据和原始场数据之间的关系）

4. 多元回归的方差分析·复相关系数

(1) 回归平方和：$U=\sum_{i=1}^{m}b_iS_{iy}$；剩余平方和：$Q=S_{yy}-\sum_{i=1}^{m}b_iS_{iy}$

(2) σ^2 的估计——剩余方差 $\widehat{\sigma^2}=\frac{Q}{n-m-1}$（**剩余方差越小，回归效果越好**）

(3) 复相关系数

y 与 $\hat{y}$ 之间的相关系数，称为复相关系数 R，表达式为：$R^2=\frac{U}{S_{yy}}=1-\frac{Q}{S_{yy}}$

R^2 又称为回归方程的**决定系数**（注意不要把决定系数和复相关系数弄混）。当 n 和 m 固定时，**复相关系数越大，表明回归效果越好**。

▲ 调整后的决定系数：$R_a^2=1-\frac{Q/(n-m-1)}{S_{yy}/(n-1)}=1-\left(\frac{n-1}{n-m-1}\right)(1-R^2)$

5. 多元回归模型的显著性检验和区间估计

(1) 回归方程的显著性检验

检验回归方程的效果，即检验 y 与 $x_1,x_2,...,x_m$ 之间是否存在线性关系。可归结为检验以下原假设：$H_0:\beta_1=\beta_2=\cdots=\beta_m$，使用的统计量为：

$$F=\frac{\frac{U}{m}}{\frac{Q}{n-m-1}}\sim F(m,\ n-m-1)$$

F 值越大（**注意：这里是单侧检验**），表明回归平方和 U 越大(相对 Q)，越应拒绝原假设。因此，若 $F>F_\alpha$，拒绝原假设，认为回归效果显著，即认为 y 与各预报因子之间存在线性关系(回归系数不全为零)。

例如，有三种不同类型的回归方程，通过计算 U、Q 得到 F 的值，进而得到相应的显著性水平 α。那么，α 的值大小如何体现回归效果呢？对于给定的 α，当 $F>F_\alpha$ 时拒绝原假设，也就是说，F 值越大，越能够通过显著性检验，从而拒绝原假设，而 F 值越大，等效为对应的 α 值越小，所以，在以“回归方程的显著性水平”为指标时，显著性水平越小，回归效果越好。

▲ **临界复相关系数**

$$F=\frac{\frac{R^2}{m}}{\frac{1-R^2}{n-m-1}}\Rightarrow R_c=\sqrt{\frac{mF_\alpha}{mF_\alpha+(n-m-1)}}$$

当 $R>R_c$ 时，即 $F>F_\alpha$，表明回归方程显著。

(2) 单个因子的方差贡献和显著性检验

- 单个因子的方差贡献率：$RC_i=\dfrac{b_iS_{iy}}{S_{yy}}$（RC：Rate of Contribution）

- 在没有引入因子前，回归平方和为 0，即 $S_{yy}=Q$。当引入因子后，U 开始增大，而 Q 会减小。因此在剔除一个因子后，U 一定会减小，Q 一定会增大。假设 m 个因子的回归方程的残差平方和为 Q_m，回归系数为 $\boldsymbol{b}$，删掉第 i 因子后新方程的残差平方和为 Q_{m-1}，回归系数为 $\boldsymbol{b}'$，于是，第 i 个因子的方差贡献为：

$$\boldsymbol{V}_i=Q_{m-1}-Q_m=U_m-U_{m-1}=\sum_{j=1}^{m}b_jS_{jy}-\sum_{j=1,j\neq i}^{m}b_j'S_{jy}$$

上式比较复杂，常采用以下公式计算：

$$V_i=\frac{b_i^2}{c_{ii}}$$

其中，b_i 为因子对应的回归系数，c_{ii} 为 $\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}$ 的对角线元素。

▲ **[注意]** $V_i=\dfrac{b_i^2}{c_{ii}}$ **是方差贡献而非方差贡献率，它表示的剔除了第 i 个因子以后，回归平方和的增量(ΔU)或者剩余平方和的增量(ΔQ)。**

- 单个因子的显著性检验

原假设 $H_0:\beta_i=0$；统计量：$F_i=\dfrac{V_i/1}{Q/(n-m-1)}\sim F(1,n-m-1)$

(3) 回归系数、回归函数、新观测值的区间估计（直接给出结论）

※ 各回归系数：$\dfrac{\hat{\beta}_i-\beta_i}{\sqrt{\hat{\sigma}^2c_{ii}}}\sim t(n-m-1)\Rightarrow\left[\hat{\beta}_i\pm t_{\alpha/2}\sqrt{\hat{\sigma}^2c_{ii}}\right]$

※ 回归函数：$\left[\hat{y}_0\pm t_{\alpha/2}\sqrt{\hat{\sigma}^2\boldsymbol{x}_0^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{x}}\right]$

※ 新值观测：$\left[\hat{y}_0 \pm t_{\alpha/2}\sqrt{\hat{\sigma}^2\left(1+\boldsymbol{x}_0^{\mathrm{T}}\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\right)^{-1}\boldsymbol{x}_0\right)}\right]$

6. 多重共线性问题

(1) 多重共线性问题

已知回归系数的方差$Var\left(\hat{\beta}_i\right)=\sigma^2 c_{ii}$，因此当$c_{ii}$较大时，将导致回归系数估计的精确度降低，不确定性增大（因为置信区间变宽了）。下面讨论c_{ii}与什么因素有关。以二元回归为例，假设所用数据均已经标准化，于是正规方程组的离差乘积阵就是相关系数阵：$\begin{bmatrix}1 & r\\ r & 1\end{bmatrix}\left(r_{12}=r_{21}=r\right)$，对应的逆矩阵为：$\begin{bmatrix}\frac{1}{1-r^2} & \frac{-r}{1-r^2}\\ \frac{-r}{1-r^2} & \frac{1}{1-r^2}\end{bmatrix}$，于是$c_{ii}=\frac{1}{1-r^2}$，也就是说，**当两个因子之间存在很强的相关性时(r 很大)，c_{ii} 的值也很大**。

对于任意多因子回归方程，可证明$c_{ii}=\frac{1}{1-R_i^2}$，其中，R_i^2为x_i对其他所有因子做回归时的决定系数。因此，**如果x_i与其它因子的任何子集之间存在着近似线性关系**，则有R_i^2和c_{ii}很大，导致回归系数$\hat{\beta}_i$估计的不确定性很大，这时称回归方程存在“**多重共线性**”问题。

(2) “多重共线性”问题的定量判断

记矩阵$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$的最大特征值与最小特征值之比为$\kappa=\frac{\lambda_{\max}}{\lambda_{\min}}$，判断准则如下：

- 当$\kappa<100$时不存在多重共线性问题；
- 当$100<\kappa<1000$时，存在中等程度多重共线性；
- 当$\kappa>1000$时，存在严重多重共线性。

7. 偏相关系数（以二元回归为例）

由于因子之间存在一定的关系，所以会出现求出的单个因子 x_i 与预报量 y 之间的相关系数与观测数据的变化规律不相符合的情况。为了准确说明单个因子与预报量真正的（简单）相关系数，引入偏相关系数的概念。

设有数据 x_1, x_2, y，下面讨论如何从 x_1 和 y 中扣除 x_2 的影响。以 x_1 和 y 为因变量，以 x_2 为自变量建立两个一元回归模型，获得其残差为：

$$e_x = x_1 - \hat{x}_1 = x_1 - (bx_2 + b_0)$$
$$e_y = y - \hat{y} = y - (cx_2 + c_0)$$

此时的 e_x, e_y 即为扣除 x_2 影响之后的 x_1 和 y，通过计算 e_x, e_y 之间的相关系数，能够真正地反应 x_1 对 y 的影响，这一相关系数便称为偏相关系数。

三、逐步回归法

1. 系数矩阵的求逆与正规方程组的求解

例如，已知 $\boldsymbol{S} = \begin{bmatrix} 10 & 7 & 4 \\ 7 & 7 & 3 \\ 4 & 3 & 4 \end{bmatrix}$，$\boldsymbol{s}_y = \begin{bmatrix} 4 \\ 4 \\ 3 \end{bmatrix}$，求解正规方程组 $\boldsymbol{Sb} = \boldsymbol{s}_y$.

(1) 求解逆矩阵 $\boldsymbol{S}^{-1}$。求解逆矩阵 $\boldsymbol{S}^{-1}$ 有必要，因为在多元回归问题求单个因子方差的贡献会用到。在不借助计算机的前提下，采用矩阵初等行变化的方法：

$$\left[\begin{array}{ccc:ccc} 10 & 7 & 4 & 1 & 0 & 0 \\ 7 & 7 & 3 & 0 & 1 & 0 \\ 4 & 3 & 4 & 0 & 0 & 1 \end{array}\right] \xrightarrow{\text{初等行变换}} \left[\begin{array}{ccc:ccc} 1 & 0 & 0 & -0.38 & -0.32 & -0.14 \\ 0 & 1 & 0 & -0.32 & 0.48 & -0.04 \\ 0 & 0 & 1 & -0.14 & -0.04 & 0.42 \end{array}\right]$$

(2) 正规方程组的求解，依然采用初等行变换。

$$\left[\begin{array}{ccc:c} 10 & 7 & 4 & 4 \\ 7 & 7 & 3 & 4 \\ 4 & 3 & 4 & 3 \end{array}\right] \xrightarrow{\text{初等行变换}} \left[\begin{array}{ccc:c} 1 & 0 & 0 & -0.18 \\ 0 & 1 & 0 & 0.52 \\ 0 & 0 & 1 & 0.54 \end{array}\right]$$

2. 紧凑型求解与求逆变换(将 1 中的求解与求逆过程合并以及节省空间)

求解求逆的计算过程中，每做一次变换，原系数阵就有一列变为单位阵元素并不再改变，而逆矩阵中则减少一列单位元素，即： 总矩阵中总有三列保持单

位阵元素。为了节省空间，在做变换时，可将系数阵中的单位阵元素用逆矩阵中对应列来代替，这就是紧凑型求解与求逆变换

$$A^{(0)}=\begin{bmatrix}10 & 7 & 4 & 4 & 1 & 0 & 0\\ 7 & 7 & 3 & 4 & 0 & 1 & 0\\ 4 & 3 & 4 & 3 & 0 & 0 & 1\end{bmatrix}$$

$$A^{(1)}=\begin{bmatrix}1 & 0.7 & 0.4 & 0.4 & 0.1 & 0 & 0\\ 0 & 2.1 & 0.2 & 1.2 & -0.7 & 1 & 0\\ 0 & 0.2 & 2.4 & 1.4 & -0.4 & 0 & 1\end{bmatrix}$$

$$A^{(2)}=\begin{bmatrix}1 & 0 & 0.33 & -0.0 & 0.33 & -0.33 & 0\\ 0 & 1 & 0.095 & 0.57 & -0.33 & 0.48 & 0\\ 0 & 0 & 2.38 & 1.29 & -0.33 & -0.095 & 1\end{bmatrix}$$

$$\underset{A^{(0)}}{\begin{bmatrix}10 & 7 & 4 & 4\\ 7 & 7 & 3 & 4\\ 4 & 3 & 4 & 3\end{bmatrix}}\Rightarrow\underset{A^{(1)}}{\begin{bmatrix}0.1 & 0.7 & 0.4 & 0.4\\ -0.7 & 2.1 & 0.2 & 1.2\\ -0.4 & 0.2 & 2.4 & 1.4\end{bmatrix}}\Rightarrow\underset{A^{(2)}}{\begin{bmatrix}0.33 & -0.33 & 0.33 & -0.00\\ -0.33 & 0.48 & 0.095 & 0.57\\ -0.33 & -0.095 & 2.38 & 1.29\end{bmatrix}}\Rightarrow\underset{A^{(3)}}{\begin{bmatrix}0.38 & -0.32 & -0.14 & -0.18\\ -0.32 & 0.48 & -0.040 & 0.52\\ -0.14 & -0.04 & 0.42 & 0.54\end{bmatrix}}$$

紧凑型求解求逆过程遵守如下运算法则：

本身：$a^{(l)}(k,k)=1/a^{(l-1)}(k,k)$

同行：$a^{(l)}(k,j)=a^{(l-1)}(k,j)/a^{(l-1)}(k,k)\qquad(j\neq k)$

同列：$a^{(l)}(i,k)=-a^{(l-1)}(i,k)/a^{(l-1)}(k,k)\qquad(i\neq k)$

其他：$a^{(l)}(i,j)=a^{(l-1)}(i,j)-\dfrac{a^{(l-1)}(i,k)a^{(l-1)}(k,j)}{a^{(l-1)}(k,k)}\qquad(i\neq k,j\neq k)$

(结合上述的变换过程很容易可以理解运算法则，且可以进行恢复运算)

3. 逐步回归

(1) 准备工作

先对 m 个备选因子与 y (共 m+1 个变量)进行标准化，如有 n 次观测，可记为 n 行(m+1)列矩阵 $\boldsymbol{Z}_{n\times(m+1)}=[\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_m,\boldsymbol{y}]$，计算其离差乘积阵 $\boldsymbol{S}^{(0)}$ 为：

$$\boldsymbol{S}^{(0)}=\begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1m} & S_{1y} \\ S_{21} & S_{22} & \cdots & S_{2m} & S_{2y} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ S_{m1} & S_{m2} & \cdots & S_{mm} & S_{my} \\ S_{y1} & S_{y2} & \cdots & S_{ym} & S_{yy} \end{bmatrix}$$

[重要说明]

▲ 上标表示当前回归方程中含有的因子个数。当对 S_{kk} **(1≤*k*≤*m*)**进行变换时，表示把第 ***k*** 因子引入；

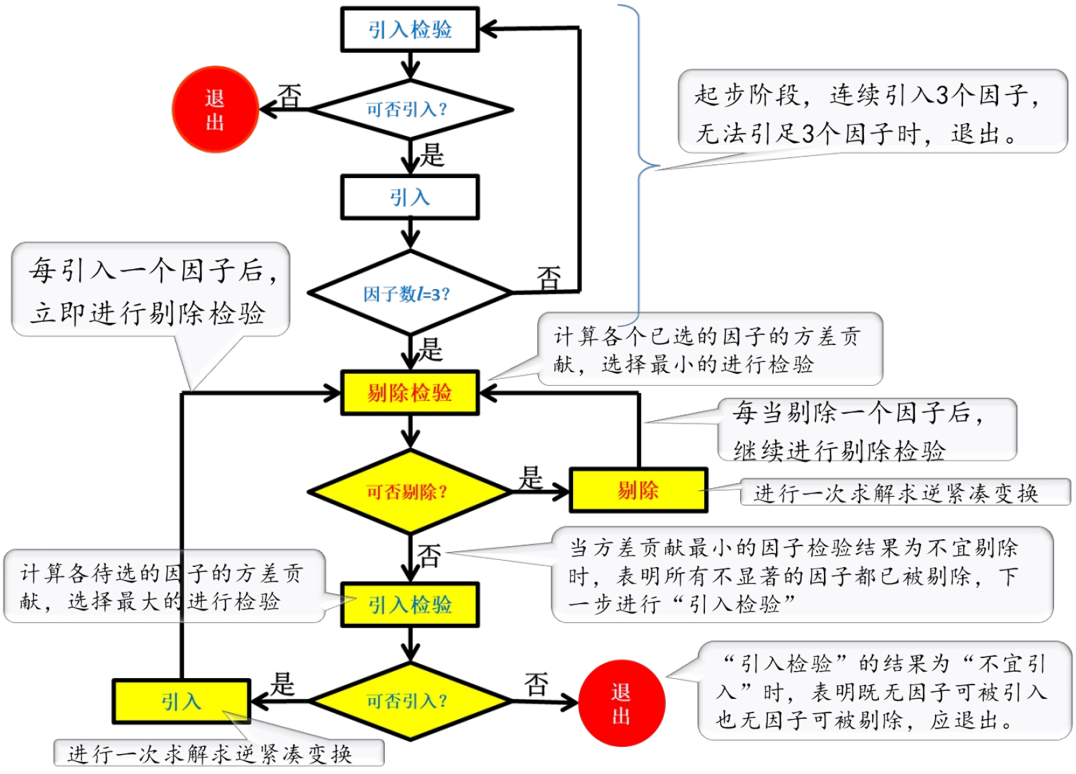
▲ 通过观察 S_{ky} 和 S_{yk} 的对称性来判断当前第 ***k*** 个因子是否已被引入方程：当 $S_{ky}=S_{yk}$ 时，第 ***k*** 因子未被引入；当 $S_{ky}=-S_{yk}$ 时，表明第 ***k*** 因子已被引入；

▲ S_{yy}=*n*-1 = $\sum_{i=1}^{n}(y_i-\bar{y})^2$，即右下角元素 $\boldsymbol{S}_{yy}$ 代表 ***y*** 的离差平方和；

▲ S_{yy} 为未引入因子时离差平方和，因此就是剩余方差；

▲ 只有在初始矩阵中，最后一列（行）才表示离差乘积和，一旦引入因子做了紧凑变换以后，最后一列便不再有离差乘积的性质。

(2) 逐步回归的步骤（程序框图呈现）



(3) 引入检验与剔除检验的比较与总结

❑**引入检验**：当前已引入l个因子，即将引入第l+1个，此时的$\boldsymbol{S}$矩阵为：

在m-l个待选因子中，因子x_i的方差贡献$V_i^{(l+1)}$：

$$V_i^{(l+1)}=\frac{[S_{iy}^{(l)}]^2}{S_{ii}^{(l)}}$$

$V_i^{(l+1)}$表示把x_i引入后(回归方程将有$l+1$个因子)，U的增量。

$$\boldsymbol{S}^{(l)}=\begin{bmatrix} S_{11}^{(l)} & \cdots & S_{1m}^{(l)} & S_{1y}^{(l)} \\ \vdots & \ddots & \vdots & \vdots \\ S_{i1}^{(l)} & \cdots & S_{ii}^{(l)} & S_{iy}^{(l)} \\ \vdots & \ddots & \vdots & \vdots \\ S_{m1}^{(l)} & \cdots & S_{mm}^{(l)} & S_{my}^{(l)} \\ S_{y1}^{(l)} & \cdots & S_{ym}^{(l)} & S_{yy}^{(l)} \end{bmatrix}$$

选择贡献最小的因子x_k,计算F统计量：

$$F_{\text{引入}}=\frac{V_k^{(l+1)}/1}{Q^{(l+1)}/(n-(l+1)-1)}=\frac{V_k^{(l+1)}/1}{(S_{yy}^{(l)}-V_k^{(l+1)})/(n-(l+1)-1)}$$

$S_{yy}^{(l)}$表示当前回归方程(l个因子,无x_i)的$Q^{(l)}$，引入x_i后将有$Q^{(l+1)}=S_{yy}^{(l)}-V_k^{(l+1)}$，方程包含$l$+1个因子。

❑**剔除检验**：

当前已引入了l个因子，要选择一个贡献最弱的因子剔除，此时的$\boldsymbol{S}$矩阵为：

在l个已选因子中(包含x_i)，考察x_i的贡献$V_i^{(l)}$：

$$V_i^{(l)}=\frac{[S_{iy}^{(l)}]^2}{S_{ii}^{(l)}}$$

$V_i^{(l)}$表示x_i在当前回归方程(共l个因子)中的方差贡献，即把x_i剔除后剩余平方和的增量。

$$\boldsymbol{S}^{(l)}=\begin{bmatrix} S_{11}^{(l)} & \cdots & S_{1m}^{(l)} & S_{1y}^{(l)} \\ \vdots & \ddots & \vdots & \vdots \\ S_{i1}^{(l)} & \cdots & S_{ii}^{(l)} & S_{iy}^{(l)} \\ \vdots & \ddots & \vdots & \vdots \\ S_{m1}^{(l)} & \cdots & S_{mm}^{(l)} & S_{my}^{(l)} \\ S_{y1}^{(l)} & \cdots & S_{ym}^{(l)} & S_{yy}^{(l)} \end{bmatrix}$$

选择贡献最小的因子x_k,计算F统计量：

$$F_{\text{剔除}}=\frac{V_k^{(l)}/1}{Q^{(l)}/(n-l-1)}=\frac{V_k^{(l)}/1}{S_{yy}^{(l)}/(n-l-1)}$$

回顾“单个因子显著性”的检验，F统计量的公式完全相同。

$S_{yy}^{(l)}$表示当前回归方程(l个因子,含x_i)的$Q^{(l)}$

[强调]

▲ 矩阵右下角元素$S_{yy}^{(l)}$始终代表当前回归方程(l个因子)的剩余平方和，即：$Q^{(l)}=S_{yy}^{(l)}$。当引入新因子后，剩余平方和为：$Q^{(l+1)}=Q^{(l)}-V_i^{(l+1)}=S_{yy}^{(l)}-V_i^{(l+1)}$

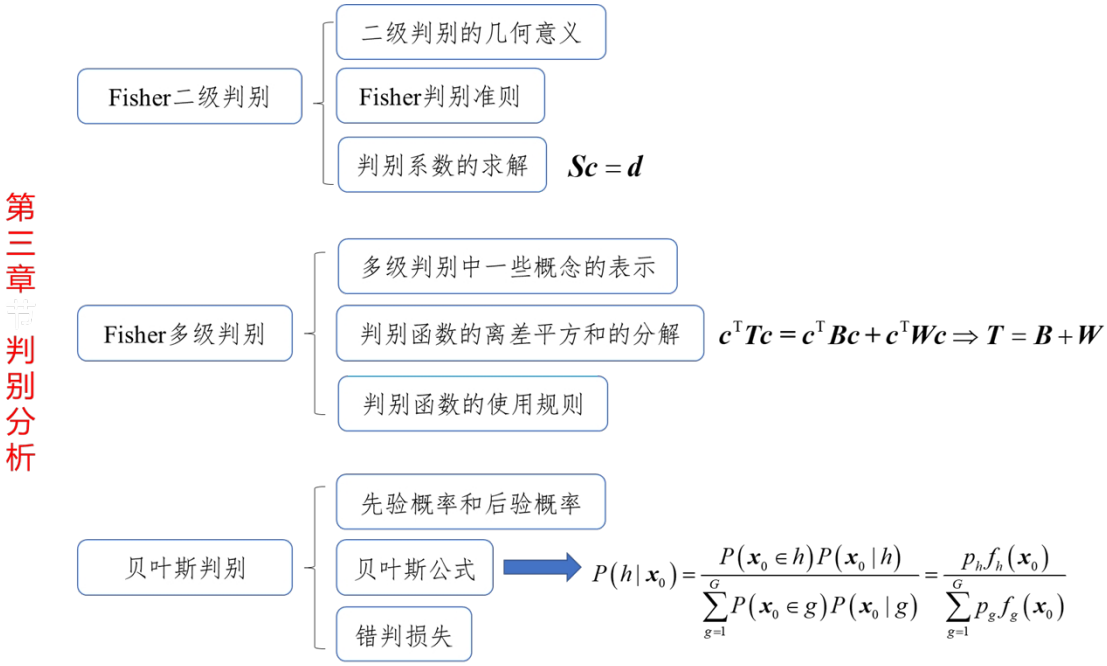
（其中$V_i^{(l+1)}$为新引入一个因子后回归平方和的增量）

▲ 引入$l\left(k_1,k_2,...,k_l\right)$个因子时，当前矩阵为$\boldsymbol{S}^{(l)}$，那么回归系数为矩阵最后一列对应元素的值：$b_{k_1}^*=S_{k_1y}^{(l)},b_{k_2}^*=S_{k_2y}^{(l)},...,b_{k_3}^*=S_{k_ly}^{(l)}$

▲ 因子的剔除需要连续剔除，直至所有不显著的因子都被剔除，但因子的引入却不能连续引入（起步阶段除外）。每引入一个因子，就进行剔除检验。

第三章 判别分析

➢ *章节思维导图*



一、 Fisher 二级判别

1. 二级判别的理解

二级判别是指预报对象只有 2 种**类别**的情况，例如“有雨”和“无雨”。预测类别，仍需先寻找因子，且寻找的因子一般都是数值型。因此，可以根据历史资料的观测，将数据分成两大类别，二级判别的关键，就是要找到一条**判别线**，使得在拿到一组因子的观测值时，能够根据判别线预测结果属于哪一类。下面以包含两个因子的情形来进一步理解二级判别模型。

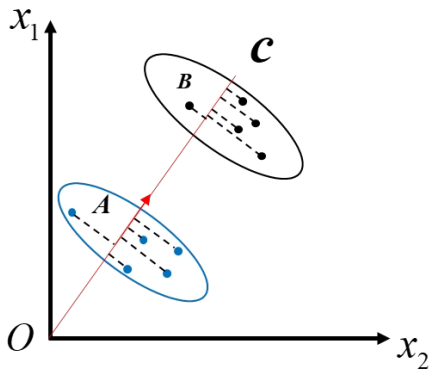
设有两个因子 x_1, x_2 和预报因子 y，则判别模型（判别线）的表达式可以写 $y = c_1x_1 + c_2x_2 = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}$，$\boldsymbol{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}, \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$，其中 $\boldsymbol{y}$ **称为判别函数，** $\boldsymbol{c_1}$ **与** $\boldsymbol{c_2}$ **为判别系数，** 列向量 $\boldsymbol{x}=[x_1, x_2]^{\mathrm{T}}$ 通常表示两因子变量的某一次观测，也称为一个“**样品**”。如果向量 $\boldsymbol{c}=[c_1, c_2]^{\mathrm{T}}$ 已知，那么对于任意一个样品 $\boldsymbol{x_0}=[a, b]^{\mathrm{T}}$，可得一个判别函数值 y_0，需要注意的是，因子 x_0 是数值型变量，所以 y_0 也是“数值”型变量，因此需设定一个判别指标 y_c，把 $y_0 > y_c$ 和 $y_0 < y_c$ 定义为不同的类别。

2. Fisher 二级判别系数的求解

(1) Fisher 判别准则

我们仍然先对二维的情形考虑（即只含有两个因子），判别函数为：

$$y = c_1x_1 + c_2x_2 = \boldsymbol{c}^{\mathrm{T}}\boldsymbol{x},\ \ \boldsymbol{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}, \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$



上述判别函数可以看作是两个向量的点乘，于是，对于每一组(x_1,x_2)都可以看成是在(c_1,c_2)方向上的投影（即$|\boldsymbol{x}|\cos(\widehat{\boldsymbol{x},\boldsymbol{c}})$），如上图所示。要使两个类别尽可能的清楚，需要找到这样一个向量$\boldsymbol{c}$，使得 A 类（或者 B 类）组内的点在$\boldsymbol{c}$方向上的投影点应尽可能近，同时 A 和 B 两类之间的距离应尽可能远。由此，我们给出 Fisher 判别准则如下：

① 组间(between-group)距离尽可能大，记作$S_B = \left[\bar{y}(A) - \bar{y}(B)\right]^2$；

② 组内(within-group)距离尽可能小，记作

$$S_W = \sum_{t=1}^{n_1}\left[y_t(A) - \bar{y}(A)\right]^2 + \sum_{t=1}^{n_2}\left[y_t(B) - \bar{y}(B)\right]^2$$

③ 将第一条和第二条规则合并为一条则，即$\lambda = \dfrac{S_B}{S_W}$达到最大。

(2) 判别系数的求解

对于 m 个因子，设判别函数为：$y = c_1x_1 + c_2x_2 + \cdots + c_mx_m$，将观测资料整理成为 m 行 n 列的原始阵，于是原始阵的每一列（代表一次观测）都称为一个“样品”。我们把原始资料阵分为 A 和 B 两大类，相应的原始阵也分为$\boldsymbol{X}(A)$和

$\boldsymbol{X}(B)$，容量分别为$n_1, n_2\ (n_1+n_2=n)$，如下：

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} = \begin{bmatrix} x_{11}(A) & x_{12}(A) & \cdots & x_{1n_1}(A) \\ x_{21}(A) & x_{22}(A) & \cdots & x_{2n_1}(A) \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1}(A) & x_{m2}(A) & \cdots & x_{mn_1}(A) \end{bmatrix} + \begin{bmatrix} x_{11}(B) & x_{12}(B) & \cdots & x_{1n_2}(B) \\ x_{21}(B) & x_{22}(B) & \cdots & x_{2n_2}(B) \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1}(B) & x_{m2}(B) & \cdots & x_{mn_2}(B) \end{bmatrix}$$

将两类的每一个样品代入到判别函数函数中，可以得到两组向量为：

$$\boldsymbol{y}(A) = \left[y_1(A), y_2(A), ..., y_{n_1}(A) \right]^{\mathrm{T}}$$
$$\boldsymbol{y}(B) = \left[y_1(B), y_2(B), ..., y_{n_2}(B) \right]^{\mathrm{T}}$$

对于第 t 次观测，对 A 和 B 两类，有：

$$y_t(A) = \boldsymbol{c}^{\mathrm{T}} \boldsymbol{x}_t(A) = \sum_{k=1}^{m} c_k x_{kt}(A)$$
$$y_t(B) = \boldsymbol{c}^{\mathrm{T}} \boldsymbol{x}_t(B) = \sum_{k=1}^{m} c_k x_{kt}(B)$$

结合 Fisher 准则和极值原理，

$$\lambda = \frac{S_B}{S_W} \text{最大} \Leftrightarrow \frac{\partial \lambda}{\partial c_k} = \frac{S_B \dfrac{\partial S_W}{\partial c_k} - S_W \dfrac{\partial S_B}{\partial c_k}}{S_W^2} = 0 \Rightarrow \frac{1}{\lambda} \frac{\partial S_B}{\partial c_k} = \frac{\partial S_W}{\partial c_k}$$

根据建立的观测资料，有

$$S_B = \left[\bar{y}(A) - \bar{y}(B) \right]^2, S_W = \sum_{t=1}^{n_1} \left[y_t(A) - \bar{y}(A) \right]^2 + \sum_{t=1}^{n_2} \left[y_t(B) - \bar{y}(B) \right]^2$$

其中 $\bar{y}(A), \bar{y}(B)$ 分别为向量 $\boldsymbol{y}(A), \boldsymbol{y}(B)$ 的均值，表示方法如下：

$$\bar{y}(A) = \frac{1}{n_1} \sum_{t=1}^{n_1} y_t(A) = \frac{1}{n_1} \sum_{t=1}^{n_1} \boldsymbol{c}^{\mathrm{T}} \boldsymbol{x}_t(A) = \frac{1}{n_1} \sum_{t=1}^{n_1} \left(\sum_{k=1}^{m} c_k x_{kt}(A) \right) = \sum_{k=1}^{m} c_k \left(\frac{1}{n_1} \sum_{t=1}^{n_1} x_{kt}(A) \right) = \sum_{k=1}^{m} c_k \overline{x_k}(A)$$

$$\bar{y}(B) = \frac{1}{n_2} \sum_{t=1}^{n_2} y_t(B) = \frac{1}{n_2} \sum_{t=1}^{n_2} \boldsymbol{c}^{\mathrm{T}} \boldsymbol{x}_t(B) = \frac{1}{n_2} \sum_{t=1}^{n_2} \left(\sum_{k=1}^{m} c_k x_{kt}(B) \right) = \sum_{k=1}^{m} c_k \left(\frac{1}{n_2} \sum_{t=1}^{n_2} x_{kt}(B) \right) = \sum_{k=1}^{m} c_k \overline{x_k}(B)$$

而 $y_t(A) = \boldsymbol{c}^{\mathrm{T}} \boldsymbol{x}_t(A) = \sum_{k=1}^{m} c_k x_{kt}(A), y_t(B) = \boldsymbol{c}^{\mathrm{T}} \boldsymbol{x}_t(B) = \sum_{k=1}^{m} c_k x_{kt}(B)$ 分别表示第 t 次观测，A 和 B 两类的判别函数值。在明白各个符号所对应的表达式后，开始进行求偏导计算。

$$\frac{\partial S_B}{\partial c_k}=\frac{\partial}{\partial c_k}\left(\left[\bar{y}(A)-\bar{y}(B)\right]^2\right)=2\left[\bar{y}(A)-\bar{y}(B)\right]\left[\frac{\partial\bar{y}(A)}{\partial c_k}-\frac{\partial\bar{y}(B)}{\partial c_k}\right]=2\left[\overline{x_k}(A)-\overline{x_k}(B)\right]\left(\sum_{k=1}^{m}c_k\left[\overline{x_k}(A)-\overline{x_k}(B)\right]\right)$$

$$\frac{\partial S_W}{\partial c_k}=\sum_{t=1}^{n_1}2\left[y_t(A)-\bar{y}(A)\right]\left[\frac{\partial y_t(A)}{\partial c_k}-\frac{\partial\bar{y}(A)}{\partial c_k}\right]+\sum_{t=1}^{n_2}2\left[y_t(B)-\bar{y}(B)\right]\left[\frac{\partial y_t(B)}{\partial c_k}-\frac{\partial\bar{y}(B)}{\partial c_k}\right]$$

$$=\sum_{t=1}^{n_1}2\left[\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}_t(A)-\boldsymbol{c}^{\mathrm{T}}\bar{\boldsymbol{x}}(A)\right]\left[x_{kt}(A)-\bar{x}_k(A)\right]+\sum_{t=1}^{n_2}2\left[\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}_t(B)-\boldsymbol{c}^{\mathrm{T}}\bar{\boldsymbol{x}}(B)\right]\left[x_{kt}(B)-\bar{x}_k(B)\right]$$

$$=2\sum_{t=1}^{n_1}\left\{c_1\left[x_{1t}(A)-\bar{x}_1(A)\right]\left[x_{kt}(A)-\bar{x}_k(A)\right]+\cdots+c_1\left[x_{mt}(A)-\bar{x}_m(A)\right]\left[x_{kt}(A)-\bar{x}_k(A)\right]\right\}$$

$$+2\sum_{t=1}^{n_2}\left\{c_1\left[x_{1t}(B)-\bar{x}_1(B)\right]\left[x_{kt}(B)-\bar{x}_k(B)\right]+\cdots+c_1\left[x_{mt}(B)-\bar{x}_m(B)\right]\left[x_{kt}(B)-\bar{x}_k(B)\right]\right\}$$

在求偏导后式子比较复杂，我们定义如下的标记：

- 记 $d_k=\bar{x}_k(A)-\bar{x}_k(B)$ 表示表示 A 类第 k 个因子的平均值与 B 类第 k 个因子的平均值之差，那么 $\frac{\partial S_B}{\partial c_k}$ 可以表示为：

$$\frac{\partial S_B}{\partial c_k}=2d_k\left(c_1d_1+c_2d_2+\cdots+c_md_m\right)$$

- 记 $s_{ik}(A)=\sum_{t=1}^{n_1}[x_{it}(A)-\bar{x}_i(A)][x_{kt}(A)-\bar{x}_k(A)]$ 表示A类中 x_i 与 x_k 的离差乘积（注：这里的下标 i 的含义为因子数目），$s_{ik}(B)=\sum_{t=1}^{n_2}[x_{it}(B)-\bar{x}_i(B)][x_{kt}(B)-\bar{x}_k(B)]$ 表示B类中 x_i 与 x_k 的离差乘积，再令 $s_{ik}=s_{ik}(A)+s_{ik}(B)$ 表示 A 和 B 类总的离差乘积和。于是 $\frac{\partial S_W}{\partial c_k}$ 可以表示为：

$$\frac{\partial S_W}{\partial c_k}=2\left(c_1s_{1k}+c_2s_{2k}+\cdots+c_ms_{mk}\right)$$

于是第 k 个方程可以写为：

$$c_1s_{k1}+c_2s_{k2}+\ldots+c_ms_{km}=\frac{1}{\lambda}(c_1d_1+c_2d_2+\ldots+c_md_m)d_k$$

对 $k=1,2,...,m$，可得到方程组如下：

$$\begin{cases}s_{11}c_1+s_{12}c_2+\ldots+s_{1m}c_m=\beta d_1\\ s_{21}c_1+s_{22}c_2+\ldots+s_{2m}c_m=\beta d_2\\ \ldots\quad\ldots\quad\ldots\quad\ldots\quad\ldots\\ s_{m1}c_1+s_{m2}c_2+\ldots+s_{mm}c_m=\beta d_m\end{cases}$$

其中 $\beta=(c_1d_1+c_2d_2+\ldots+c_md_m)/\lambda$，上述方程与下面的方程的解等价：

$$\begin{cases} s_{11}c_1+s_{12}c_2+\ldots+s_{1m}c_m=d_1 \\ s_{21}c_1+s_{22}c_2+\ldots+s_{2m}c_m=d_2 \\ \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \\ s_{m1}c_1+s_{m2}c_2+\ldots+s_{mm}c_m=d_m \end{cases} \xrightarrow{\text{矩阵形式}} \boldsymbol{Sc}=\boldsymbol{d}$$

上述方程组为求解判别系数的正规方程组。其中 $\boldsymbol{S}$ 是两个类别的“类内离差乘积阵”之和（类内：每个类别的组内），当资料为距平变量时，有：

$$\boldsymbol{S}=\boldsymbol{S}(A)+\boldsymbol{S}(B)=\boldsymbol{X}_d(A)\boldsymbol{X}_d^{\mathrm{T}}(A)+\boldsymbol{X}_d(B)\boldsymbol{X}_d^{\mathrm{T}}(B)$$

(3) 判别函数显著性检验

检验两类总体的**均值是否有显著差异**，即原假设 H_0: $\mu_1-\mu_2=0$

统计量：$F=\left(\frac{n_1+n_2-m-1}{(n_1+n_2-2)m}\right)\left(\frac{n_1n_2}{n_1+n_2}\right)D^2 \sim F(m,\ n_1+n_2-m-1)$

其中 D^2 称为马氏距离，计算公式：$D^2=\boldsymbol{d}^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{d}$；$\boldsymbol{V}=\frac{1}{n_1+n_2-2}\boldsymbol{S}$

其中 $\boldsymbol{d}=\overline{\boldsymbol{X}}_A-\overline{\boldsymbol{X}}_B$，即 m 个因子中，每个因子在两类中的均值差组成的向量；

3. Fisher 多级判别

(1) 多级判别的初步认识

假设有 m 个变量，进行了 n 次观测（即获得了 n 个样品，每个样品有 m 个因子）。n 个样品共分为了 G 类，每类样品的数量分别为 $n_1,n_2,\ldots,n_G$（注意他们不一定要相等），具体如下图。

	第1类				.....	第g类				.....	第G类				均值
x_1	x_{11}^1	x_{12}^1	...	$x_{1n_1}^1$	...	x_{11}^g	x_{12}^g	...	$x_{1n_g}^g$	...	x_{11}^G	x_{12}^G	...	$x_{1n_G}^G$	$\overline{x_1}$
x_2	x_{21}^1	x_{21}^1	...	$x_{2n_1}^1$	...	x_{21}^g	x_{21}^g	...	$x_{2n_g}^g$	...	x_{21}^G	x_{21}^G	...	$x_{2n_G}^G$	$\overline{x_2}$
⋮	⋮	⋮	⋱	⋮	...	⋮	⋮	⋱	⋮	...	⋮	⋮	⋱	⋮	⋮
x_m	x_{m1}^1	x_{m1}^1	...	$x_{mn_1}^1$	...	x_{m1}^g	x_{m1}^g	...	$x_{mn_g}^g$	...	x_{m1}^G	x_{m1}^G	...	$x_{mn_G}^G$	$\overline{x_m}$
y	y_1^1	y_2^1	...	$y_{n_1}^1$	...	y_1^g	y_2^g	...	$y_{n_g}^g$	...	y_1^G	y_2^G	...	$y_{n_1}^G$	$\bar{y}$

	第g类				均值(重心)
x_1	x_{11}^g	x_{12}^g	$\cdots$	$x_{1n_1}^g$	$\overline{x_1^g}$
x_2	x_{21}^g	x_{21}^g	$\cdots$	$x_{2n_1}^g$	$\overline{x_2^g}$
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\cdots$
x_m	x_{m1}^g	x_{m1}^g	$\cdots$	$x_{mn_1}^g$	$\overline{x_m^g}$
y	y_1^g	y_2^g	$\cdots$	$y_{n_1}^g$	$\overline{y^g}$

- 第 g 类中第 $k(1\le k\le n_g)$个样品为：$\boldsymbol{x}_k^g=\left[x_{1k}^g,x_{2k}^g,...,x_{mk}^g\right]$
- 第 g 类中第 k 个样品的判别函数值为：$y_k^g=\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}_k^g$
- 第 g 类样品的均值向量为：$\overline{\boldsymbol{x}}^g=\left[x_1^g,x_2^g,...,x_m^g\right]$
- 第 g 类样品的判别函数的均值(重心)：$\overline{y}^g=\boldsymbol{c}^{\mathrm{T}}\overline{\boldsymbol{x}}^g$
- 全体样品的均值向量：$\overline{\boldsymbol{x}}=\left[\overline{x}_1,\overline{x}_2,...,\overline{x}_3\right]$
- 全体样品(n 个)判别函数的均值：$\overline{y}=\boldsymbol{c}^{\mathrm{T}}\overline{\boldsymbol{x}}$

(2) 判别函数离差平方和的分解

- **判别函数总的离差平方和——每一个样品的判别函数值与所有样品判别函数的均值之间的离差平方和**

$$S_T=\sum_{g=1}^{G}\sum_{k=1}^{n_g}\left(y_k^g-\overline{y}\right)^2=\sum_{g=1}^{G}\sum_{k=1}^{n_g}\left(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}_k^g-\boldsymbol{c}^{\mathrm{T}}\overline{\boldsymbol{x}}\right)^2=\sum_{g=1}^{G}\sum_{k=1}^{n_g}\left(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}_k^g-\boldsymbol{c}^{\mathrm{T}}\overline{\boldsymbol{x}}\right)\left(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}_k^g-\boldsymbol{c}^{\mathrm{T}}\overline{\boldsymbol{x}}\right)=\boldsymbol{c}^{\mathrm{T}}\left[\sum_{g=1}^{G}\sum_{k=1}^{n_g}\left(\boldsymbol{x}_k^g-\overline{\boldsymbol{x}}\right)\left(\boldsymbol{x}_k^g-\overline{\boldsymbol{x}}\right)^{\mathrm{T}}\right]\boldsymbol{c}$$

记$\boldsymbol{T}=\sum_{g=1}^{G}\sum_{k=1}^{n_g}\left(\boldsymbol{x}_k^g-\overline{\boldsymbol{x}}\right)\left(\boldsymbol{x}_k^g-\overline{\boldsymbol{x}}\right)^{\mathrm{T}}$，称$T$为 m 个因子总的离差交叉乘积阵，也就是每个样品与总样品均值差的平方和(每个样品是一个向量，总体样品就是均值向量)。

[注]$\sum_{g=1}^{G}\sum_{k=1}^{n_g}\left(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}_k^g-\boldsymbol{c}^{\mathrm{T}}\overline{\boldsymbol{x}}\right)^2=\sum_{g=1}^{G}\sum_{k=1}^{n_g}\boldsymbol{c}^{\mathrm{T}}\left(\boldsymbol{x}_k^g-\overline{\boldsymbol{x}}\right)\boldsymbol{c}^{\mathrm{T}}\left(\boldsymbol{x}_k^g-\overline{\boldsymbol{x}}\right)$，而$\boldsymbol{c}^{\mathrm{T}}\left(\boldsymbol{x}_k^g-\overline{\boldsymbol{x}}\right)$是一个标量，所以可以进行转置且不会影响结果。故对第二个表达式进行转置操作，即可变换得到我们最后的结果。

- **判别函数的组间离差平方和**

$$S_B=\sum_{g=1}^{G}\left(\bar{y}^g-\bar{y}\right)^2=\sum_{g=1}^{G}\left(\boldsymbol{c}^{\mathrm{T}}\bar{\boldsymbol{x}}^g-\boldsymbol{c}^{\mathrm{T}}\bar{\boldsymbol{x}}\right)^2=\boldsymbol{c}^{\mathrm{T}}\left[\sum_{g=1}^{G}\left(\bar{\boldsymbol{x}}^g-\bar{\boldsymbol{x}}\right)\left(\bar{\boldsymbol{x}}^g-\bar{\boldsymbol{x}}\right)^{\mathrm{T}}\right]\boldsymbol{c}$$

记$\boldsymbol{B}=\sum_{g=1}^{G}\left(\bar{\boldsymbol{x}}^g-\bar{\boldsymbol{x}}\right)\left(\bar{\boldsymbol{x}}^g-\bar{\boldsymbol{x}}\right)^{\mathrm{T}}$，称$B$为组间离差乘积阵。

- **判别函数（合并）组内离差平方和——所有类别每一个组内的离差平和的总和**

$$S_W=\sum_{g=1}^{G}\sum_{k=1}^{n_g}\left(y_k^g-\bar{y}^g\right)^2=\sum_{g=1}^{G}\sum_{k=1}^{n_g}\left(\boldsymbol{c}^{\mathrm{T}}\boldsymbol{x}_k^g-\boldsymbol{c}^{\mathrm{T}}\bar{\boldsymbol{x}}^g\right)^2=\boldsymbol{c}^{\mathrm{T}}\left[\sum_{g=1}^{G}\sum_{k=1}^{n_g}\left(\boldsymbol{x}_k^g-\bar{\boldsymbol{x}}^g\right)\left(\boldsymbol{x}_k^g-\bar{\boldsymbol{x}}^g\right)^{\mathrm{T}}\right]\boldsymbol{c}$$

记$\boldsymbol{W}=\sum_{g=1}^{G}\sum_{k=1}^{n_g}\left(\boldsymbol{x}_k^g-\bar{\boldsymbol{x}}^g\right)\left(\boldsymbol{x}_k^g-\bar{\boldsymbol{x}}^g\right)^{\mathrm{T}}$，称$W$为合并组内离差乘积阵

综上，我们可以得到如下关系：$\boldsymbol{c}^{\mathrm{T}}\boldsymbol{T}\boldsymbol{c}=\boldsymbol{c}^{\mathrm{T}}\boldsymbol{B}\boldsymbol{c}+\boldsymbol{c}^{\mathrm{T}}\boldsymbol{W}\boldsymbol{c}\Rightarrow\boldsymbol{T}=\boldsymbol{B}+\boldsymbol{W}$

(3) Fisher 多级判别准则

和二级判别一样，我们的目标还是要使得$\lambda=\frac{S_B}{S_W}$达到最大，根据极值原理，需要对矩阵进行求导。其结果为：$\left(\boldsymbol{W}^{-1}\boldsymbol{B}-\lambda\boldsymbol{I}\right)\boldsymbol{c}=0$，也就是说，系数向量$\boldsymbol{c}$即为矩阵$\boldsymbol{W}^{-1}\boldsymbol{B}$的特征向量，而$\lambda=\frac{S_B}{S_W}$为$\boldsymbol{W}^{-1}\boldsymbol{B}$的特征值。

设$\boldsymbol{W}^{-1}\boldsymbol{B}$共有 s 个$(s\le\min(m,G-1))$大于零的特征值：$\lambda_1\ge\lambda_2\ge\cdots\ge\lambda_s>0$。$\boldsymbol{W}^{-1}\boldsymbol{B}$的对应于“最大特征值$\lambda_1$”的特征向量$\boldsymbol{c}_1$可以使$\frac{S_B}{S_W}$最大，此外，其他特征向量$\boldsymbol{c}_i(1<i\le s)$都可以使$\frac{S_B}{S_W}$达到**极大值**。共存在 s 个判别函数，相应特征值的大小$(\lambda=\frac{S_B}{S_W})$反映了判别函数的判别能力，可称为“判别效率”。

- 具体求导推到过程涉及新的知识点，这里只给出了求导结果，相关知识点的学习可以参考以下网址：

  [1] 矩阵求导、几种重要的矩阵及常用的矩阵求导公式

  [2] 标量、向量、矩阵求导（两种布局方式）

(4) 判别规则

- **判别规则(一)——优先使用第一判别函数**

  计算y_0与每一个类别的距离，离哪个类别最近就把它划入哪个类别。此外还需要考虑方差效应，因为方差越大意味着某类别的范围越大，因此计算出距离后还要再除以方差，找到最小的那一个。如果没有唯一最小(最近的类别不唯一)，那么将剩下的使用第二判别函数进行分类，依次类推，找到距离最小的那一个类别。

- **判别规则(二)——取前 K(特征值累计权重 ≥ 0.7)个判别函数综合判别**

判别系数	第1类重心的判别函数向量：$\overline{\boldsymbol{y^1}}$	…	第G类重心的判别函数向量：$\overline{\boldsymbol{y^G}}$	新样品x_0的判别函数向量：$\overline{\boldsymbol{y_0}}$
判别系数1$(\boldsymbol{c_1})$	$\overline{y_1^1}=\boldsymbol{c}_1^{\mathrm{T}}\overline{\boldsymbol{x}^1}$	…	$\overline{y_1^G}=\boldsymbol{c}_1^{\mathrm{T}}\overline{\boldsymbol{x}^G}$	$y_{10}=\boldsymbol{c}_1^{\mathrm{T}}\boldsymbol{x}_0$
判别系数2$(\boldsymbol{c_2})$	$\overline{y_2^1}=\boldsymbol{c}_2^{\mathrm{T}}\overline{\boldsymbol{x}^1}$	…	$\overline{y_2^G}=\boldsymbol{c}_2^{\mathrm{T}}\overline{\boldsymbol{x}^G}$	$y_{20}=\boldsymbol{c}_2^{\mathrm{T}}\boldsymbol{x}_0$
……	…	…	…	…
判别系数k$(\boldsymbol{c}_k)$	$\overline{y_k^1}=\boldsymbol{c}_k^{\mathrm{T}}\overline{\boldsymbol{x}^1}$	…	$\overline{y_k^G}=\boldsymbol{c}_k^{\mathrm{T}}\overline{\boldsymbol{x}^G}$	$y_{k0}=\boldsymbol{c}_{\mathrm{k}}^{\mathrm{T}}\boldsymbol{x}_0$

每一类都有一个均值向量，即这一类的重心。那么，这些重心在不同的判别系数映射下，可以计算得到不同的判别函数值，于是每一类下，可以得到一个由 K 个判别函数值构成的向量，作为这一类在 K 维空间中的位置。对于新样品，可以计算其对应的判别函数向量，计算这两个向量之前的距离，找到距离最小的归类，即：

$$\min_{1\le g\le G}\left\{\left(\boldsymbol{y}_0-\overline{\boldsymbol{y}}^g\right)^{\mathrm{T}}\left(\boldsymbol{y}_0-\overline{\boldsymbol{y}}^g\right)\right\}\left(\min_{1\le g\le G}\left\{\sum_{i=1}^{k}\left(y_{i0}-\overline{y}_i^g\right)^2\right\}\right)$$

[注]我们把每一类 m 个因子的均值组成的向量称为这一类的均值向量或者重心，重心与判别系数相乘得到的叫做该类判别函数的重心(是这一类多个样品的判别函数的均值)。判别规则(二)的表格中，$\overline{\boldsymbol{y}}^G$ 是由**第 G 类重心对应的不同判别函数值**构成的向量，而这些判别函数值也是均值，因而 $\overline{\boldsymbol{y}}^G$ 又叫做**判别函数的重心**。

二、 贝叶斯判别

1. 先验概率和后验概率

(1) 先验概率

在采样之前，我们就对任一样品(不管其为何值)它来自各总体的概率 p_1，p_2，…,p_G 已有所了解，可根据历史样品计算某类样品数占总样品数的比例， 得到某一总体中的样品被抽到的概率。

$$p_g=\frac{n_g}{\sum_{i=1}^{G}n_i}=\frac{n_g}{n}$$

(2) 后验概率

当样品 $\boldsymbol{x}$ 的值已知时($\boldsymbol{x}=[x_1,x_2,\ldots,x_m]^{\mathrm{T}}$)， 再来判断 $\boldsymbol{x}$ 属于 A_g 总体的概率，称为“后验概率”，记为 $P(g|\boldsymbol{x})$，这是一种条件概率。

2. 贝叶斯公式(离散情况)

设一共有 $1,2,3,\cdots,G$ 类总体，每一类总体的先验概率为 $P(\boldsymbol{x}\in g)=p_g\,(1\le g\le G)$，概率密度函数为 $f_g(\boldsymbol{x})$。概率密度函数的意义为：若抽到一个来自第 g 类的样品，其值为 $\boldsymbol{x}_0$ 的概率为 $f_g(\boldsymbol{x}_0)$。若有一样品 $\boldsymbol{x}_0$，求其来自第 h 类的概率，此时，根据贝叶斯公式可以写成如下形式：

$$P(h\,|\,\boldsymbol{x}_0)=\frac{P(\boldsymbol{x}_0\in h)P(\boldsymbol{x}_0\,|\,h)}{\sum_{g=1}^{G}P(\boldsymbol{x}_0\in g)P(\boldsymbol{x}_0\,|\,g)}=\frac{p_h f_h(\boldsymbol{x}_0)}{\sum_{g=1}^{G}p_g f_g(\boldsymbol{x}_0)}$$

3. 错判损失

如果对来自 $\mathrm{A_g}$ 类的 $\boldsymbol{x}$ 判别的结果是“A_h”，错判造成的**损失**记为: $L(h|g)$，且:

$$\begin{cases}L(h\,|\,g)=0, & \text{当}h=g\text{时}\\ L(h\,|\,g)>0, & \text{当}h\ne g\text{时}\end{cases}$$

可用二者乘积 $P(g\,|\,\boldsymbol{x})L(h\,|\,g)$ 表征损失的严重性。可见样品 $\boldsymbol{x}$ 来自 $\mathrm{A_g}$ 总体的概率 $P(g\,|\,\boldsymbol{x})$ 越大，错判造成的损失也越大。

[注]

(1) 注意对错判损失函数的理解：$L(h|g)$表示真实情况属于第g类，却错判成第h类的损失；

(2) 错判损失不会影响先验概率，考虑错判损失，主要是为了决策（应该将样品判给损失小的类）。

4. 样品属于某一类的概率计算

求解思路：由贝叶斯公式可知，分母部分与类别无关，所以要使$P(h|\boldsymbol{x}_0)$达到最大，则使$p_h f_h(\boldsymbol{x}_0)$最大。

- 多元概率密度分布：$f_g(\boldsymbol{x})=\dfrac{1}{(2\pi)^{m/2}|\boldsymbol{V}|^{\frac{1}{2}}}\exp[-\dfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}_g)^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_g)]$
- 求$p_h f_h(\boldsymbol{x}_0)\Leftrightarrow$求$y_h=\ln p_h f_h(\boldsymbol{x}_0)$
- $y_h=\ln p_h+\boldsymbol{c}_h^{\mathrm{T}}\boldsymbol{x}+c_{0h}$，其中各参数为：$\begin{cases}\boldsymbol{c}_h=\boldsymbol{V}^{-1}\boldsymbol{\mu}_h=[c_{1h},c_{2h},\ldots,c_{mh}]^{\mathrm{T}}\\ c_{0h}=-\dfrac{1}{2}\boldsymbol{\mu}_h^{\mathrm{T}}\boldsymbol{V}^{-1}\boldsymbol{\mu}_h=-\dfrac{1}{2}\boldsymbol{\mu}_h^{\mathrm{T}}\boldsymbol{c}_h\end{cases}$

（p_h为先验概率，$\boldsymbol{V}$为m个因子的协方差阵）

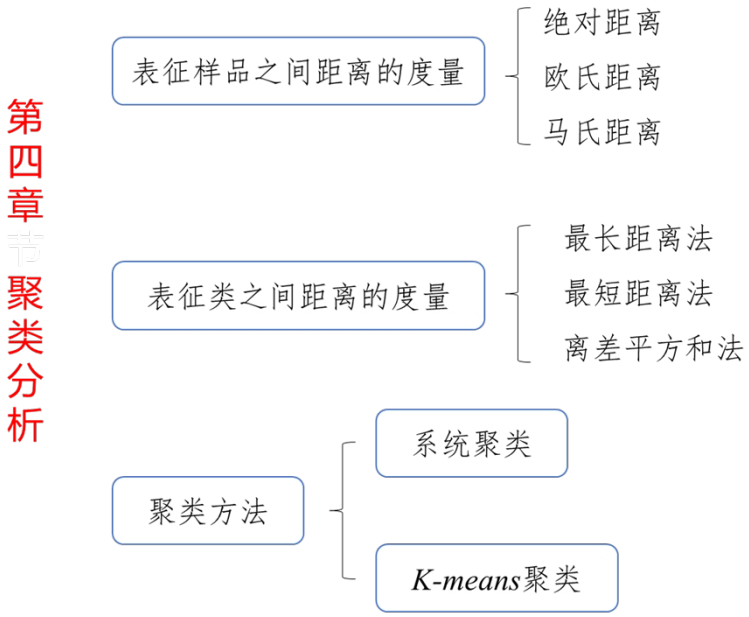
综上，可以求出该类样品属于每一类的概率：

$$P(h|\boldsymbol{x}_0)=\frac{p_h f_h(\boldsymbol{x}_0)}{\sum_{g=1}^{G}p_g f_g(\boldsymbol{x}_0)}=\frac{e^{y_h(\boldsymbol{x}_0)}}{\sum_{g=1}^{G}e^{y_g(\boldsymbol{x}_0)}}$$

找到概率最大的那一个归类（这是在错判损失相同的情况下）

第四章 聚类分析

➢ *章节思维导图*



一、相似性的度量

要对 n 个样品进行分类，首先要衡量任意两个样品之间的接近程度，有两类指标：**距离系数和相似系数**。已知 m 个因子的 n 次观测如下：

$$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n]=\begin{bmatrix} x_{11},x_{12},\ldots,x_{1n} \\ x_{21},x_{22},\ldots,x_{2n} \\ \ldots\ldots\ldots\ldots\ldots \\ x_{m1},x_{m2},\ldots,x_{mn} \end{bmatrix}$$

1. 距离系数

(1) 绝对距离

两样品各因子之差的绝对值之和，即 $\boldsymbol{X}$ 矩阵中两列向量作差、取绝对值、然后求和。

$$d_{ij}=\sum_{k=1}^{m}|x_{ki}-x_{kj}|,\quad (i,j=1,2,\ldots,n)$$

(2) 欧氏距离

$\boldsymbol{X}$ 矩阵中两列向量 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 的差向量各元素的平方和的平方根。

$$d_{ij}=\sqrt{\sum_{k=1}^{m}(x_{ki}-x_{kj})^2}=[(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}(\boldsymbol{x}_i-\boldsymbol{x}_j)]^{\frac{1}{2}},\quad (i,j=1,2,...,n)$$

(3) 马氏距离

马氏距离考虑到各指标之间的相关性，又称为**协方差距离**。

$$d_{ij}=\sqrt{(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}V^{-1}(\boldsymbol{x}_i-\boldsymbol{x}_j)}$$

$\boldsymbol{V}$为各因子的协方差阵，可用样本计算协方差阵来估计

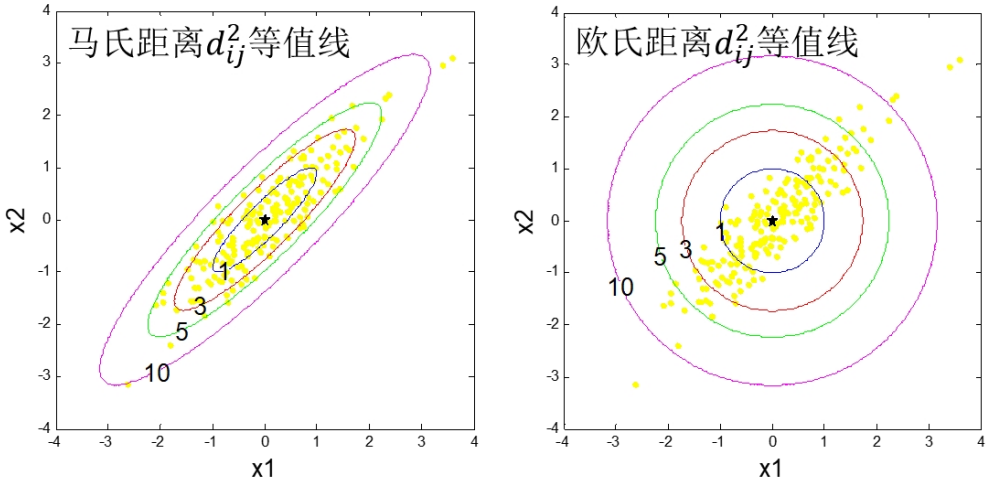
[注] 三类距离的比较

(1) 绝对距离和欧氏距离易受数据量纲的影响，因此在使用这两种距离度量时，应该对数据进行标准化处理；

(2) 马氏距离不受指标量纲的影响(即利用距平和标准化数据算得的马氏距离相同)，还考虑了各指标之间的相关性；

(3) 若各因子变量已经过标准化处理(方差为 1)且各因子相互独立，于是协方差阵 $\boldsymbol{V}$ 为单位阵，这时的马氏距离等于欧氏距离；

(4) 欧式距离与马氏距离的等值线分别如下图所示：



(5) 马氏距离与多元概率密度函数分布的关系：

$$f(\boldsymbol{x})=\frac{1}{(2\pi)^{m/2}|\boldsymbol{V}|^{1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{V}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right]==\frac{1}{(2\pi)^{m/2}|\boldsymbol{V}|^{1/2}}\exp\left(-\frac{D^2}{2}\right)$$

从这个关系可以看出，当马氏距离越小的时候，对应的概率密度函数值越大，即样品和均值的距离越近。

2. 相似系数

$\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 两个样品(列向量)是 m 维空间中的两个列向量，则 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 之间的相似程度可用两个向量之间的夹角余弦表示：

$$\cos\theta_{ij}=\frac{\boldsymbol{x}_i\cdot\boldsymbol{x}_j}{|\boldsymbol{x}_i||\boldsymbol{x}_j|}=\frac{\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j}{|\boldsymbol{x}_i||\boldsymbol{x}_j|}=\frac{\sum_{k=1}^{m}x_{ki}x_{kj}}{\sqrt{\sum_{k=1}^{m}x_{ki}^2}\sqrt{\sum_{k=1}^{m}x_{kj}^2}}$$

显然，$cos\theta_{ij}$ 的取值范围为[-1, 1]。$cos\theta_{ij}$ 越大，表明两个样品之间的相似程度越高，因此，为了符合“距离越小表示样品(类)间距离越近”的认识，常常把 1-$cos\theta_{ij}$ 作为两样品(类)之间的距离。

[注]

(1) 相关系数通常针对两个“指标”(两个变量)来计算，计算过程包含求距平。相似系数通常针对两个“样品”(两次观测)来计算，计算过程无需求距平；

(2) 相似系数多用于比较**多个变量**的**两组**空间分布的**相似程度**。相关系数多用于比较**两个变量**的**多次时间观测**(即两个时间序列)的**相关程度**。

二、 系统聚类法

1. 系统聚类法步骤

(1) 最开始，把 n 个样品各成一类，即 G_1，G_2，…, G_n 类，

(2) 然后两两计算类与类之间的距离(此时其实计算的还是两个样品之间的距离)，选择距离最小的两类合并成新的一类(新类中含有两个样品)，

(3) 然后重新计算新类与其他类之间的距离（类间距离），

(4) 找距离最小的两类合并，如此进行下去，直至所有样品都合成一大类为止。

▲ 系统聚类法的关键在于计算距离，而前面提到的距离系数(绝对距离、欧氏距离、马氏距离等)是**两**样品间的距离，而不是**类间的距离**。

▲ 常见的计算类间距离的方法有：

“最短距离”法、“最长距离”法、“离差平方和”

2. 最短(长)距离法

我们通过一个例子来理解这个方法。已知六个样品如下：

$$\begin{bmatrix}1\\0\end{bmatrix},\ \begin{bmatrix}-1\\-1\end{bmatrix},\ \begin{bmatrix}2\\4\end{bmatrix},\ \begin{bmatrix}1\\3\end{bmatrix},\ \begin{bmatrix}-2\\2\end{bmatrix},\ \begin{bmatrix}0\\3\end{bmatrix}$$

样品距离采用绝对距离，按最长距离法进行分类，并画出聚类图。

- **Step 1：计算初始各类间的距离(即样品间的距离)；**

$D_{(0)}$	G1	G2	G3	G4	G5	G6
G1	\					
G2	3	\				
G3	5	8	\			
G4	3	6	2	\		
G5	5	4	6	4	\	
G6	4	5	3	1	3	\

- **Step 2：G4 与 G6 合并为 G7，按最长距离法计算类间距离；**

$D_{(1)}$	G1	G2	G3	G5	G7(4,6)
G1	\				
G2	3	\			
G3	5	8	\		
G5	5	4	6	\	
G7(4,6)	4	6	3	4	\

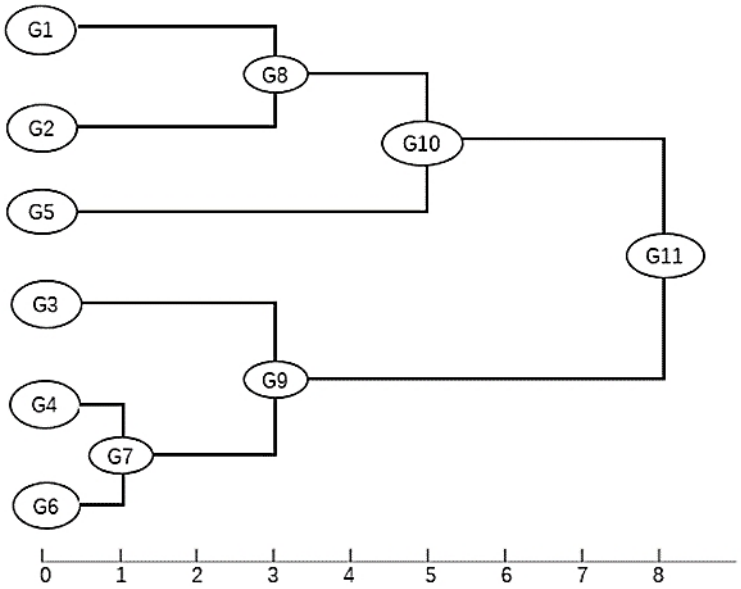
- **Step 4：G1 与 G2 合并为 G8， G3 与 G7 合并为 G9；**

$D_{(2)}$	G5	G8(1,2)	G9(3,4,6)
G5	\		
G8(1,2)	5	\	
G9(3,4,6)	6	8	\

- **Step 5：G5 和 G8 合并为 G10；**

$D_{(3)}$	G9(1,2,5)	G10(3,4,6)
G9(1,2,5)	\	
G10(3,4,6)	8	\

- **Step 6：G9 和 G10 合并为 G11，画出聚类图如下：**



(横轴表示合并时的类间距离)

[注]最长距离法、最短距离法由于计算类间距离，其中最长距离法最容易混淆。在最开始计算距离时，因为每一类都是一个样品本身，而计算样品间的距离时，不能因为是“采用最长距离法”就选择距离最大的进行合并——这显然是不符合常理的。

3. 离差平方和法

(1) 离差平方和法的思想

同一类别内部各样品间的离差平方和应较小，类与类之间的离差平方和应较大。设有 m 个指标(因子)，观测到容量为 n 的样本，如下：

$$\boldsymbol{X}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_n]=\begin{bmatrix} x_{11},x_{12},\ldots,x_{1n} \\ x_{21},x_{22},\ldots,x_{2n} \\ \ldots\ldots\ldots\ldots\ldots \\ x_{m1},x_{m2},\ldots,x_{mn} \end{bmatrix}$$

这 n 个样品可分为 k 类， G_1， G_2，…, G_k，每类的样品数为 n_g(g=1,2,…, k)，总

和为 n。对于第 g 类的资料阵（m 行 n_g 列，从 $\boldsymbol{X}$ 中抽取 n_g 列）可表示为：

$$\boldsymbol{X}_g=\begin{bmatrix} x_{11}^g & x_{12}^g & \cdots & x_{1n_g}^g \\ x_{21}^g & x_{22}^g & \cdots & x_{2n_g}^g \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1}^g & x_{m2}^g & \cdots & x_{mn_g}^g \end{bmatrix}$$

则第 g 类的类内离差平方和为：

$$S_g=\sum_{t=1}^{n_g}(\boldsymbol{x}_t^g-\overline{\boldsymbol{x}}^g)^{\mathrm{T}}(\boldsymbol{x}_t^g-\overline{\boldsymbol{x}}^g)$$

其中 $\overline{\boldsymbol{x}}^g$ 为第 g 类的重心(均值向量)，$\boldsymbol{x}_t^g$ 为第 g 类的样品。从而 k 个类别总的类内离差平方和为：

$$S=\sum_{g=1}^{k}\sum_{t=1}^{n_g}(\boldsymbol{x}_t^g-\overline{\boldsymbol{x}}^g)^{\mathrm{T}}(\boldsymbol{x}_t^g-\overline{\boldsymbol{x}}^g)$$

我们的目的是要固定一个 k 时，使 S 达到极小。

(2) 实际计算方法

把 n 个样品分成 k 类的方案很多，要比较所有的分法使得 S 最小，不现实。因此，只好放弃在一切分类中寻求 S 的极小值，提出使 S 达到局部极小的办法。

设 G_p 与 G_q 两类的类内离差平方和分别为 S_p 和 S_q，若 G_p 与 G_q 合并成 G_r 类后的离差平方和为 S_r，则此次合并导致总离差平方和的增量为

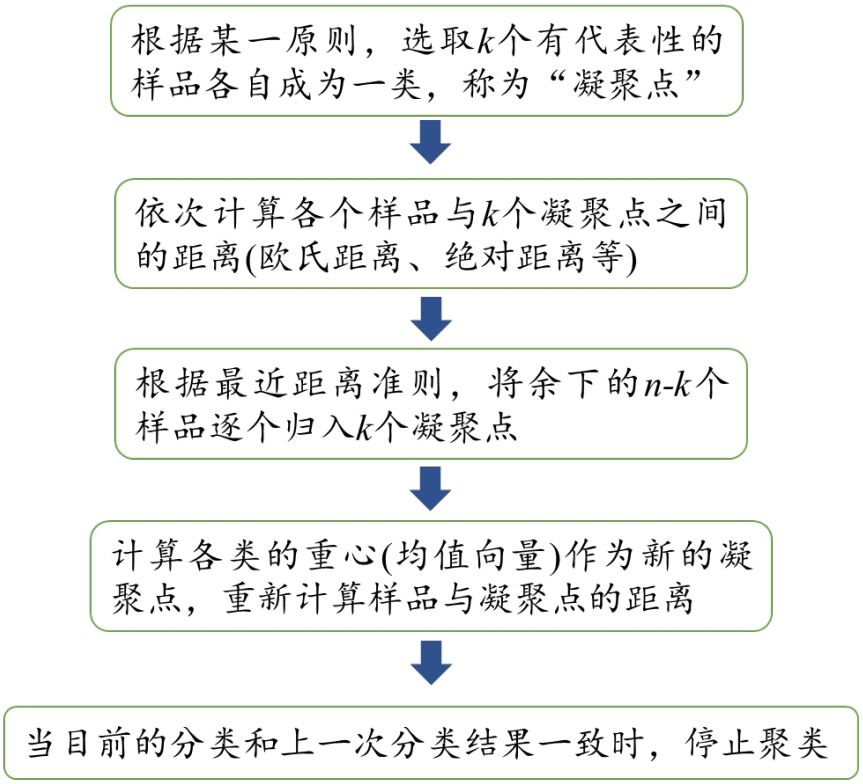
$$D_{pq}^2=S_r-(S_p+S_q)$$

聚类的原则是：**选择使 D^2_{pq} 最小的两类合并**，因此 D^2_{pq} 可认为是两类之间的距离，可以证明 D^2_{pq} 可由 G_p 与 G_q 两类的重心之差的平方和来表示：

$$D_{pq}^2=\frac{n_pn_q}{n_p+n_q}(\overline{\boldsymbol{x}}^p-\overline{\boldsymbol{x}}^q)^{\mathrm{T}}(\overline{\boldsymbol{x}}^p-\overline{\boldsymbol{x}}^q)$$

三、(定)*K-means* 聚类法

聚类步骤：



通过一个例子来进一步了解。

现有 4 个样样品，每个样品采用 2 个指标。现要使用定 *K-means* 聚类法把 4 个样品分成 2 类，假设初始的划分结果是“A、B 作为一类，C、D 作为另一类”，取绝对距离作为样品距离。

	x_1	x_2
A	5	3
B	-1	1
C	1	-2
D	-3	-2

- **Step 1**：目前已经分成了两类：A 和 B，C 和 D，计算二者的均值（凝聚点）分别为 P1(2，2) 和 P2 (-1，-2)；
- **Step 2**：计算四个样品同上述凝聚点 P1 和 P2 的距离，如下图所示：

	P1	**P2**
A	4	11
B	4	3
C	5	2
D	9	2

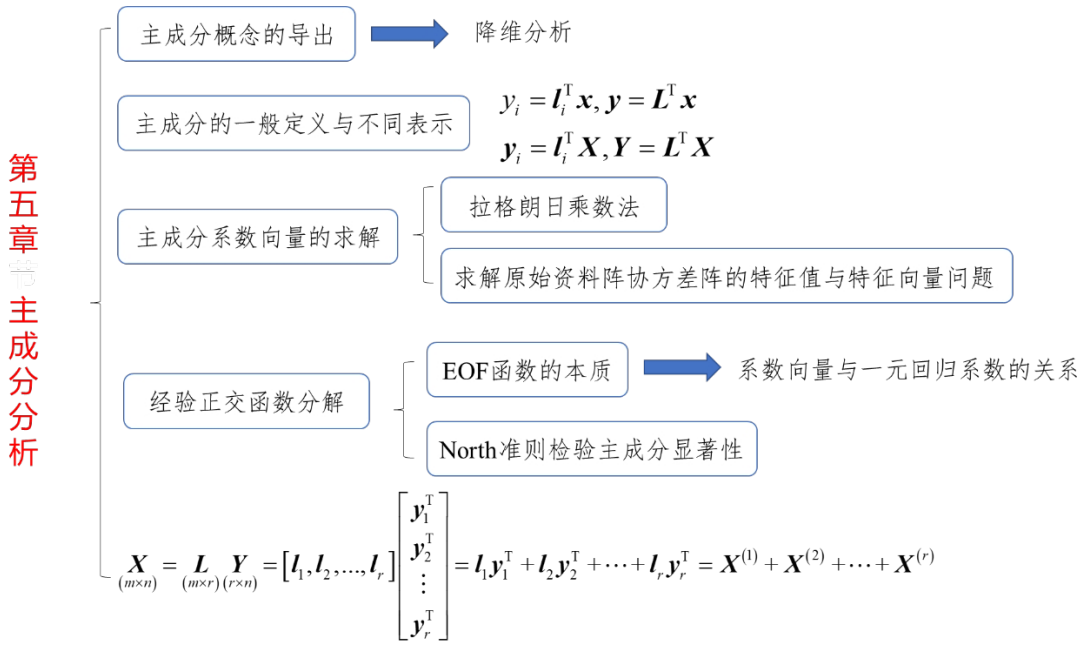
- **Step 3**: 通过第二步可以知道，应该将 A 划为一类，将 B，C，D 划为一类；
- **Step 4**: 重新计算凝聚点为：P3(5，3) 和 P4 (-1，-1)，计算样品同 P3 和 P4 的距离，如下图所示：

	P3	**P4**
A	0	10
B	8	2
C	9	3
D	13	3

- **Step 5**: 发现新的分类情况与前一次相同，停止聚类。最终分类结果为：A 为一类，B，C 和 D 为一类

第五章 主成分分析

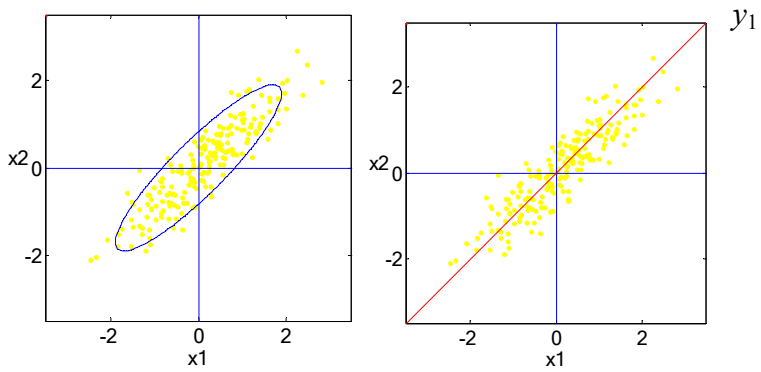
➢ *章节思维导图*



一、主成分概念的导出

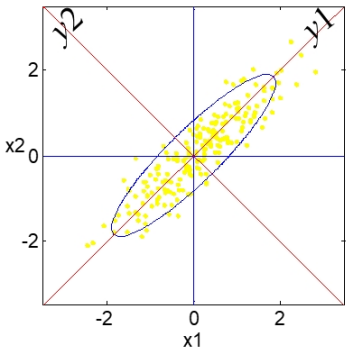
1. 主成分的概念

假设有两个因子 x_1, x_2，经过 n 次观测共获得 n 个样品，绘制出的点聚图如下左图所示(由点聚图可知存在两个因子存在相关性，且由二维正态分布的概率密度函数可知，该样品概率密度的等值线为一组组等值线)，显然，n 个样品的信息由 x_1, x_2 全部表达。



现在从另一个角度来考虑这个问题。取椭圆等值线的长轴作为新的坐标轴 y_1(如上右图所示)，可以发现，各点在红线上的坐标值可以在很大程度上表达 x_1, x_2 的变化信息，也就是说：x_1, x_2 两个变量的信息可以近似地用 1 个新变量 y_1

来表达，于是降低了变量的维度，简化了问题。需要指出的是，**y_1 可以反映 x_1,x_2 随时间变化的大部分信息**，若要表达全部信息，需再考虑以椭圆的短轴作为坐标轴的坐标信息，记作 y_2，如下图所示：



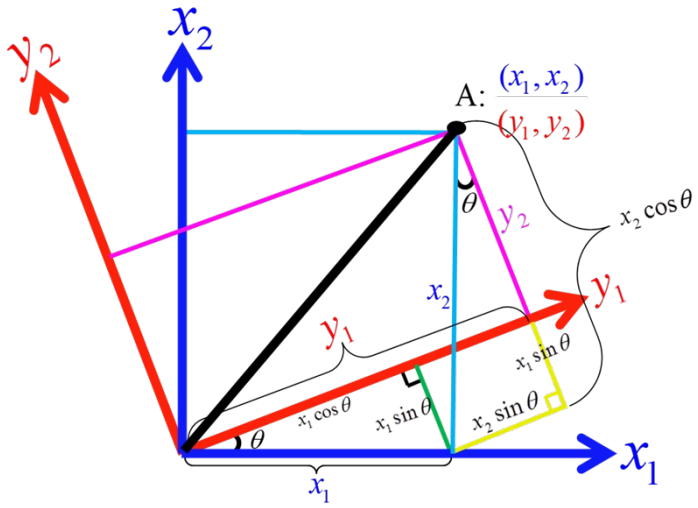
y_1,y_2 具有如下特征：

(1) y_2 的离散程度（方差）远小于 y_1；(2) y_2 与 y_1 相互独立(相关系数为 0)

称 y_1 为原变量 x_1,x_2 的“第一主成分”， y_2 为原变量 x_1,x_2 的“第二主成分”

2. 主成分表达式的导出

我们从坐标转换的角度来求出主成分的表达式。设坐标逆时针旋转了 θ 角度。



根据上图，我们可以得到新坐标和就坐标之间的关系如下：

$$\begin{cases} y_1 = x_1\cos\theta + x_2\sin\theta \\ y_2 = -x_1\sin\theta + x_2\cos\theta \end{cases} \Rightarrow \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$$

记 $\boldsymbol{l}_1=\begin{bmatrix}l_{11}\\l_{21}\end{bmatrix}=\begin{bmatrix}\cos\theta\\\sin\theta\end{bmatrix}$，$\boldsymbol{l}_2=\begin{bmatrix}l_{12}\\l_{22}\end{bmatrix}=\begin{bmatrix}-\sin\theta\\\cos\theta\end{bmatrix}$，$\boldsymbol{x}=\begin{bmatrix}x_1\\x_2\end{bmatrix}$，其中 $\boldsymbol{l}_1,\boldsymbol{l}_2$ 称为主成分的系数向量，因此两个主成分又可以表示为：$y_1=\boldsymbol{l}_1^{\mathrm{T}}\boldsymbol{x}, y_2=\boldsymbol{l}_2^{\mathrm{T}}\boldsymbol{x}$。主成分的系数向量具有如下性质：

$$\boldsymbol{l}_i^{\mathrm{T}}\boldsymbol{l}_j=\begin{cases}1, & i=j\\0, & i\neq j\end{cases}$$

即每一个主成分对应一个系数向量，且这个系数向量为单位化的向量，且不同主成分的系数向量之间是正交的。

▲ 坐标变化后的总离差平方变化情况

$$\sum_{t=1}^{n}(y_{1t}-\bar{y}_1)^2+\sum_{t=1}^{n}(y_{2t}-\bar{y}_2)^2=\sum_{t=1}^{n}y_{1t}^2-n\bar{y}_1^2+\sum_{t=1}^{n}y_{2t}^2-n\bar{y}_2^2=\sum_{t=1}^{n}(y_{1t}^2+y_{2t}^2)-n(\bar{y}_1^2+\bar{y}_2^2)$$

$$\sum_{t=1}^{n}(x_{1t}-\bar{x}_1)^2+\sum_{t=1}^{n}(x_{2t}-\bar{x}_2)^2=\sum_{t=1}^{n}x_{1t}^2-n\bar{x}_1^2+\sum_{t=1}^{n}x_{2t}^2-n\bar{x}_2^2=\sum_{t=1}^{n}(x_{1t}^2+x_{2t}^2)-n(\bar{x}_1^2+\bar{x}_2^2)$$

对于第一项，有：$\sum_{t=1}^{n}(y_{1t}^2+y_{2t}^2)=\sum_{t=1}^{n}(x_{1t}^2+x_{2t}^2)$

对于第二项：把 $\bar{y}_1=\bar{x}_1\cos\theta+\bar{x}_2\sin\theta$ 和 $\bar{y}_2=-\bar{x}_1\sin\theta+\bar{x}_2\cos\theta$ 代入，

易得：$n(\bar{y}_1^2+\bar{y}_2^2)=n(\bar{x}_1^2+\bar{x}_2^2)$

结论：**经过坐标转换后，原来样品的总的离差平方和保持不变，只是将原变量的离差平方和进行重新分配。**

3. 主成分的一般定义及不同表示

设 $\boldsymbol{x}=[x_1, x_2, \ldots, x_m]^{\mathrm{T}}$ 是一个由 m 个随机变量组成的向量，设 $\boldsymbol{x}$ 的数学期望为 $\boldsymbol{0}$，$\boldsymbol{x}$ 的第 i 个主成分的定义为：

$$y_i=\boldsymbol{l}_i^{\mathrm{T}}\boldsymbol{x}\quad(\boldsymbol{l}_i^{\mathrm{T}}\boldsymbol{l}_i=1, i=1,2,\ldots,m)$$

且满足以下条件：

(1) 在一切 $y_i=\boldsymbol{l}_i^{\mathrm{T}}\boldsymbol{x}$ 中，方差贡献最大者称为第一主成分 $y_1=\boldsymbol{l}_1^{\mathrm{T}}\boldsymbol{x}$；

(2) 第二主成分是指在一切 $y_i=\boldsymbol{l}_i^{\mathrm{T}}\boldsymbol{x}$ 中，与第一主成分 y_1 无关，并且方差最大者；

(3) 第 k 主成分是指在一切 $y_i=\boldsymbol{l}_i^{\mathrm{T}}\boldsymbol{x}$ 中，与前 k-1 个成分 $y_1,y_2,\ldots,y_{k-1}$ 无关，并且

方差最大者。

▲ 原变量有 m 个因子，则主成分最多有 m 个；

▲ 主成分与符号无关。如果 $\boldsymbol{l}_1$ 是第一主成分的系数向量， $y_1=\boldsymbol{l}_1^{\mathrm{T}}\boldsymbol{x}$ 是第一主成分，那么，$-\boldsymbol{l}_1$ 也是第一主成分的系数向量，$-y_1=-\boldsymbol{l}_1^{\mathrm{T}}\boldsymbol{x}$ 也是 $\boldsymbol{x}$ 的第一主成分；

▲ 主成分的不同表示(与维度有关)

① 若 m 个因子进行了一次观测(得到了一组样品) $\boldsymbol{x}=[x_1, x_2, \ldots, x_m]^{\mathrm{T}}$，那么第 k 主成分可以表示为：$y_k=l_{1k}x_1+l_{2k}x_2+\cdots+l_{mk}x_m=\boldsymbol{l}_k^{\mathrm{T}}\boldsymbol{x}\left(k\le m\right)$

② 若 m 个因子一共有 r 个主成分$\left(r\le m\right)$，他们的表达式如下：

$$\begin{cases} y_1=l_{11}x_1+l_{21}x_2+\cdots+l_{m1}x_m=\boldsymbol{l}_1^{\mathrm{T}}\boldsymbol{x} \\ y_2=l_{12}x_1+l_{22}x_2+\cdots+l_{m2}x_m=\boldsymbol{l}_2^{\mathrm{T}}\boldsymbol{x} \\ \cdots\cdots \\ y_r=l_{1r}x_1+l_{2r}x_2+\cdots+l_{mr}x_m=\boldsymbol{l}_r^{\mathrm{T}}\boldsymbol{x} \end{cases}$$

写成矩阵，即：

$$\boldsymbol{y}=\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_r \end{bmatrix},\boldsymbol{L}=\begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1r} \\ l_{21} & l_{22} & \cdots & l_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ l_{m1} & l_{m2} & \cdots & l_{mr} \end{bmatrix}\Rightarrow \boldsymbol{y}=\begin{bmatrix} \boldsymbol{l}_1^{\mathrm{T}}\boldsymbol{x} \\ \boldsymbol{l}_2^{\mathrm{T}}\boldsymbol{x} \\ \vdots \\ \boldsymbol{l}_r^{\mathrm{T}}\boldsymbol{x} \end{bmatrix}=\begin{bmatrix} \boldsymbol{l}_1^{\mathrm{T}} \\ \boldsymbol{l}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{l}_r^{\mathrm{T}} \end{bmatrix}\boldsymbol{x}=\boldsymbol{L}^{\mathrm{T}}\boldsymbol{x}$$

③ 若 m 个因子一共进行了 n 次观测，得到了资料阵 $\boldsymbol{X}$，且共有 r 个主成分$\left(r\le\min\{m,n\}\right)$。资料阵表示为：

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \ldots & x_{1n} \\ x_{21} & x_{22} & \ldots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \ldots & x_{mn} \end{bmatrix}$$

此时，原来②中的主成分向量就变成矩阵：

$$\boldsymbol{Y}=\boldsymbol{L}^{\mathrm{T}}\boldsymbol{X}=\begin{bmatrix} \boldsymbol{l}_1^{\mathrm{T}}\boldsymbol{X} \\ \boldsymbol{l}_2^{\mathrm{T}}\boldsymbol{X} \\ \vdots \\ \boldsymbol{l}_r^{\mathrm{T}}\boldsymbol{X} \end{bmatrix}=\begin{bmatrix} \boldsymbol{y}_1^{\mathrm{T}} \\ \boldsymbol{y}_2^{\mathrm{T}} \\ \vdots \\ \boldsymbol{y}_r^{\mathrm{T}} \end{bmatrix}=\begin{bmatrix} y_{11} & y_{12} & \cdots & y_{1n} \\ y_{21} & y_{22} & \cdots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{r1} & y_{r2} & \cdots & y_{rn} \end{bmatrix}$$

二、主成分系数向量的求解

[注] 根据主成分的定义，**原始资料阵应满足均值为 0**，如果不满足这一条件，进行主成分系数向量求解会导致错误结果。

1. 第一主成分系数向量的导出

对 m 个变量进行 n 次观测，得到 m 行 n 列的**距平**资料阵 $\boldsymbol{X}$ 如下：

$$\boldsymbol{X}=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix}$$

这时第一主成分 $\boldsymbol{l}_1^{\mathrm{T}}\boldsymbol{X}$ 为一行向量，记为$\boldsymbol{y}_1^{\mathrm{T}}$，而第一主成分的方差可以表示为：

$$s_{y_1}^2=\frac{\left(y_1-\overline{y}_1\right)^{\mathrm{T}}\left(y_1-\overline{y}_1\right)}{n-1}=\frac{y_1^{\mathrm{T}}y_1}{n-1}=\frac{\left(\boldsymbol{l}_1^{\mathrm{T}}\boldsymbol{X}\right)\left(\boldsymbol{l}_1^{\mathrm{T}}\boldsymbol{X}\right)^{\mathrm{T}}}{n-1}=\frac{\boldsymbol{l}_1^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{l}_1}{n-1}=\boldsymbol{l}_1^{\mathrm{T}}\boldsymbol{S}\boldsymbol{l}_1$$

其中 $\boldsymbol{S}=\frac{1}{n-1}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ 为 m 个变量的协方差阵。根据主成分定义，**第 1 主成分y_1^{T}应该在 $\boldsymbol{l}_1^{\mathrm{T}}\boldsymbol{l}_1=1$ 的约束条件下方差最大**——显然，这是一个条件极值问题，可以采用拉格朗日乘数法求解(由于过程涉及矩阵及向量的求导，所以只给最后化简的结果，具体过程可参考 Fisher 判别一节提供的学习链接)。构造拉格朗日函数：

$$Q\left(\boldsymbol{l}_1,\lambda\right)=\boldsymbol{l}_1^{\mathrm{T}}\boldsymbol{S}\boldsymbol{l}_1-\lambda\left(\boldsymbol{l}_1^{\mathrm{T}}\boldsymbol{l}_1-1\right)$$

根据 $\frac{\partial Q}{\partial \boldsymbol{l}_1}=0$，求得 $\boldsymbol{S}\boldsymbol{l}_1=\lambda\boldsymbol{l}_1$，即 λ 就是 $\boldsymbol{S}$ 的**特征值**，$\boldsymbol{l}_1$ 是对应于 λ 的**特征向量**。

于是，第 1 主成分 y_1 的方差又可以表示为：

$$s_{y_1}^2==\boldsymbol{l}_1^{\mathrm{T}}\boldsymbol{S}\boldsymbol{l}_1=\lambda\boldsymbol{l}_1^{\mathrm{T}}\boldsymbol{l}_1=\lambda$$

考虑约束条件，则结论如下：

主成分y_1的方差就是协方差阵 $\boldsymbol{S}$ 的最大特征值 λ，系数向量 $\boldsymbol{l}_1$ 就是对应于最大特征值 λ_1 的特征向量

2. 主成分的一些结论

(1) m 个变量的主成分的系数向量就是其协方差阵 $\boldsymbol{S}=\frac{1}{n-1}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$ 的特征值 $\boldsymbol{\lambda_1\geq\lambda_2\geq\ldots\geq\lambda_r}$ 所对应的单位化的特征向量 $\boldsymbol{l}_1$，$\boldsymbol{l}_2$, …，$\boldsymbol{l}_r$。第 k 个特征值 λ_k，就是第 k 主成分 y_k 的方差；

(2) **不同主成分是相互正交的**(协方差为 0)，各个主成分的方差就是 $\boldsymbol{S}$ 的特征值，所以，r **个主成分** $\boldsymbol{y}=[y_1, y_2, \ldots, y_r]^{\mathrm{T}}$ **的协方差阵为对角阵**，即：

$$\boldsymbol{S}_{yy}=\begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1r} \\ s_{21} & s_{22} & \cdots & s_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ s_{r1} & s_{r2} & \cdots & s_{rr} \end{bmatrix}=\begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_r \end{bmatrix}$$

(3) m 个主成分的**方差总和**与原 m 个变量的**方差总和**相等，$\sum_{i=1}^{r}\lambda_i=\sum_{i=1}^{r}s_{ii}$；

(4) 两个常用的指标：

- 方差贡献率：$VF_k=\dfrac{\lambda_k}{\sum_{i=1}^{r}\lambda_i}$；
- 累计方差贡献率：$CVF_p\left(p<r\right)=\dfrac{\sum_{i=1}^{p}\lambda_i}{\sum_{i=1}^{r}\lambda_i}$

▲ 距平资料 $\boldsymbol{X}_{\mathrm{d}}$ 与标准化资料 $\boldsymbol{X}^{*}$的主成分——**二者并不相同**

- 对于标准化资料阵：各变量方差相等，因此第一主成分所表达的信息与相互关联的变量个数有关，将反应大多数变量的信息。
- 对于距平资料阵：各变量方差不一定相等，因此，第一主成分所表达的信息会综合考虑变量方差的大小与变量个数。

三、经验正交函数分解(*Empirical Orthogonal Function decomposition, EOF* 分解)

1. 问题的引入

对于某一要素例如气温，如何考察某一**空间区域**的气温距平的时间变化规律？如果该区域各空间点的气温距平总是大致呈同位相变化（往往是面积较小的区域），可对其进行空间平均，然后分析空间均值的时间变化序列。例如：Nino3.4 指数，但是，如果该区域各点的**气温距平并非总是呈同位相变化**，则空间平均容易使不同区域的反位相变化相抵消，因而不可取，这时候，主成分分析就可以发挥作用了。主成分分析在时空数据分析中又叫经验正交函数分解。

对于地球格点(或站点)数据，**每个空间点可看作一个变量**，如果空间点数目很多，且相互之间通常存在相关性(尤其地理位置临近的变量)，因而可进行主成分分析，提取主成分的**时间**变化序列，同时对应的系数向量可表征相应的**空间**模

态。因此，这相当于把原多变量资料阵进行了时空分解

2. *EOF* 表达式的导出

已知 m 行 n 列的资料阵为 $\boldsymbol{X}$，主成分的表达式为：$\boldsymbol{Y}=\boldsymbol{L}^{\mathrm{T}}\boldsymbol{X}$，易知 $\boldsymbol{L}$ 是正交矩阵即 $\boldsymbol{L}^{\mathrm{T}}\boldsymbol{L}=\boldsymbol{I}$，所以上是可以写为：$\boldsymbol{X}=\boldsymbol{L}\boldsymbol{Y}$，具体如下：

$$\underset{(m\times n)}{\boldsymbol{X}}=\underset{(m\times r)}{\boldsymbol{L}}\underset{(r\times n)}{\boldsymbol{Y}}=[\boldsymbol{l}_1,\boldsymbol{l}_2,\ldots,\boldsymbol{l}_r]\begin{bmatrix}\boldsymbol{y}_1^{\mathrm{T}}\\ \boldsymbol{y}_2^{\mathrm{T}}\\ \vdots\\ \boldsymbol{y}_r^{\mathrm{T}}\end{bmatrix}=\boldsymbol{l}_1\boldsymbol{y}_1^{\mathrm{T}}+\boldsymbol{l}_2\boldsymbol{y}_2^{\mathrm{T}}+\cdots+\boldsymbol{l}_r\boldsymbol{y}_r^{\mathrm{T}}=\boldsymbol{X}^{(1)}+\boldsymbol{X}^{(2)}+\cdots+\boldsymbol{X}^{(r)}$$

称 $\underset{(m\times 1)}{\boldsymbol{l}_k}$ 为第 k 模态的空间函数(空间向量)，$\underset{(1\times n)}{\boldsymbol{y}_k}$ 为第 k 模态的时间函数(时间系数)，资料阵 $\boldsymbol{X}$ 就是 r 个与之形状相同(m 行 n 列)的矩阵相加所构成的。

每个 $\boldsymbol{l}_i$ 都是单位化的，但时间序列 $\boldsymbol{y}_i$ 却是按方差从大到小排列的，即越前面的 $\boldsymbol{X}^{(i)}$ 就越重要（总方差越大），如果取前 p 个 EOF 模态(主成分)，$(p<=m)$，这时就得到对原资料阵 $\boldsymbol{X}$ 的一种**估计(重构)**：

$$\widehat{\boldsymbol{X}}=\sum_{i=1}^{p}\boldsymbol{X}^{(i)}=\sum_{i=1}^{p}\boldsymbol{l}_i\boldsymbol{y}_i$$

▲ *PCA* 与 *EOF* 的比较

	主成分分析（PCA）	经验正交函数（EOF）分解
$\boldsymbol{l}_k$ (m行1列)	第k主成分的系数向量	第k模态的空间函数(空间向量)
$\boldsymbol{y}_k$ (1行n列)	第k主成分的时间序列	第k模态的时间函数(时间系数)

3. 空间函数的本质

对 $\boldsymbol{X}=\boldsymbol{L}\boldsymbol{Y}$ 两边同时乘以 $\frac{1}{n-1}\boldsymbol{Y}^{\mathrm{T}}$，得：$\frac{1}{n-1}\boldsymbol{X}\boldsymbol{Y}^{\mathrm{T}}=\frac{1}{n-1}\boldsymbol{L}\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}$

$$\text{等式左边}=\frac{1}{n-1}\boldsymbol{X}\begin{bmatrix}\boldsymbol{y}_1^{\mathrm{T}}\\ \boldsymbol{y}_2^{\mathrm{T}}\\ \vdots\\ \boldsymbol{y}_r^{\mathrm{T}}\end{bmatrix}^{\mathrm{T}}=\frac{1}{n-1}[\boldsymbol{X}\boldsymbol{y}_1,\boldsymbol{X}\boldsymbol{y}_2,\ldots,\boldsymbol{X}\boldsymbol{y}_r];$$

等式右边$=\frac{1}{n-1}\boldsymbol{LYY}^{\mathrm{T}}=[\boldsymbol{l}_1,\boldsymbol{l}_2,...,\boldsymbol{l}_r]\begin{bmatrix}\lambda_1 & 0 & \cdots & 0\\ 0 & \lambda_2 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \lambda_r\end{bmatrix}=[\lambda_1\boldsymbol{l}_1,\lambda_2\boldsymbol{l}_2,...,\lambda_r\boldsymbol{l}_r]$

于是，对应位置有$\frac{1}{n-1}\boldsymbol{Xy}_i=\lambda_i\boldsymbol{l}_i$，所以，第 j 模态的空间向量为：$\boldsymbol{l}_i=\frac{\boldsymbol{Xy}_i}{\lambda_i(n-1)}$，

若对于$\boldsymbol{l}_j$的第 i 个元素，有$l_{ij}=\frac{\frac{1}{(n-1)}\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{y}_i}{\lambda_i}=\frac{s_{x_iy_i}}{s_{y_i}^2}$——回归系数的表达式，因此，

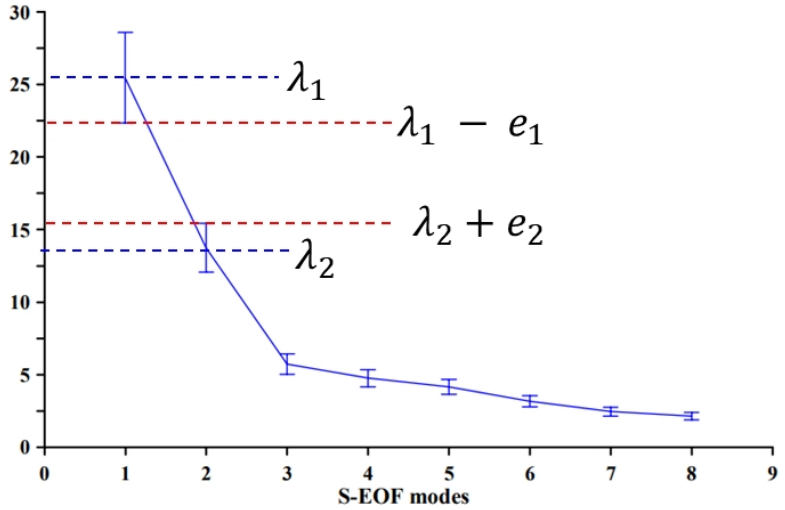
$l_{ij}=\frac{s_{x_iy_i}}{s_{y_i}^2}$就是以**“第 i 个变量 x_i ”为预报量，以“第 j 主成分 y_j ”为因子**的一元回归方程的回归系数，回归方程为：$\hat{x}_i=l_{ij}y_j$。

4. 主成分的显著性检验——North 准则

第 j 模态特征值 λ_j 的误差范围: $e_j=\lambda_j\sqrt{\frac{2}{n}}$

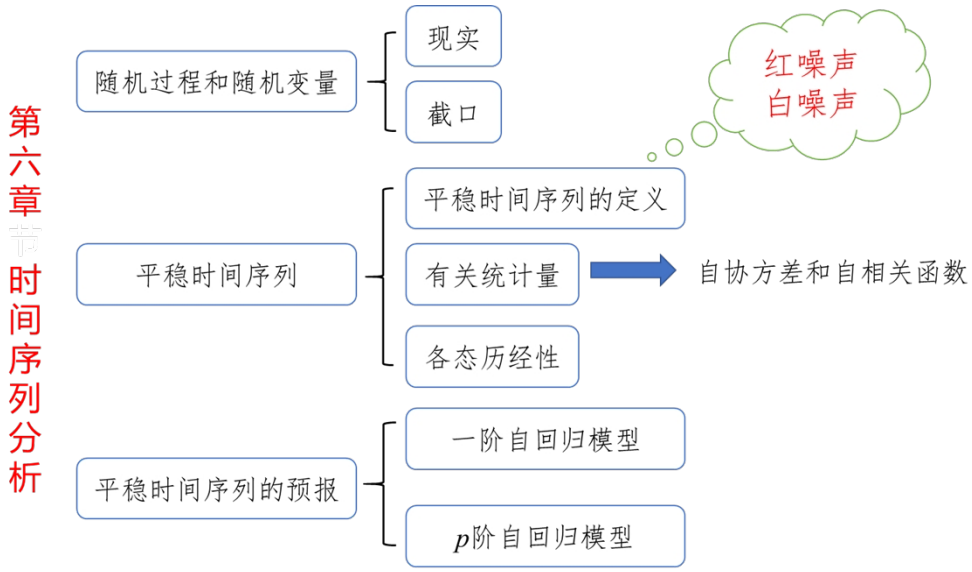
如果第 j+1 模态要显著区别于第 j 模态，须有：$\lambda_j-e_j\gg\lambda_{j+1}+e_{j+1}$

通常绘出各模态特征值(或方差贡献)及其误差棒(如下图)对各模态进行检验——**误差棒不相交原则**。下图中第四模态与第三模态的误差棒存在交集，所以可以判定，第四模态无法显著区别于第三模态，故取前三模态即可。



第六章 时间序列分析

➢ *章节思维导图*



一、 随机过程的相关概念

1. 随机过程与随机变量

随机变量(*Random Variable*)是“静态的”，是对某些随机现象的数值表达，例如某地一年一年月平均气温的观测序列。而随机过程(*Stochastic Process*)“动态的”，它依赖于参数(参数通常为时间)的一组随机变量的全体，例如气温的变化，虽气温存在春夏秋冬的周期性规律，但是每年的过程都不完全相同，具有随机性。

2. 现实与截口

月 / 年	一	二	三	四	五	六	七	八	九	十	十一	十二
1951	-2.2	-0.5	4.3	9.7	16.0	20.7	23.6	25.4	21.9	17.7	8.4	4.5
1952	0.4	-1.4	3.7	10.4	16.2	20.5	24.2	24.1	21.4	15.6	8.8	-0.4
1953	-2.3	0.0	4.7	10.7	16.1	20.6	24.7	25.8	22.4	18.5	8.4	2.5
1954	0.3	0.5	3.6	9.7	14.6	19.0	22.0	24.6	21.3	15.1	11.1	-1
…	…	…	…	…	…	…	…	…	…	…	…	…

▲ 青岛市逐年各月平均气温

在青岛市逐年逐月平均气温的变化是一个随机过程，而某一年气温观测序列仅是该随机过程的一部分，我们**把随机过程 *X(t)*的某一次观测过程称为一个“现实”，一般用 *x(t)*表示**。对于同一个月份，不同的年份其月平均气温有所差异，

此时，某月的气温在年际变化上表现为一个随机变量，我们**把随机过程在某一时刻 t_j 表现为一个随机变量的时刻 t_j 称为“截口”，记作 $X(t_j)$**。

3. 随机过程的统计特征(了解)

(1) 均值函数

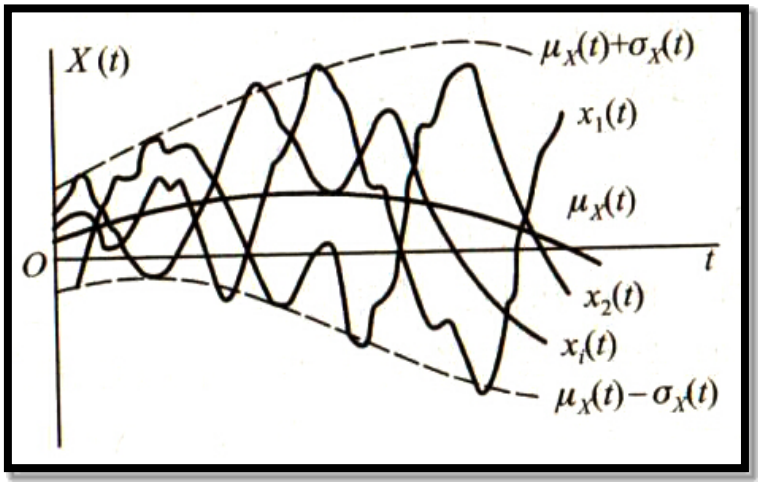
$$E(X(t))=\int_{-\infty}^{\infty}xf(x,t)\,dx=\mu(t)$$

对每一个截口(每个截口处随机过程是一个随机变量)求均值，这个均值又构成时间序列，即随机过程的均值是时间的函数

(2) 方差函数

$$D(X(t))=E[X(t)-\mu(t)]^2=\sigma^2=\int_{-\infty}^{\infty}[x-\mu(t)]^2f(x,t)dx$$

方差函数也是时间 t 的确定性的函数， 反映了每个截口处取值的变动情况，即相对于均值函数的离散程度。



(3) 协方差函数和相关函数

随机过程在任两个时刻(截口)t_1 和 t_2 表现为两个随机变量，它们的协方差为：

$$K(t_1,t_2)=E[(X(t_1)-\mu(t_1))(X(t_2)-\mu(t_2))]$$

$K(t_1,t_2)$称为**自协方差函数**，简称为协方差函数。当 $t_1=t_2=t$ 时，自协方差函数就是随机过程在第 t 时刻的方差。

t_1 和 t_2 时刻的**相关系数**为：$\rho(t_1,t_2)=\dfrac{K(t_1,t_2)}{\sigma(t_1)\sigma(t_2)}$，$\rho(t_1,t_2)$称为自相关函数，简称为相关函数，表示随机过程 $X(t)$在不同时刻 t_1 和 t_2 之间线性相关的程度。

二、平稳随机过程及其统计特征

1. 宽平稳随机过程的定义

当随机过程的统计特性不随时间的推移而变化，满足：

(1) 均值函数是与 t 无关的常数，$E[X(t)]=\mu$

(2) 协方差函数仅仅与时间间隔 τ 有关，而与 t 的起始点位置无关

$$K(t,t+\tau)=E[(X(t)-\mu)(X(t+\tau)-\mu)]=K(\tau)$$

这种随机过程称为：“**广义平稳随机过程**”或“**宽平稳随机过程**”，相应的时间序列资料称为：“宽平稳时间序列”。

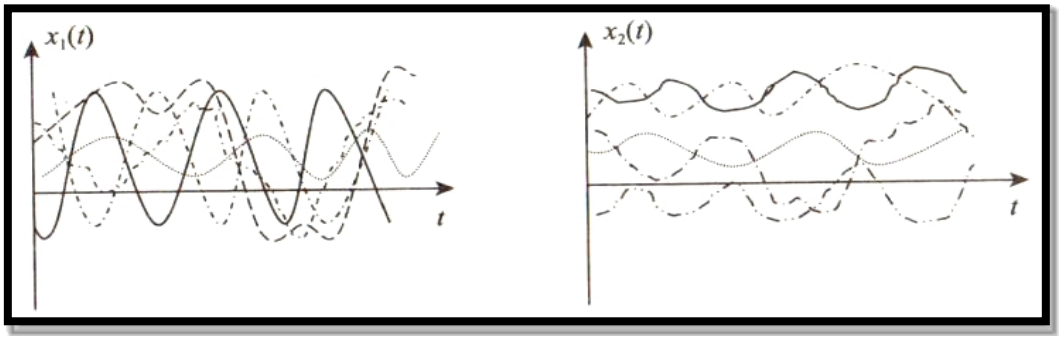
2. 宽平稳随机过程统计量的估计

$$\hat{\mu}_t=\frac{1}{n}\sum_{i=1}^{n}x_{it};\hat{\sigma}_t^2=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{it}-\hat{\mu}_t\right)^2$$

$$\hat{K}\left(t_i,t_j\right)=\frac{1}{n-1}\sum_{i=1}^{n}\left(x_{it_i}-\hat{\mu}_{t_i}\right)\left(x_{it_j}-\hat{\mu}_{t_j}\right);\hat{\rho}\left(t_i,t_j\right)=\frac{\hat{K}\left(t_i,t_j\right)}{\sigma_{t_i}\sigma_{t_j}}$$

实际上就是对某一截口进行估计，上式中的 n 为现实的个数。

3. 平稳随机过程的各态历经性



- **左图**：每个现实都围绕着随机过程的均值波动，且他们的平均振幅都差不多，一个现实就可近似代表整个随机过程的属性。
- **右图**：每个现实都围绕不同的数学期望波动，且振幅也不一致，仅靠一个现实无法代表整个随机过程的特性

类似左图的平稳随机过程，对于任意一个现实，只要观测时间足够长，就**可把该现实的时间平均作为整个随机过程总体均值的近似值**，具有这种性质的平稳

随机过程就称其具有“**各态历经性**”

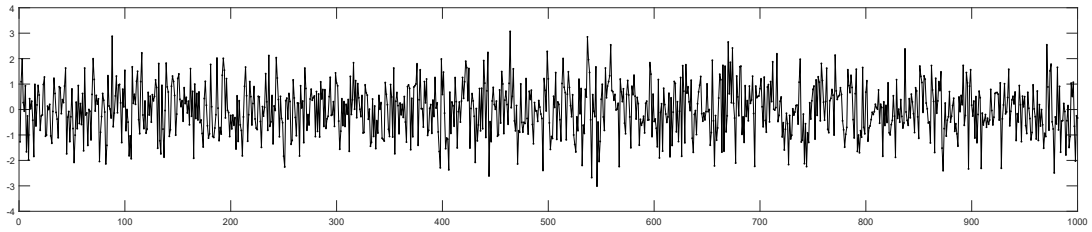
▲ 平稳随机是序列的应用——白噪声过程

对于一个**零均值**的随机过程 a，若其方差满足

$$K_a(t,\ t+\tau)=\begin{cases}\sigma_a^2 & (\tau=0)\\ 0 & (\tau\neq 0)\end{cases}$$

白噪声序列的特点：

表现在**任何两个时点的随机变量都不相关**，序列中没有任何可以利用的动态规律，因此**不能用历史数据对未来进行预测和推断**。



▲ 白噪声采样

4. 平稳时间序列的两个重要的统计量

(1) 时滞为 τ 的**自协方差函数**的估计

$$\hat{K}(\tau)=\frac{1}{n-\tau}\sum_{t=1}^{n-\tau}(x_t-\bar{x})(x_{t+\tau}-\bar{x})$$

(2) 时滞为 τ 的**自相关函数**的估计

$$\hat{\rho}(\tau)=\frac{\hat{K}(\tau)}{\sigma(t)\sigma(t+\tau)}=\frac{\hat{K}(\tau)}{s^2}=\frac{\frac{1}{n-\tau}\sum_{t=1}^{n-\tau}(x_t-\bar{x})(x_{t+\tau}-\bar{x})}{\frac{1}{n}\sum_{t=1}^{n}(x_t-\bar{x})^2}$$

[注]

- τ=0 时，自协方差函数为即为序列的方差，自相关函数为 1；
- 在自相关系数的计算中，$\sigma(t)=\sigma(t+\tau)=s$，因为对于同一个时间序列，无论截取多长，我们总是认为其方差和总时间序列的方差一致；
- 进一步理解自协方差的运算：

τ=1	1	2	3	4	5	6	7	8	9	10
	1	2	3	4	5	6	7	8	9	10

τ=2	1	2	3	4	5	6	7	8	9	10
	1	2	3	4	5	6	7	8	9	10

(上图为一个 1-10 的时间序列，填色部分为计算自协方差时用到的序列，可以发现基本是固定首尾进行计算)

▲ 关于计算两个时间序列的超前滞后的协方差

τ=1 (a超前b一个时间单位)	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10
	b1	b2	b3	b4	b5	b6	b7	b8	b9	b10

τ=-2 (a滞后b两个时间单位)	a1	a2	a3	a4	a5	a6	a7	a8	a9	a10
	b1	b2	b3	b4	b5	b6	b7	b8	b9	b10

(填色部分为计算使用的序列)

三、平稳时间序列的预报

1. 一阶自回归模型

表示要素在某一时刻与**前一时刻**之间的线性回归模型，称为一阶自回归模型，记为 AR(1)，对随机时间序列 x_t(设**已中心化**，其均值为 0，方便公式推导)，有：

$$x_t = \varphi_1 x_{t-1} + a_t$$

式中，φ_1为回归系数，a_t为白噪音，表示 t 时刻的随机因素对x_t的贡献，用前一时刻的 x_{t-1} 乘以上式两边，然后取数学期望，得：

$$E(x_t x_{t-1}) = \hat{\varphi}_1 E(x_{t-1} x_{t-1}) + E(x_{t-1} a_t) \Leftrightarrow \frac{E(x_t x_{t-1})}{E(x_{t-1} x_{t-1})} = \hat{\varphi}_1 + \frac{E(x_{t-1} a_t)}{E(x_{t-1} x_{t-1})}$$

从而得到 $\hat{\varphi}_1 = \rho_1$，ρ_1 **表示时滞为 1 的自相关函数。符合一阶自回归模型的随机过程称为红噪声过程。**

▲ 时滞为 τ 时的一阶自回归模型

时滞为 τ 的一阶自回归模型可以看成一个递推公式，对于第 t-1 时刻，有 $x_{t-1} = \rho_1 x_{t-2} + a_{t-1}$，回代 $x_t = \rho_1 x_{t-1} + a_t$，有：

$$x_t=\rho_1\left(\rho_1x_{t-2}+a_{t-1}\right)+a_t=\rho_1^2x_{t-2}+a_t+\rho_1a_{t-1}$$

依次类推，第 t 时刻的 x_t 与第 t-τ 时刻的 $x_{t-\tau}$ 的关系可表示为：

$$x_t=\rho_1^\tau x_{t-\tau}+a_t+\rho_1a_{t-1}+\rho_1^2a_{t-2}+\cdots+\rho_1^{\tau-1}a_{t-\tau+1}=\rho_1^\tau x_{t-\tau}+\sum_{k=0}^{\tau-1}\rho_1^ka_{t-k}$$

用前τ时刻的$x_{t-\tau}$乘以上式两边，然后取数学期望，得

$$E\left(x_tx_{t-\tau}\right)=\rho_1^\tau E\left(x_{t-\tau}x_{t-\tau}\right)\Leftrightarrow\rho_1^\tau=\frac{E\left(x_tx_{t-\tau}\right)}{E\left(x_{t-\tau}x_{t-\tau}\right)}=\rho_\tau$$

当$\tau\to0$时，有$\lim\limits_{\tau\to0}\rho_1^\tau=0$，此时有

$$x_t\approx a_t+\rho_1a_{t-1}+\rho_1^2a_{t-2}+\cdots+\rho_1^{\tau-1}a_{t-\tau+1}$$

这表明，**某时刻的要素还可看成是前期无穷个白噪声共同影响的结果**

2. *p* 阶自回归模型(类比多元回归方程学习)

设具有各态历经性的**标准化**平稳时间序列为：x_1, x_2, x_3 ，…， x_n。要预报第 t 时刻的值，利用前期第 t-1, t-2, …, t-p, 共 p 个时刻的值作为因子变量 p 阶自回归模型 AR(p)可以写为：

$$x_t=\varphi_1x_{t-1}+\varphi_2x_{t-2}+\ldots+\varphi_px_{t-p}+a_t$$

上述的自回归系数通过尤拉-沃克(***Yule-Walker***)方程进行求解：

$$\begin{bmatrix}1&\rho_1&\rho_2&\cdots&\rho_{p-1}\\\rho_1&1&\rho_1&\cdots&\rho_{p-2}\\\rho_2&\rho_1&1&\cdots&\rho_{p-3}\\\vdots&\vdots&\vdots&\vdots&\vdots\\\rho_{p-1}&\rho_{p-2}&\rho_{p-3}&\cdots&1\end{bmatrix}\begin{bmatrix}\varphi_1\\\varphi_2\\\varphi_3\\\vdots\\\varphi_p\end{bmatrix}=\begin{bmatrix}\rho_1\\\rho_2\\\rho_3\\\vdots\\\rho_p\end{bmatrix}$$

[注]

- 在多元回归方程中，标准化距平的回归系数与距平资料的回归系数存在关系式：$b_j^*=\frac{\sqrt{s_{jj}}}{\sqrt{s_{yy}}}b_j\left(j=1,2,\ldots,m\right)$，而在此处多元自回归方程中，由于是同一个时间序列，所以任何两个子序列都具有相同的标准差，所以自回归**标准化变量自回归系数=距平变量自回归系数**；

➢ 多元自回归方程的显著性检验同多元回归一样，采用 ***F* 分布**。

对一阶自回归模型的进一步认识

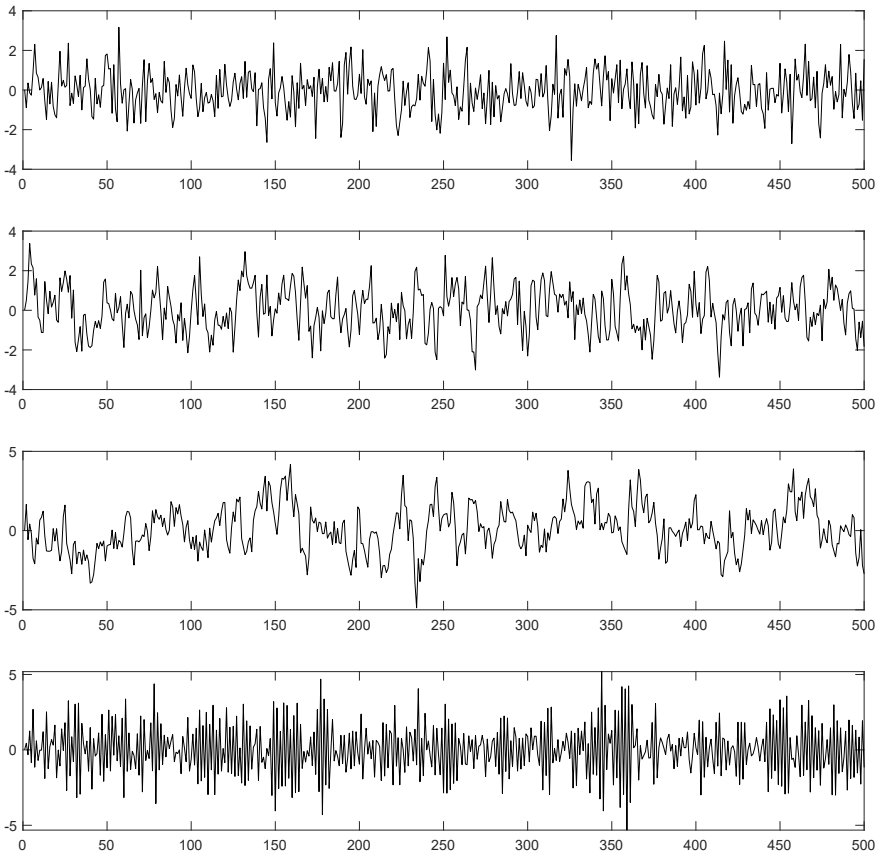
已知下列 4 个时间序列，将其与相应的图片进行匹配。

(a) 白噪声

(b) AR(1): $x_t = 0.8x_{t-1} + a_t$

(c) AR(1): $x_t = 0.5x_{t-1} + a_t$

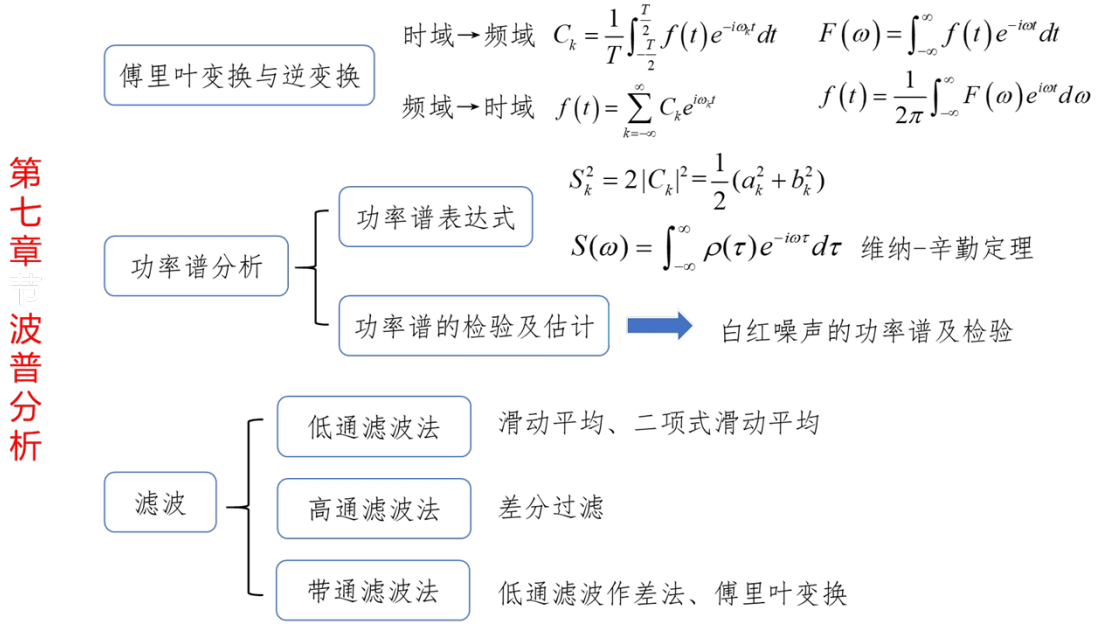
(d) AR(1): $x_t = -0.8x_{t-1} + a_t$



自上而下：依次为：a, c, b, d。 观察四个选项，发现(b)(c)的表达式是相近的，再来看图，第二个图与第三个图比较相近。对于一阶自回归模型的回归系数，有关系式$\rho_\tau = \rho_1^\tau$，如果时滞为 1 时的自相关系数较大，那么从t到$t+\tau$这段时间内，自相关系数ρ_1^τ减弱得比较慢，即某个时刻，观测值为正异常，那么在相对较长的一段时间里，这种正异常将会保持，在图上会有一写比较“宽”的部分。

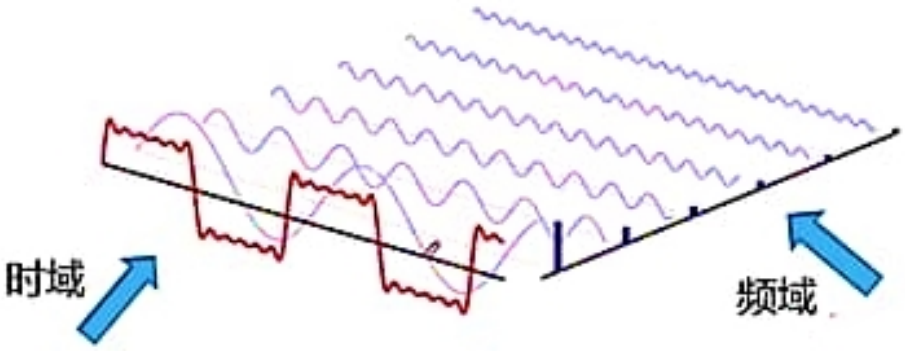
第七章 波谱分析

➢ *章节思维导图*



一、傅里叶变换与逆变换

1. 对傅里叶变换逆变换的理解



➢ **在时域方向看：**

只能看到一个随时间变化的时域信号，看不到信号背后包含多少个正余弦函数；

➢ **从频域方向看：**

(1) 这个信号究竟包含哪些频率分量（正余弦函数）

(2) 每一个频率分量的幅值是多少（幅度）

(3) 每一个频率分量的起始点（相位）

2. 离散形式的傅里叶变换与逆变换

三角形式的傅里叶级数展开为 $f(t)=a_0+\sum_{k=1}^{\infty}(a_k\cos\omega_k t+b_k\sin\omega_k t)$，其中，

- k为波数(时间段 T 内含有的波的个数), 因此，波数为 k 的谐波(第 k 谐波)周期 $T_k=\frac{T}{k}$
- 圆频率和频率的关系：$\omega_k=2\pi f=2\pi\frac{k}{T}$
- 傅里叶系数的求解公式：

$$\begin{cases} a_0=\dfrac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}}f(t)dt \\ a_k=\dfrac{2}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}}f(t)\cos\omega_k t dt \\ b_k=\dfrac{2}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}}f(t)\sin\omega_k t dt \end{cases}$$

根据欧拉公式 $e^{i\theta}=\cos\theta+i\sin\theta$，可以得到：$\cos\theta=\frac{e^{i\theta}+e^{-i\theta}}{2},\sin\theta=\frac{e^{i\theta}-e^{-i\theta}}{2}$，

将其回代到傅里叶三角级数展开式中：

$$f(t)=a_0+\sum_{k=1}^{\infty}\left(\frac{a_k-ib_k}{2}e^{i\omega_k t}+\frac{a_k+ib_k}{2}e^{-i\omega_k t}\right)$$

令 $C_0=a_0, C_k=\frac{a_k-ib_k}{2}$, $C_{-k}=\frac{a_k+ib_k}{2}$, 根据傅里叶系数的基本计算式和$\omega_k=2\pi\frac{k}{T}$

可知，$C_k=C_{-k},\omega_k=-\omega_k$，因此，可以写成：$f(t)=\sum_{k=-\infty}^{\infty}C_k e^{i\omega_k t}$，称 $C_k=\frac{a_k-ib_k}{2}$

为“复(数)谱”。

把 $f(t)=\sum_{k=-\infty}^{\infty}C_k e^{i\omega_k t}$ 同时乘以 $e^{-i\omega_{k'}t}$，有 $\int_{-\frac{T}{2}}^{\frac{T}{2}}f(t)e^{-i\omega_{k'}t}dt=\sum_{k=-\infty}^{\infty}C_k\int_{-\frac{T}{2}}^{\frac{T}{2}}e^{i(\omega_k-\omega_{k'})t}dt$，

根据傅里叶函数的正交性：$\int_{-\frac{T}{2}}^{\frac{T}{2}}e^{i(\omega_k-\omega_{k'})t}dt=\begin{cases}T, & k=k'\\ 0, & k\neq k'\end{cases}$，得到复谱的表达式如下：

$$C_k=\frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}}f(t)e^{-i\omega_k t}dt$$

3. 连续形式的傅里叶变换与逆变换

$$F(\omega)=\int_{-\infty}^{\infty}f(t)e^{-i\omega t}dt;\ f(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}F(\omega)e^{i\omega t}d\omega$$

[小结]

- **傅里叶变换：从时域到频域转化**
- **傅里叶逆变换：从频域到时域的转化**

二、功率谱分析

进行功率谱分析的目的：找到其主要波动的谐波

1. 功率谱的概念

记$X(\omega)$为$x(t)$的傅里叶变换，当$T\to\infty$时，平均功率S可以在频域表示为

$$S=\lim_{T\to\infty}\frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}}|x(t)|^2\,dt=\lim_{T\to\infty}\frac{1}{2\pi T}\int_{-\infty}^{\infty}|X(\omega)|^2\,d\omega=\int_{-\infty}^{\infty}\lim_{T\to\infty}\frac{|X(\omega)|^2}{2\pi T}d\omega$$

定义$S_\omega=\lim\limits_{T\to\infty}\frac{|X(\omega)|^2}{2\pi T}$为 $x(t)$的功率谱密度。

下面来求时间函数$f(t)$的功率谱表达式。已知$f(t)=\sum\limits_{k=-\infty}^{\infty}C_k e^{i\omega_k t}=\sum\limits_{k'=-\infty}^{\infty}C_{k'}e^{i\omega_{k'}t}$

(因为区间为$(-\infty,+\infty)$所以$f(t)$可以分别用用k,k'表示)，于是：

$$\frac{1}{T}\int_{-T/2}^{T/2}[f(t)]^2\,dt=\frac{1}{T}\int_{-T/2}^{T/2}\sum_{k=-\infty}^{\infty}\sum_{k'=-\infty}^{\infty}C_kC_{k'}e^{i(\omega_k+\omega_{k'})t}dt=\sum_{k=-\infty}^{\infty}\sum_{k'=-\infty}^{\infty}C_kC_{k'}\frac{1}{T}\int_{-T/2}^{T/2}e^{i(\omega_k+\omega_{k'})t}dt$$

利用$\frac{1}{T}\int_{-T/2}^{T/2}e^{i(\omega_k+\omega_{k'})t}dt=\begin{cases}1, & (k+k'=0)\\ 0, & (k+k'\neq 0)\end{cases}$，上述式子可进一步化简为：

$$\frac{1}{T}\int_{-T/2}^{T/2}[f(t)]^2dt=\sum_{k=-\infty}^{\infty}C_kC_{-k}=\sum_{k=-\infty}^{\infty}C_kC_k^*\quad(C_k^*\text{为}C_k\text{的共轭})$$

对于一个复数，它本身与其共轭的乘积为模的平方，故有：

$$C_kC_k^*=|C_k|^2=\frac{1}{4}(a_k^2+b_k^2)$$

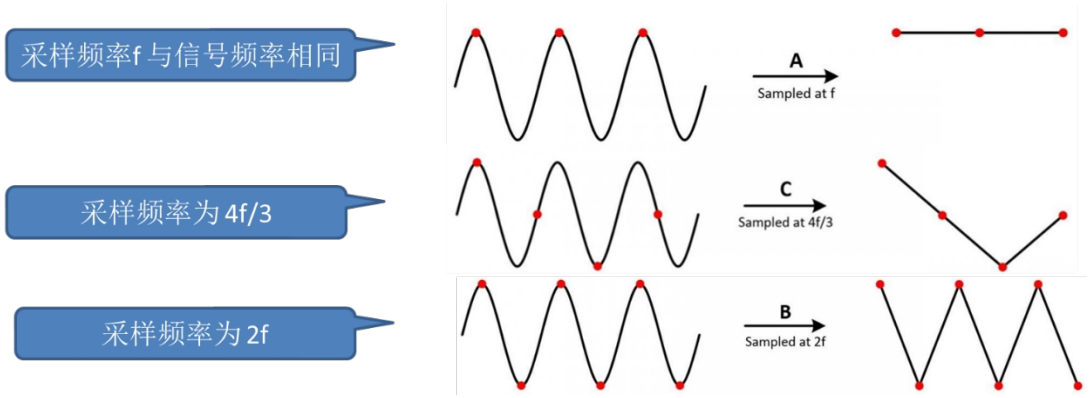
由于振幅谱$|C_{-k}^2|$与$|C_k^2|$相同，离散功率谱正负半轴对称，因此可以将谱值乘以

2，只画正半轴的部分，此时离散功率谱的表达式为：

$$S_k^2 = 2|C_k|^2 = \frac{1}{2}(a_k^2 + b_k^2)$$

而平均功率也可以写成：$\frac{1}{T}\int_{-T/2}^{T/2}[f(t)]^2 dt = \sum_{k=-\infty}^{\infty} C_k C_k^* = 2\sum_{k=0}^{\infty}|C_k^2| = \sum_{k=0}^{\infty} S_k^2$

▲ 奈奎斯特采样定理



采样频率应不小于($\geq$)信号频率的 2 倍，才能获得准确的信号周期，否则会产生混叠效应（虚假周期）

2. 离散功率谱的估算及检验

对于离散功率谱，最终是要计算 $S_k^2 = \frac{1}{2}\left(a_k^2 + b_k^2\right)$，下面给出一个计算步骤。

设一实测的气象时间序列为 $x_1, x_2, \ldots, x_n$,

(1) 针对不同的波数计算对应的傅里叶系数；

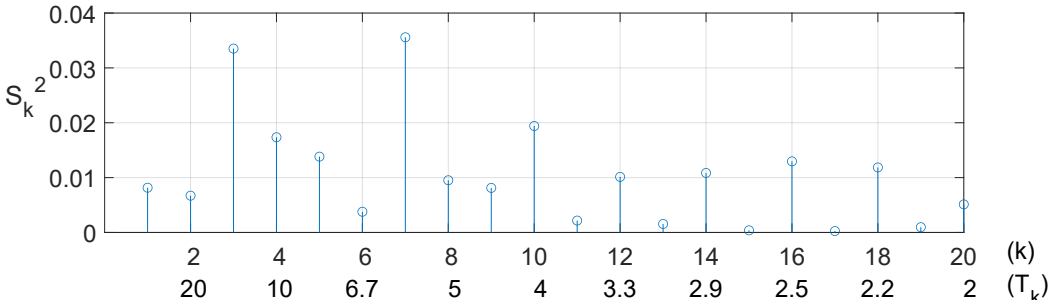
$$\begin{cases} a_0 = \frac{1}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)dt \\ a_k = \frac{2}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)\cos\omega_k t dt \\ b_k = \frac{2}{T}\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)\sin\omega_k t dt \end{cases} \xrightarrow{\text{改成离散形式}} \begin{cases} a_k = \frac{2}{n}\sum_{t=1}^{n} x_t \cos 2\pi\frac{k}{n}t \\ b_k = \frac{2}{n}\sum_{t=1}^{n} x_t \sin 2\pi\frac{k}{n}t \end{cases} \left(k = 1,2,\ldots,\left[\frac{n-1}{2}\right]\right)$$

(2) 求出 a_k 和 b_k 后，计算各谐波的功率谱值；

(3) 以波数为横坐标，对应的功率普值为纵坐标绘制功率谱图。在谱图上，为了方便分析，常在波数坐标下面标上对应的周期或频率值。$T_k = \frac{n}{k} = \frac{1}{f_k}$

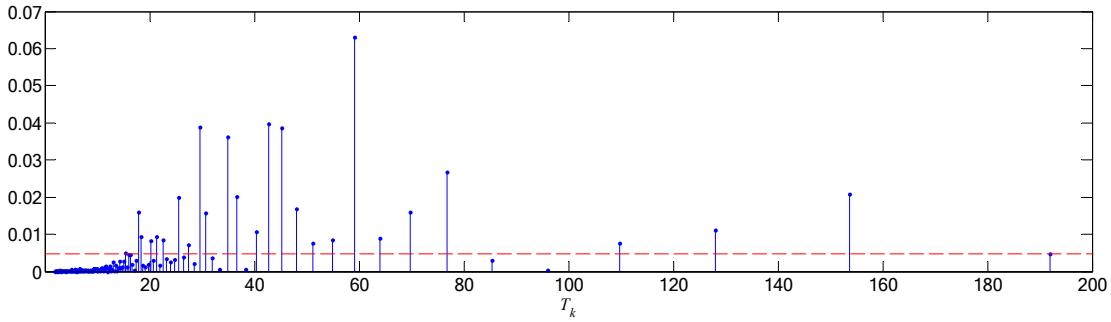
(4) 对离散功率谱进行检验。要检验某一频率(或波数 k)的谱值 $S_k^2=(a_k^2+b_k^2)/2$ 所

对应的周期 $T_k=n/k$ 是否显著，原假设 H_0 为： $E(a_k)=E(b_k)=0$，检验统计量采用：$F=\dfrac{s_k^2/2}{\left(s^2-s_k^2\right)/(n-2-1)}\sim F(2,n-2-1)$，其中 s^2 为原序列的方差。有时候采用临界谱值检验的方法更加方便：$S_c^2=\dfrac{2F_c s^2}{2F_c+(n-2-1)}$。



▲ 功率谱图示意

功率谱值越大，对应的周期 T_k 越显著。



▲ 1950-2013 年 Nino3.4 区月平均海温序列的离散功率谱图(横轴为月份)

由上图可以知道，El Nino 具有 2-7 年的周期

3. 连续功率谱的估算及检验

(1) 连续功率谱的功率谱密度函数

$$F(\omega)=\int_{-\infty}^{\infty}f(t)e^{-i\omega t}dt \quad \text{（傅里叶变换，从时域到频域）}$$

$$f(t)=\frac{1}{2\pi}\int_{-\infty}^{\infty}F(\omega)e^{i\omega t}d\omega \quad \text{（逆变换，从频域到时域）}$$

$$\int_{-\infty}^{\infty}[f(t)]^2dt=\int_{-\infty}^{\infty}f(t)[\frac{1}{2\pi}\int_{-\infty}^{\infty}F(\omega)e^{i\omega t}d\omega]dt=\frac{1}{2\pi}\int_{-\infty}^{\infty}F(\omega)[\int_{-\infty}^{\infty}f(t)e^{i\omega t}dt]d\omega$$
$$=\frac{1}{2\pi}\int_{-\infty}^{\infty}F(\omega)F(-\omega)d\omega$$

从而连续功率谱的密度函数为：$S(\omega)=F(\omega)F(-\omega)$

(2) 连续功率谱的估算(维纳—辛勤定理)

设 x_t 为标准化时间序列，数学期望为 0， 方差为 1，则它的自相关函数为：

$$\rho(\tau)=\int_{-\infty}^{\infty}x(t)x(t+\tau)dt$$

上式中 τ 为滞后时间长度，将 $x(t+\tau)$ 用频率积分的形式表示，有：

$$\rho(\tau)=\int_{-\infty}^{\infty}x(t)[\frac{1}{2\pi}\int_{-\infty}^{\infty}F(\omega)e^{i\omega(t+\tau)}d\omega]dt=\int_{-\infty}^{\infty}x(t)[\frac{1}{2\pi}\int_{-\infty}^{\infty}F(\omega)e^{i\omega t}e^{i\omega\tau}d\omega]dt$$
$$=\int_{-\infty}^{\infty}F(\omega)[\frac{1}{2\pi}\int_{-\infty}^{\infty}x(t)e^{i\omega t}dt]e^{i\omega\tau}d\omega$$
$$=\frac{1}{2\pi}\int_{-\infty}^{\infty}F(\omega)F(-\omega)e^{i\omega\tau}d\omega$$

因此可以得到**维纳—辛勤定理：对平稳过程，功率谱密度 *S*(ω)和自相关函数 *ρ*(τ) 是一傅氏变换对。**

$$\begin{cases}\rho(\tau)=\dfrac{1}{2\pi}\displaystyle\int_{-\infty}^{\infty}S(\omega)e^{i\omega\tau}d\omega\\ S(\omega)=\displaystyle\int_{-\infty}^{\infty}\rho(\tau)e^{-i\omega\tau}d\tau\end{cases}$$

(3) 连续功率谱的计算步骤及检验

- 需首先计算样本的自相关系数 $r(\tau)$ ，(τ= 0, 1, 2, …, m)， m 为最大落后长度(一般取 m 取在 $n/3$~$n/10$ 之间)；
- 利用“梯形法”对 τ 求积分，计算粗谱；
- 把上一步计算的粗谱进行**平滑**，以减小误差；

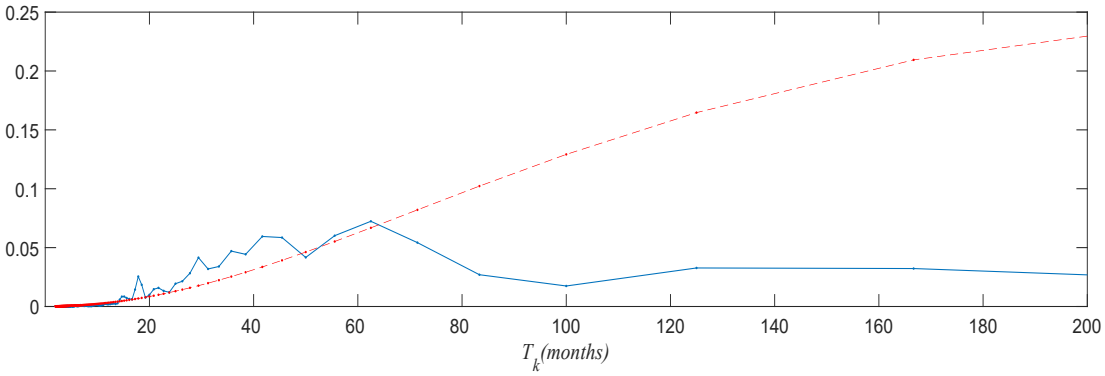
$$\begin{cases}S_0=\dfrac{1}{2}\hat{S}_0+\dfrac{1}{2}\hat{S}_1\\ S_l=\dfrac{1}{4}\hat{S}_{l-1}+\dfrac{1}{2}\hat{S}_l+\dfrac{1}{4}S_{l+1}\qquad(1\le l\le m-1)\\ S_m=\dfrac{1}{2}\hat{S}_{m-1}+\dfrac{1}{2}\hat{S}_m\end{cases}$$

- 以波数 k（或频率 f、周期 T）为横坐标轴，以平滑功率谱密度估计值 S_l 为纵

坐标，作谱图，波数与周期的关系为：$\begin{cases}\omega_k=\dfrac{\pi k}{m}\\ T_k=\dfrac{2m}{k}\end{cases}$

- 假设所要检验的某一气象要素的总体谱为某一非周期性的随机过程谱，在这一假设成立的条件下，**某一频率的谱估计值与所假设的过程的平均谱估计之比**，遵从自由度为 f 的 χ^2 分布，统计量为：$\dfrac{S_k}{\overline{S}_k/f}\sim\chi^2(f)\left(f=\dfrac{(2n-\frac{m}{2})}{m}\right)$

  ($\overline{S}_k$ 为所假设的非随机过程的第 k 个波的功率谱的平均值)，常常也采用临界值检验的方法进行检验比较方便。



▲ Nino3.4 指数的功率谱及临界谱值(红线)

可见 ENSO 的显著周期为 2-5 年

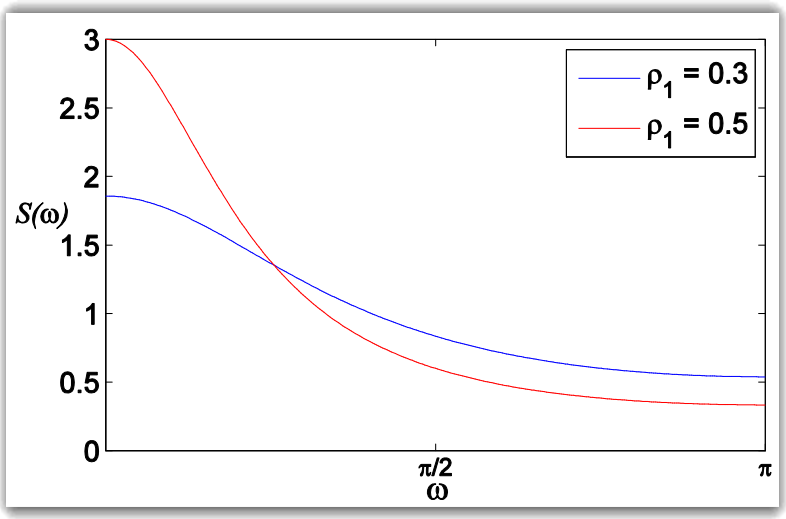
(4) 红噪声的功率谱

红噪声的功率谱可以写为：$S(\omega)=\int_{-\infty}^{\infty}\rho(\tau)e^{-i\omega\tau}d\tau=\int_{-\infty}^{\infty}\rho_1^{|\tau|}e^{-i\omega\tau}d\tau$，令 $z=e^{i\omega}$，将上述积分形式写成级数求和的形式，有：

$$S(\omega)=\sum_{\tau=-\infty}^{\infty}\rho_1^{|\tau|}z^{-\tau}$$

类比几何级数：$\sum_{n=1}^{\infty}x^n=\dfrac{1}{1-x}$，分别讨论当 $\tau<0,\tau=0,\tau>0$ 时的情形，上述式子可以继续化简为：

$$S(\omega,\rho)=1+\left(\frac{1}{1-\rho_1 z}-1\right)+\left(\frac{1}{1-\rho_1 z^{-1}}-1\right)$$
$$=\frac{1}{1-\rho_1 z}+\frac{1-(1-\rho_1 z^{-1})}{1-\rho_1 z^{-1}}=\frac{1-\rho_1 z^{-1}+\rho_1 z^{-1}(1-\rho_1 z)}{(1-\rho_1 z)(1-\rho_1 z^{-1})}$$
$$=\frac{1-\rho_1^2}{1-\rho_1 z^{-1}-\rho_1 z+\rho_1^2}=\frac{1-\rho_1^2}{1-\rho_1(e^{-i\omega}+e^{i\omega})+\rho_1^2}$$
$$=\frac{1-\rho_1^2}{1-2\rho_1\cos\omega+\rho_1^2}$$



▲ 红噪声功率谱

可以取 $\omega=\frac{\pi}{2}$，得到 $S\left(\frac{\pi}{2},\rho_1\right)=1-\frac{2}{1+\frac{1}{\rho_1^2}}$（与 ρ_1 反比），据此可以判断哪一个是自相关系数较大的红噪声过程。

▲ 白噪声的功率谱密度为常数

三、滤波方法

1. 脉冲响应函数与滤波器

用一脉冲函数 $\delta(t)$作为输入，经过滤波器系统后，它的输出记为 $h(t)$， 称为

脉冲响应。若用 L 表示线性系统由 $\delta(t)$ 到 $h(t)$ 的转换，可记作： $h(t)=L[\delta(t)]$。

现在对任意时间函数 $f(t)$，根据脉冲函数的筛选性质（$\int_{-\infty}^{\infty}\delta(t-t_0)f(t)dt=f(t_0)$），可表示成与脉冲函数乘积的积分形式：

$$g(t)=L[f(t)]=\int_{-\infty}^{\infty}f(\tau)L[\delta(t-\tau)]d\tau=\int_{-\infty}^{\infty}f(\tau)h(t-\tau)d\tau$$

称 $h(t)$为系统的“**脉冲响应函数**”(滤波器的核心)

把脉冲响应函数 $h(t)$的谱记为 $H(\omega)$，其中 $H(\omega)=\int_{-\infty}^{\infty}h(t)e^{-i\omega t}dt$，记 $F(\omega)$和 $G(\omega)$分别为输入函数 $f(t)$和输出函数 $g(t)$的谱，对于输出谱 $G(\omega)$，可以写成：

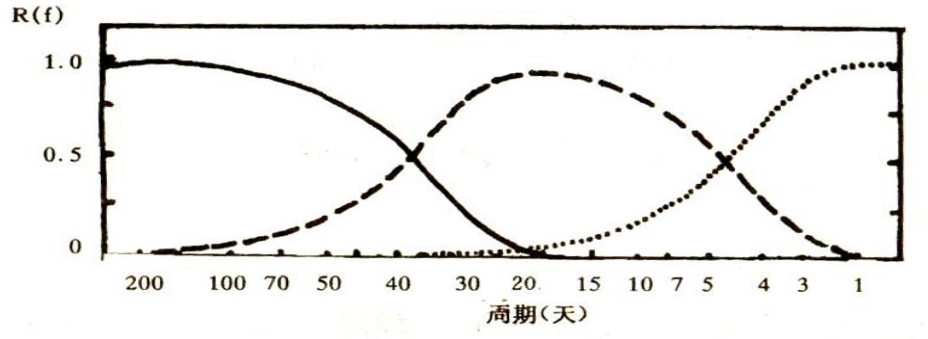
$$\begin{aligned}G(\omega)&=\int_{-\infty}^{\infty}g(t)e^{-i\omega t}dt=\int_{-\infty}^{\infty}[\int_{-\infty}^{\infty}f(\tau)h(t-\tau)d\tau]e^{-i\omega t}dt\\&=\int_{-\infty}^{\infty}f(\tau)[\int_{-\infty}^{\infty}h(t-\tau)e^{-i\omega(t-\tau)}d(t-\tau)]e^{-i\omega\tau}d\tau=\int_{-\infty}^{\infty}f(\tau)H(\omega)e^{-i\omega\tau}d\tau\\&=H(\omega)\int_{-\infty}^{\infty}f(\tau)e^{-i\omega\tau}d\tau\\&=H(\omega)F(\omega)\end{aligned}$$

于是输出函数的功率谱密度为：

$$S_g(\omega)=2|G(\omega)|^2=2|H(\omega)|^2|F(\omega)|^2=|H(\omega)|^2S_f(\omega)\Leftrightarrow|H(\omega)|^2=\frac{S_g(\omega)}{S_f(\omega)}$$

上式表明：某一频率 ω 的振动，经过滤波后， 振动方差变为原来的$|H(\omega)|^2$倍；若$|H(\omega)|<1$，表明该频率的振动方差有所削弱。若$|H(\omega)|=0$，表明该频率的振动完全被削除。

2. 常见的滤波方法



(1) **低通(low-pass)滤波**：使过滤后的序列主要包含低频振动分量，把高频的分量

滤掉，响应函数如图中实线(‘—’)所示；

(2) **高通(high-pass)滤波**：使过滤后的序列主要包含高频振动分量，把低频的分量滤掉，响应函数如图中点线(‘…’)所示；

(3) **带通(band-pass)滤波**：把高频和低频的信号都滤掉，只保留某一段频带的信号，响应函数如图中虚线(‘---’)所示。

3. 低通滤波器

(1) 等权滑动平均

已知 $g(t)=\int_{-\infty}^{\infty} f(\tau)h(t-\tau)d\tau$，令 $\tau=t+\lambda$，有 $g(t)=\int_{-\infty}^{\infty} x(t+\lambda)h(\lambda)d\lambda$，截取滑动长度 k，则上述积分可写成求和形式：$g(t)=\sum_{i=-k}^{k} h_i x_{t+i}$，其中 h_i 称为滑动权重系数，满足条件 $\sum_{i=-k}^{k} h_i=1$，我们通过响应函数谱来考察经过滑动平均后不同频率的削弱情况：

$$H(\omega)=\int_{-\infty}^{\infty} h(t)e^{-i\omega t}dt \approx \sum_{j=-k}^{k} h_j e^{-i\omega j}=\sum_{j=-k}^{k} h_j\cos\omega j - i\sum_{j=-k}^{k} h_j\sin\omega j=\sum_{j=-k}^{k} h_j\cos\omega j+0$$

$$=h_0+2\sum_{j=1}^{k} h_j\cos\omega j \Leftrightarrow H(f)=h_0+2\sum_{j=1}^{k} h_j\cos 2\pi fj$$

对于“等权重”的 ***m*** 点滑动，滑动窗口的长度 ***m*=2*k*+1**，权重为 $h_j=\frac{1}{2k+1}=\frac{1}{m}$，于是 $H(f)=\frac{1}{m}\left[1+2\sum_{j=1}^{k} h_j\cos 2\pi fj\right]$，对 $2\sum_{j=1}^{k} h_j\cos 2\pi fj$ 进一步化简如下：

$$2\sum_{j=1}^{k}\cos 2\pi fj=\frac{\sum_{j=1}^{k} 2\cos 2\pi fj\sin\pi f}{\sin\pi f}=\frac{\sum_{j=1}^{k}[\sin\pi f(2j+1)-\sin\pi f(2j-1)]}{\sin\pi f}$$

$$=\frac{\sin\pi f(2k+1)-\sin\pi f}{\sin\pi f}=\frac{\sin\pi f(2k+1)}{\sin\pi f}-1=\frac{\sin\pi fm}{\sin\pi f}-1$$

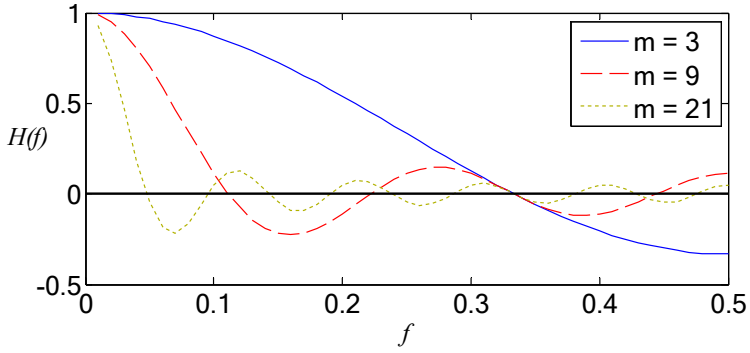
从而可以得到 $H(f)=\frac{\sin\pi fm}{m\sin\pi f}$

对上述导出的公式做一个简单的讨论：

◆ 对于**周期等于滑动间隔**(m)的振动，由于频率为 f=1/m，因此，$H(f)$=0,表示这

种振动分量达到完全削除；

- 对于周期等于 m/i 的高频振动(频率为 i/m，i=1,2,3,…)，也有 $H(f)$=0，振动达到完全削除；
- 对于周期大于 m 的低频振动，频率 f<1/m，因此响应 $H(f)$<1，这类周期也有不同程度的削弱，周期越大，削弱程度越小；对于无限长的周期(f→0)，频率响应 $H(f)$—>1， 因此，过滤后无任何削弱。

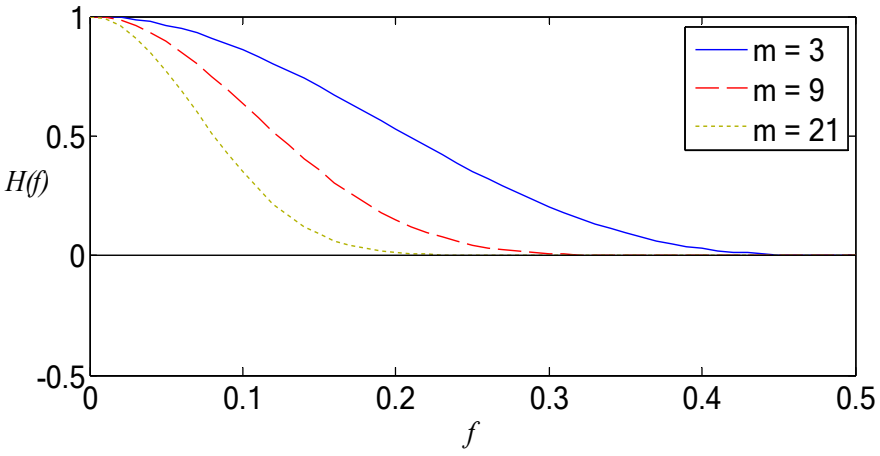


(2) 二项式系数滑动(不等权)

将二项式的系数作为滑动平均的系数权重，其中 m 点(m=2k+1)滑动对应于 $(a+b)^{m-1}$（即$(a+b)^{2k}$）的二项式系数(记住两个常见的)：

- 三点滑动平均：m=3 时，k=1，二项式 $(a+b)^2$ 的系数为：1, 2, 1，故权重系数为：1/4, 1/2, 1/4；
- 五点滑动平均：m=5 时，k=2，二项式$(a+b)^4$ 的系数为：1, 4, 6 , 4, 1，故权重系数为：1/16, 4/16, 6/16, 4/16, 1/16

二项式系数滑动的频率响应函数为$H(f)=\cos^m(\pi f)$，频率响应函数图如下：



可见，二项式系数滑动与等权滑动相比，是更有效的低通滤波器：能够通过更多的低频信号，同时也能滤掉更多的高频信号。

4. 高通滤波器

(1) **一阶差分过滤**

一阶差分滤波可以表示为：$g(t)=\nabla f(t)=f(t)-f(t-1)$，有：

$$\begin{aligned}G(\omega)&=\int_{-\infty}^{+\infty}g(t)e^{-i\omega t}dt=\int_{-\infty}^{+\infty}\left[f(t)-f(t-1)\right]e^{-i\omega t}dt\\&=\int_{-\infty}^{+\infty}f(t)e^{-i\omega t}dt-\int_{-\infty}^{+\infty}f(t-1)e^{-i\omega t}dt\\&=F(\omega)-e^{-i\omega}\int_{-\infty}^{+\infty}f(t-1)e^{-i\omega(t-1)}d(t-1)\\&=F(\omega)\left(1-e^{-i\omega}\right)\end{aligned}$$

由上式可知一阶差分过滤的响应函数为：$H(\omega)=1-e^{-i\omega}$，它的模为：

$$|H(f)|=|1-\cos\omega+i\sin\omega|=\sqrt{(1+\cos\omega)^2+\sin^2\omega}=\sqrt{2(1-\cos\omega)}=2\left|\sin\frac{\omega}{2}\right|=2|\sin\pi f|$$

(2) 二阶及高阶差分过滤

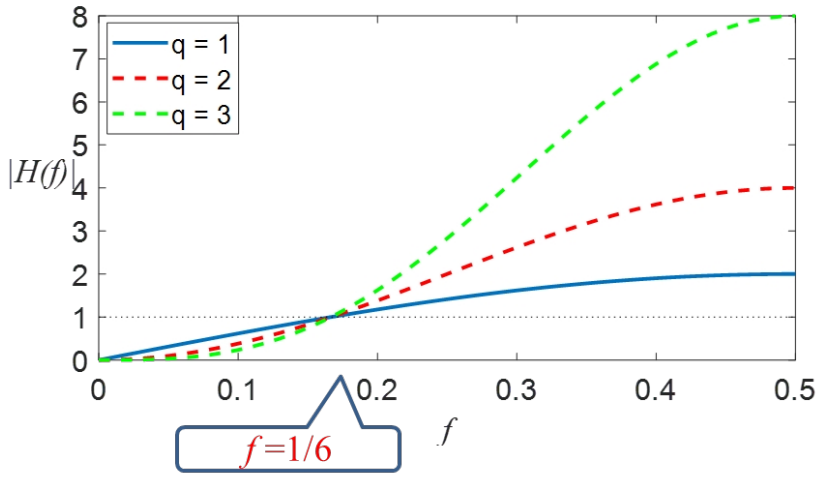
对于二阶差分过滤，有

$$g(t)=\nabla\left[\nabla f(t)\right]=\nabla\left[f(t)-f(t-1)\right]=f(t)-2f(t-1)+f(t-2)$$

可证明，对于“q 阶差分过滤”，响应函数为：

$$|H(f)|=\left(2|\sin\pi f|\right)^q$$

由于 f 的取值范围为：[0, 1/2]，因此，$\mathbf{0\le|H(f)|\le 2^q}$，低频的振动分量（$f\to 0$）得到削弱，高频分量（$f\to 0.5$）得到增强，如下图所示：



从上图可以看到，f=1/6 时（**周期= 6 倍采样间隔**）， $H(f)$=1，即振动分量无任何削弱;频率 f>1/6 的分量振幅增强；频率 f<1/6 的分量振幅削弱；随着差分过滤阶数的升高，对低频的削弱更强，但是对高频振幅的放大也更强。

5. 带通滤波器

(1) **采用两个简单的低通滤波器相减**

- 对原序列 x 先做 5 年滑动平均，削去周期小于 5 年的波，余下周期大于 5 年的波，得到新序列记为 x_1；
- 再把原序列 x 做 9 年滑动平均，除去周期小于 9 年的波，余下周期大于 9 年的波，得到的新序列记为 x_2；
- 令 $x_3 = x_1 - x_2$，得到的新序列 x_3 包含了 5-9 年周期的振动，相当于实现了带通滤波。

这种滤波器的响应函数为：$\mathbf{H_3(f) = H_1(f) - H_2(f)}$

(2) **利用傅里叶逆变换进行带通滤波**

我们滤波时想保留频率在 $[f_1, f_2]$ 之间的振动(f_1<f_2)，这时 f_1 和 f_2 可称为**截止(截断)频率**，可把响应函数设计为：

$$H(f)=\begin{cases}0 & f<f_1\\ 1 & f_1\le f\le f_2\\ 0 & f>f_2\end{cases}$$

最终目的是要得到$g(t)=\int_{-\infty}^{\infty}x(t+\lambda)h(\lambda)d\lambda$，具体步骤如下：

- 根据过滤要求，定义相应的分段函数H(f)；
- 计算脉冲响应函数$h(t)$权重系数；

$$h(t)=\int_{-\infty}^{\infty}H(f)e^{i2\pi ft}df=2\int_{0}^{\infty}H(f)\cos(2\pi ft)df$$

上式离散化进行求解，f的取值范围为：[0, 1/2]。f 的离散采样可取间隔为1/2n, 即 $\boldsymbol{f=0, 1/2n, \ 2/2n, \ 3/2n, \ \ldots, \ n/2n}$, 于是，用梯形法写出以上积分的离散形式(注意，响应函数具有对称性：h_i=h_{-i})：

$$h_i=\frac{1}{2n}[H(0)+2\sum_{f=\frac{1}{2n}}^{\frac{n}{2n}}H(f)\cos(2\pi fi)]$$

- 根据$g(t)=\int_{-\infty}^{\infty}x(t+\lambda)h(\lambda)d\lambda$，将其写成离散形式有$g_t=\sum_{i=-K}^{K}h_ix_{t+i},(t=1,2,...,n)$，从而得到滤波以后的值。