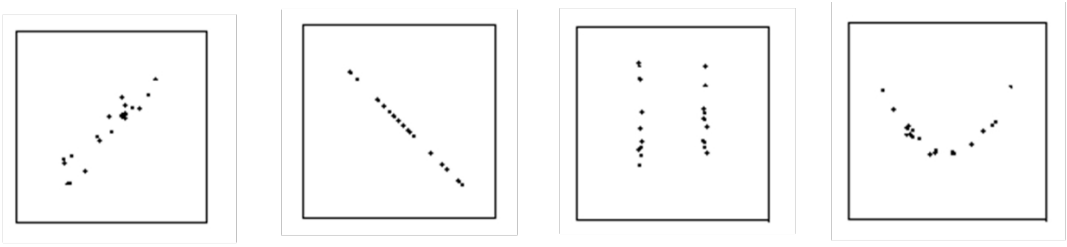
习题一

出题人：蒋斌

涉及章节：第一章《气象资料的基本整理》；第二章《回归分析》

一、 选择题

1. 下面对散点图的线性相关性表述错误的是_____.



A. 正相关(0<*r*<1)　B. 完全负相关(*r*=-1)　C. 零相关(*r*=0)　D. 非线性相关(*r*>0)

2. 在分析气候变化时，某站由 1980 年前 30 年实测资料计算均值得到平均气温为 25.1℃，再由后 20 年的实测资料计算得到平均气温为 24.8℃，若要确定该站气温在后 20 年是否存在显著差异，则采用什么分布检验？_____.

   A.t 分布　B.χ 分布　C.F 分布　D.正态分布

3. 某站一月降水量服从正态分布。根据 1980 年前 30 年实测资料计算得到降水量的标准差为$s_1 = 40mm$，再根据后 20 年的实测资料计算得到降水量的标准差为$s_2 = 37mm$，若要确定该站一月份降水的年际变化率是否存在显著差异，则采用的统计量服从____.

   A.t 分布　B.χ 分布　C.F 分布　D.正态分布

4. 已知变量 x 的方差为 4，其标准化距平 x_s 与距平变量 y_d 之间存在回归关系 $y_d = 1.5x_s$，当测得变量 x 的距平值为 $x_d = 0.2$ 时，变量 y 的距平估计 y_d = _____.

   A.0.075　B.0.15　C.3　D.0.3

5. 下列说法正确的是_____.

A. 随机变量 x 和 y 的相关系数为 r，则 $2x+1$ 与 $-3y+5$ 的相关系数也为 r

B. 某 m 个因子经过 n 次观测后得到的距平资料阵为 $\boldsymbol{X}_{n\times m}$，其协方差阵为 $\boldsymbol{S}=\frac{1}{n-1}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}$

C. 在逐步回归中，每剔除一个因子后，就需要再引入一个因子

D. 相关很显著是指显著性水平α值很小，两总体存在相关的概率很高

6. 对 m 个备选因子与预报量 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 $\boldsymbol{Z}_{n\times(m+1)}=[\boldsymbol{x}_1,\boldsymbol{x}_2,...,\boldsymbol{x}_m,\boldsymbol{y}]$，计算得到初始离差乘积阵 $\boldsymbol{S}^{(0)}$，经过逐步回归后，初始矩阵变为 $\boldsymbol{S}^{(l)}$，l 表示当前引入因子的个数。下列说法错误的是_____.

A. 引入或剔除某一个因子时，需要对 $\boldsymbol{S}$ 进行一次求解求逆紧凑变换

B. 当方差贡献最小的因子检验结果为不宜剔除时，表明所有不显著的因子都已被剔除，下一步应该进行“引入检验”

C. $\boldsymbol{S}^{(l)}$ 最后一列中 $S_{1y}^{(l)},S_{2y}^{(l)},...,S_{my}^{(l)}$ 表示的含义为第 l 次变换中各因子与预报量的离差乘积

D. 引入检验与剔除检验均用到 F 分布

7. [多选]下列说法正确的是_____.

A. 距平资料的均值为零，标准化资料的方差为 1

B. 标准化资料的协方差阵等同于相关系数阵

C. 一元回归方程中，当 y 和 x 都是距平资料时，回归方程可写为 $\hat{y}=bx$，此时 b 的含义表示当 x 的变化为 1 个单位时，y 距平的估计值

D. 一元回归中，回归变量 y 的离差平方和可以分解为回归平方和、剩余平方和，对于固定的样本容量 n，回归平方和越大表示回归的效果越好

8. [多选]下列说法正确的是_____.

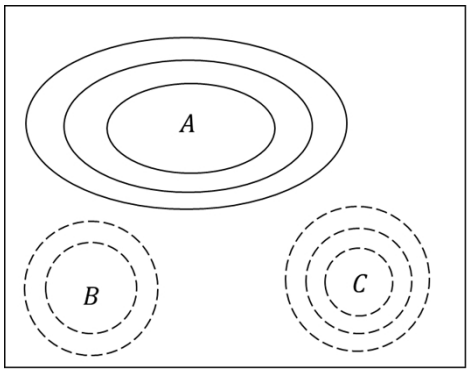
A. 在一元回归中，有 $\hat{y}_0 = \hat{\mu}_{y_0} = b_0 + bx_0$，对新观测值的估计与对回归函数值 $b_0 + bx_0$ 的估计是等价的

B. 一元回归模型的检验与相关系数的检验是等价的

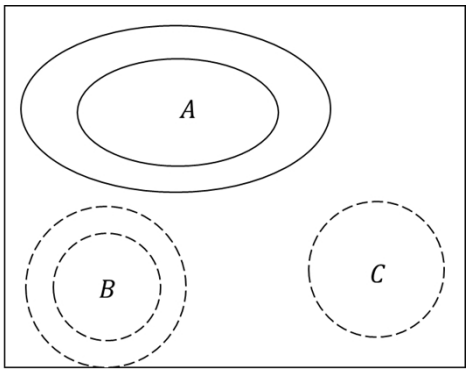
C. 若预报变量 y 与多个预报因子有关，为了考察单个因子与 y 的关系，可以将每个因子和 y 计算得到相应的简单相关系数

D. 在一元回归模型 $y = \beta x + \beta_0$ 中，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

9. [多选]图(a)表示某一时刻某地的气温的距平场分布，图(b)为该气温距平场对应的标准化距平场分布。图(a)中相邻等值线之间的差值为±0.5℃，图(b) 中相邻等值线之间的差值为±0.2℃。两幅图中，0℃等值线均已略去，黑色实线表示等值线为正值，虚线表示等值线的值为负值，下列说法正确的是_____.



(a)



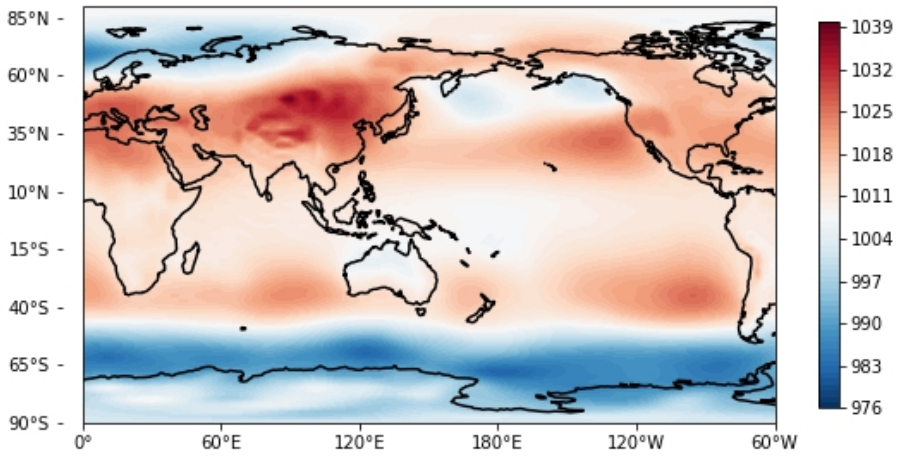
(b)

A.B 点的标准差比 A 点小　　B. C 点的方差最大

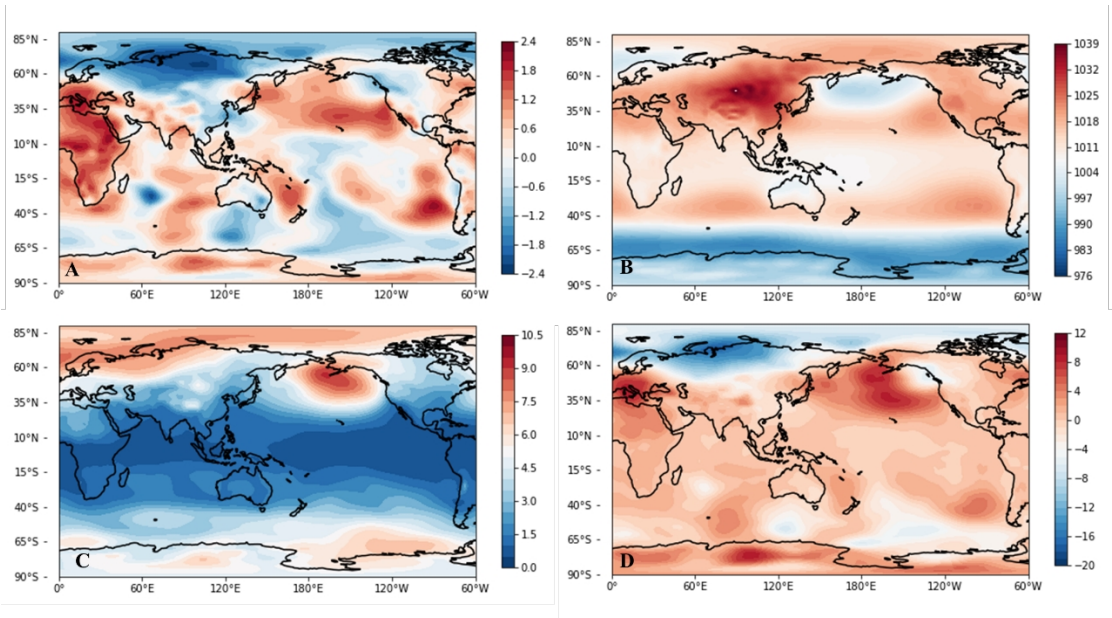
C. A 点的气温和 C 点的气温呈负相关　　D. B 点的气温和 C 点的气温呈正相关

二、 填空题

1. 下图为 2020 年 1 月份全球海平面气压场分布图。



根据上图判断，_____表示 1950-2021 年 1 月份海平面平均气压场分布；_____表示 1950-2021 年 1 月份海平面平均气压场的标准差；_____表示 2020 年 1 月份的海平面气压的距平场；_____表示 2020 年 1 月份的海平面气压的标准化距平场。



2. 已知 x, y 均是标准化数据，他们的协方差阵为 $\begin{bmatrix} 1 & 3 \\ 3 & 1 \end{bmatrix}$，若对 x 做关于 y 的回归，建立回归方程 $y = bx$，则回归系数为_____.

3. 基于一组历史观测资料(样本容量 n=100)计算得：原始变量 x 的均值为 1.5，标准差为 0.5，其标距平变量记为 x_d，y 的均值为 2.0，标准差为 4.0，标准化变量记为 y_s，利用这些历史数据拟合回归方程，$y_s = ax_d + b$，已知 a=0.5，问：

b=_____，x 与 y 的相关系数 r=_____，剩余平方和 Q=_____。若某次测得 x 的距平为x_d=-1.5，则可估计 y 的距平y_d =____。若要建立原始变量 x 与 y 的回归方程$y = cx + d$, 则 c=_____, d=_____。

4. 对标准化变量的离差乘积阵进行逐步回归操作(样本容 n=10)。在某一步，剔除了一个因子之后当前回归方程中的因子个数为 l，剔除变换后的矩阵为：

$$S^{(l)} = \begin{bmatrix} 0.101 & -0.010 & -0.818 & -0.065 & 0.641 \\ -0.010 & 0.101 & 0.125 & -0.965 & 0.694 \\ 0.818 & -0.125 & 3.357 & -1.278 & 0.127 \\ 0.065 & 0.965 & -1.278 & 0.519 & -0.0542 \\ -0.641 & -0.694 & 0.127 & -0.0542 & 0.170 \end{bmatrix}$$

(1) 当前引入的因子个数为 l=_____.

(2) 当前回归方程的回归平方和为_____，剩余平方和为_____，复相关系数为_____.

(3) 若继续引入第三个因子，剩余平方和将变为_____.

5. 对两个变量 x, y 进行 102 次观测，对 x 建立关于 y 的回归方程为 $y = 1.5x - 0.2$，其中 x 的标准差为 $s_x = 2$，y 的标准差为 $s_y = 5$，在显著性水平 $\alpha = 0.05$ 下采用 t 分布对该回归方程进行显著性检验，则用到的统计量的值为 t=_____，当_____时，可以认为回归方程通过了显著性检验.

答案

一、选择题

1. D (非线性相关等价于线性相关中的零相关)
2. A（检验两总体均值之间是否存在差异使用的统计量及服从的分布：

$$t = \frac{\overline{x}_1 - \overline{x}_2 - (\mu_1 - \mu_2)}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t(n_1 + n_2 - 2)$$

）

3. C(检验两总体的方差是否相等，采用的检验统计量为$\frac{s_1^2/s_2^2}{\sigma_1^2/\sigma_2^2}\sim F\left(n_1-1,n_2-2\right)$)

4. B(注意要将x_d转换为标准化距平才能代公式，还要注意题目给的是方差)

5. D (A. 相关系数的经过线性表示后，原来的相关系数与经过线性变换后的相关系数满足的关系为：$r_{XY}=\frac{cov\left(X,Y\right)}{\sqrt{D\left(X\right)D\left(Y\right)}},r_{aX+b,cY+d}=\frac{cov\left(aX+b,cY+d\right)}{\sqrt{D\left(aX+b\right)D\left(cY+d\right)}}=\frac{ab}{|ab|}\frac{cov\left(X,Y\right)}{\sqrt{D\left(X\right)D\left(Y\right)}}$ B. 协方差阵一定方阵，其阶数为因子的个数。C.逐步回归中，因子的每次只能引入一个，接着是进行连续剔除)

6. C(只有在初始矩阵中，最后一列（行）才表示离差乘积和，假设引入$l\left(k_1,k_2,...,k_l\right)$个因子时，当前矩阵为$\boldsymbol{S}^{(l)}$，那么回归系数即为矩阵最后一列对应元素的值：$b_{k_1}^*=S_{k_1y}^{(l)},b_{k_2}^*=S_{k_2y}^{(l)},...,b_{k_3}^*=S_{k_ly}^{(l)}$，由于这是根据标准化距平得出的资料阵，所以该回归系数为标准化变量的回归系数)

7. ABCD

8. BD(A. 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。C.偏相关系数不是简单的拿单个预报因子与预报变量做相关，例如要得到y与某因子x_i的偏相关系数，应该要剔除处x_i之外的其他因子对x_i和y的影响后，建立经过调整后的x_i',y'相关系数，才是偏相关系数)

9. AB(距平和标准化距平之间的关系为：$x_d=x_ss$，通过观察图，结果如下：

	A	B	C
距平	+1.5℃	-1℃	-1.5℃
标准化距平	+0.4	-0.4	-0.2
标准差	3.75℃	2.5℃	7.5℃

由此可知，AB 选项正确。两地气温的相关性应该通过计算相关系数来判定，所以 CD 选项错误)

二、填空题

1. B C D A（根据 Colorbar 的数值、结合统计量的性质来判断即可）

2. 3(根据公式$b = r_{xy}\frac{s_y}{s_x}$计算。因为是标准化数据，协方差阵即相关系数阵，又因为标准化数据的标准差为 1，从而回得到回归系数)

3. b=0；r=0.25；Q=1485(1600)；y_d=-3.0；c=2；d=-1

[解析]

(1) 标准化变量与距平资料的均值都为 0，因此$b = E(\hat{y}_s) - aE(x_d) = 0$

(2) 根据关系式$a s_y = r_{xy}\frac{s_y}{s_x}$, 其中$s_x = 0.5$，$a = 0.5, s_y = 4.0$，从而 x 与 y 的相关系数为 0.25；或者转化为标准化数据之间的回归方程，此时相关系数就是回归方程的系数。

(3) 剩余平方和为$Q = S_{yy}(1 - r^2) = (n-1)s_{yy}(1 - r^2) = 1485$

(4) $y_s = y_d / s_y$, $y_d = a s_y x_d = -3.0$

(5)根据已知的回归方程：$y_s = a x_d$，其中$y_s = \frac{y - \bar{y}}{s_y}, x_s = x - \bar{x}$，所以有：

$$y = a s_y x + \left(\bar{y} - a s_y \bar{x}\right)$$

对比可得，$c = 2, d = -1$

4. $l = 2$；$U = 8.83, Q = 0.17, R = 0.99$；0.165

[解析]

(1) 根据最后一行及列对应位置元素的反对性来判断，可知引入了两个因子；

(2) 矩阵的右下角的元素始终代表当前回归方程的剩余平方和，所以$Q = 0.17$；

由于是标准化数据，有 $S_{yy} = (n-1)s_y^2 = 9$，故回归平方和$U = S_{yy} - Q = 8.83$；

复相关系数为$R = \sqrt{\frac{U}{S_{yy}}} \approx 0.99$

(3) 引入一个因子等同于对矩阵做一次紧凑型求解变换，最右下角的元素值变为：

$$Q' = 0.17 - \frac{0.127^2}{3.357} \approx 0.165$$

5. 7.5 $t_{0.025}(100) < 7.5$或$t_{0.025}(100) > -7.5$ (回归系数与相关系数的关系为：$b = r_{xy}\frac{s_y}{s_x}$；回归方程的检验用到的统计量为：$t = \frac{r}{\sqrt{1-r^2}}\sqrt{n-2} \sim t(n-2)$，由题意可以求得相关系数为$r_{xy} = b\frac{s_x}{s_y} = 1.5 \times \frac{2}{5} = 0.6$，一共进行了 102 次观测，代入检验统计量中，求得$t = 7.5$)

习题二

出题人：蒋斌

涉及章节：第三章《判别分析》；第四章《聚类分析》；第五章《主成分分析》

一、选择题

1. 下列说法错误的是_____.

A. 在 Fisher 二级判别中，被预报的因子有两种判别结果

B. 在 Fisher 二级判别中，已知 A 类和 B 类的样本容量分别为n_1和n_2，它们的类内离差乘积阵为 S，则类内协方差阵的无偏估计为$\boldsymbol{V}=\dfrac{1}{n_1+n_2-1}\boldsymbol{S}$

C. 已知判别系数为$\boldsymbol{c}$，第 g 类的重心(均值)向量为$\bar{\boldsymbol{x}}_g$，则第 g 类判别函数的重心为$y=\boldsymbol{c}^{\mathrm{T}}\bar{\boldsymbol{x}}_g$

D. 根据以往经验和分析，在实验或采样时不需要查看该样品就可以得到其属于某一类的概率，这种概率称为“先验概率”

2. 下列说法错误的是_____.

A. 已知总体可分为 A 和 B 两类，它们的判别函数的重心分别为y_A,y_B，现有一样品，经过计算后得到其判别函数值为y_0，当$|y_0-y_A|<|y_0-y_B|$时，可以将该样品划入 A 类

B. 贝叶斯公式的本质为“由果溯因”，找到导致某一结果发生的各种原因的概率，从而找到导致该结果发生的最可能原因

C. 错判损失函数$\boldsymbol{L}(\boldsymbol{h}|\boldsymbol{g})$ 表示真实情况属于第$\boldsymbol{g}$类，却判成第$\boldsymbol{h}$类的损失，错判损失函数会影响由贝叶斯公式求出的后验概率

D. 某件事已经发生，想要计算这件事发生的原因是由某个因素引起的概率，这种概率称为“后验概率”或者“条件概率”

3. 下列说法错误的是_____.

A. 绝对距离和欧氏距离易受数据量纲的影响，当样品之间的量纲不同时，应该线对样品进行标准化处理

B. 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

C. 相关系数通常针对两个变量(每个变量有 n 次观测)来计算，而相似系数通常针对两次观测(一次观测有 m 个变量)来计算

D. 相似系数的取值范围为[-1, 1]，相似系数越小，表示两个样品之间的距离越近

4. 下列说法错误的是_____.

A. 定 *k-means* 聚类的方法中，第一步需要人为设定 k 个凝聚点

B. 在定 *k-means* 聚类中，如果本次聚类结果和上一次结果相同，则停止聚类

C. 在系统聚类中，聚类图可以反映两类合并时的距离

D. 在系统聚类中，经过第一次计算距离，如果采用最长距离法合并类，则应该将距离最长的两类合并

5. 下列说法错误的是_____.

A. 距平资料 $\boldsymbol{X}_{\mathrm{d}}$ 与标准化资料 $\boldsymbol{X}^*$ 的主成分相同

B. 各个主成分的方差就是观测资料阵的协方差阵 $\boldsymbol{S}$ 对应的特征值， 不同主成分是相互正交的，即不同主成分的协方差均为 0

C. m 个原变量进行了 n 次观测，主成分最多有 r 个，其中 $r \leq min\{m,n\}$

D. 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

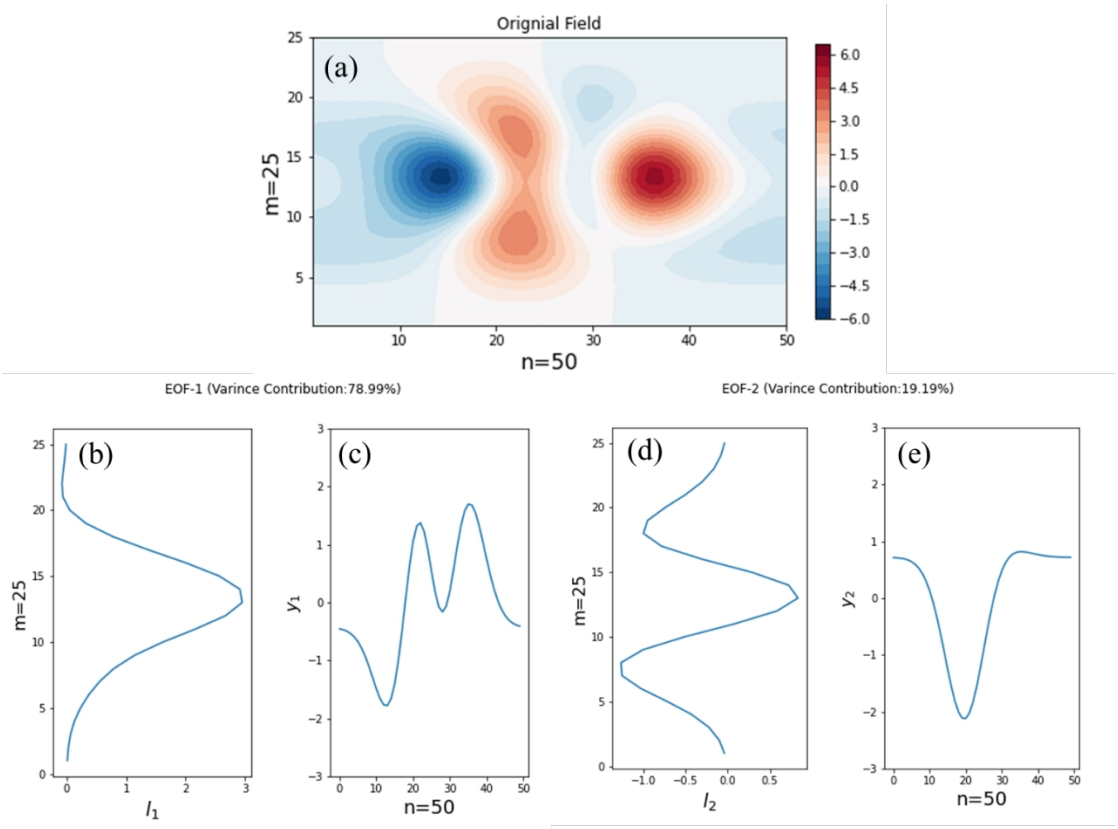
6. [多选] 图(a)为某距平资料的原始场，m=25 表示原变量个数，n=50 为观测次数， 图(c)和图(e)分别该场对应的第一主成分(PC1)和第二主成分(PC2)，图(b)和图(d)分别为 PC1 和 PC2 对应的时间系数，下列说法正确的是_____.

A. 图(b)表明第一模态中原变量 10~15 对应的估计值存在极大或者极小

B. 图(c)反映了第一模态有三个大值中心，且其中两个与另一个反位相

C. 图(d)反映了原变量观测值的变化规律为“小-大-小”或“大-小-大”

D. 图(e)表明，大约在第 20 次观测中 m 个原变量的变化特征与图(d)反映的特征相反



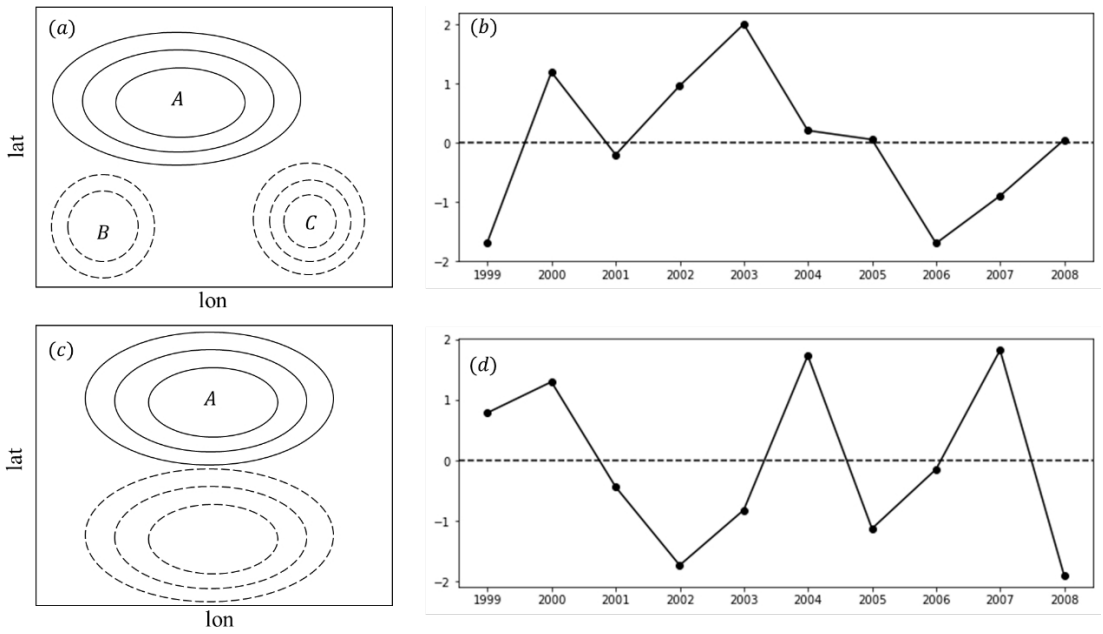
7. [多选]现对某区域 1999-2008 年 12 月份的气温距平场进行 EOF 分析。图(a)和图(b)分别为气温距平场对应的第一模态和第一主成分，图(c)和图(d)分别为气温距平场对应的第二模态和第二主成分。在图(a)和(c)中，实等值线表示正值，虚等值线表示负值，下列说法正确的是_____.

A. C 地的气温比 B 地的气温更冷

B. A 地 2006 年经历了比往年更冷的一个冬天

C. 2004 年，该区域的气温异常为“南暖背冷”

D. 以上说法均不正确



二、填空题

1. 采用 m 个因子对某地有雨或无雨进行判别分析，总共有 n 次历史观测，记为 n 行 m 列资料阵 $\boldsymbol{X}$，均值向量(m 行 1 列)记为$\bar{\boldsymbol{x}}$，距平资料阵记为 $\boldsymbol{X}_{\mathrm{d}}$，其中 n_1 次有雨，n_2 次无雨，即 $n_1+n_2=n$。有雨时的因子资料阵(n_1 行 m 列)记为 $\boldsymbol{P}$，其均值向量记为(m 行 1 列) $\bar{\boldsymbol{p}}$，距平资料阵记为 $\boldsymbol{P}_{\mathrm{d}}$, 无雨时的因子资料阵($n_2$ 行 m 列)记为 $\boldsymbol{Q}$，其均值向量记为(m 行 1 列) $\bar{\boldsymbol{q}}$ ，距平资料阵记为 $\boldsymbol{Q}_{\mathrm{d}}$。

(1) 总(合并)类内离差乘积阵可表示为 _____；

(2) Fisher 二级判别系数向量可表示为 $\boldsymbol{c}$=_____；

(3) 判别函数的组间距离的平方可表示为_____；

(4) 判别函数的组内离差平方和可表示为_____；

(5) 判别函数的组间离差平方和可表示为_____；

(6) 对于某次新的因子观测记为(列向量)$\boldsymbol{x}_0$，如果_____，则可预测有雨，否则预测无雨。

2. 已知 5 个原变量共有 3 个主成分，已知主成分的协方差阵为 $\begin{bmatrix}2&0&0\\0&3&0\\0&0&5\end{bmatrix}$，则第二主成分的方差贡献为_____，原变量的总方差为_____.

3. 采用最长距离法进行合并。下图为 6 个样品通过系统聚类得出的距离矩阵，

则___和___合为一类。

$D_{(0)}$	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	\

4. m 个变量共有 r $(r \le m)$个主成分，其系数向量矩阵为 $\boldsymbol{L}=[\boldsymbol{l}_1,\boldsymbol{l}_2,...,\boldsymbol{l}_r]$，现在观测得到一组样品 $\boldsymbol{x}=[x_1, x_2, ..., x_m]^{\mathrm{T}}$，那么第一主成分可以表示为_____.

5. 对 m 个变量一共进行了 n 次观测，得到了资料阵 $\boldsymbol{X}_{(m\times n)}$，且共有 r 个主成分 $(r \le \min\{m,n\})$，数向量矩阵为 $\boldsymbol{L}=[\boldsymbol{l}_1,\boldsymbol{l}_2,...,\boldsymbol{l}_r]$，则主成分可以表示为_____，其中第 k 主成分的观测序列 $\boldsymbol{y}_k^{\mathrm{T}}$ =_____.(第 4 题，第 5 题的主成分系数向量均为列向量)

三、计算题

1. 有以下 6 个样品（每个样品有 2 个指标），标号依次为 1-6

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

(1) 以相似系数度量样品距离，写出距离阵并找出距离最近的一组样品。

(2) 以绝对距离度量样品距离，按最长距离法分类，并画出聚类图。

2. 已知任意一个样品可分为三类：G_1, G_2, G_3，某样品属于这三类的先验概率分别为 $p_1=0.35, p_2=0.48, p_3=0.17$．现抽得一个样品 $\boldsymbol{x}_0$，它在三类中的概率密度分别为 $f_1(\boldsymbol{x}_0)=0.65, f_2(\boldsymbol{x}_0)=0.83, f_3(\boldsymbol{x}_0)=0.37$，考虑如下错判损失：

$L(G_1 \mid G_1)=0$	$L(G_2 \mid G_1)=4$	$L(G_3 \mid G_1)=3$
$L(G_1 \mid G_2)=5$	$L(G_2 \mid G_2)=0$	$L(G_3 \mid G_2)=7$
$L(G_1 \mid G_3)=2$	$L(G_2 \mid G_3)=10$	$L(G_3 \mid G_3)=0$

例如：$L(G_2 \mid G_1)$表示真实类别为G_1却判成了G_2

判断$\boldsymbol{x}_0$属于哪一类的概率最大。

3. 利用计算机(Matlab 或 Python)对气象资料进行主成分分析或者 EOF 分解时，常常会涉及“时空转换”。若已知某距平资料阵为$\boldsymbol{X}_{(m\times n)}$，试推导“时空转换”的原理。

答案

一、选择题

1. B(求解 Fisher 判别的正规方程组中的合并类内离差乘积矩阵$\boldsymbol{S}$的协方差阵的无偏估计为：$\boldsymbol{V}=\dfrac{1}{n_1+n_2-2}\boldsymbol{S}$)
2. C(错判损失函数考虑了因为错判所造成的损失，从而影响我们在判别时的决策，它不会因影响由贝叶斯公式求出的后验概率)
3. D(相似系数越大，表示两个样品之间的距离越近)
4. D(注意题中的表述——“第一次计算距离”,那么此时每一类就只有一个样品，此时计算的距离还是样品间距，对于样品之间的距离，一定是“距离越短越容易合并成一类”)
5. A(主成分分析的目的就是要挑出所用因子中方差较大的几个因子作为第一主成分表达，距平资料当然可以满足上述要求，但是对于标准化资料，各变量方差相等，将反应大多数变量的信息)
6. ABCD
7. BC(第一模态指的是系数向量$\boldsymbol{l}_1$的空间分布,第一主成分指的是$\boldsymbol{y}_1$的时间序列

分布，第二模态、第二主成分同理。于是第一、二模态气温距平的估计值分别为 $\boldsymbol{x}_{d1}=\boldsymbol{l}_1^T\boldsymbol{y}_1, \boldsymbol{x}_{d2}=\boldsymbol{l}_2^T\boldsymbol{y}_2$，据此去判断符号及其实际意义。)

二、填空题

1. (1) $\boldsymbol{P}_d^T\boldsymbol{P}_d+\boldsymbol{Q}_d^T\boldsymbol{Q}_d$；(2) $\left(\boldsymbol{P}_d^T\boldsymbol{P}_d+\boldsymbol{Q}_d^T\boldsymbol{Q}_d\right)^{-1}(\overline{\boldsymbol{p}}-\overline{\boldsymbol{q}})$；(3) $({}_C{}^T\overline{\boldsymbol{p}}-{}_C{}^T\overline{\boldsymbol{q}})^2$ ；(4) $\boldsymbol{c}^T(\boldsymbol{P}_d^T\boldsymbol{P}_d+\boldsymbol{Q}_d^T\boldsymbol{Q}_d)\boldsymbol{c}$(或者将此式展开写也对)；(5) $n_1\left(\boldsymbol{c}^T\overline{\boldsymbol{p}}-\boldsymbol{c}^T\overline{\boldsymbol{x}}\right)^2+n_2\left(\boldsymbol{c}^T\overline{\boldsymbol{q}}-\boldsymbol{c}^T\overline{\boldsymbol{x}}\right)^2$；(6) $|{}_C{}^T\boldsymbol{x}_0-{}_C{}^T\overline{\boldsymbol{p}}|<|{}_C{}^T\boldsymbol{x}_0-{}_C{}^T\overline{\boldsymbol{q}}|$(注意本题的资料阵摆放形式)
2. 30%；10(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 S 的特征值，所以，r 个主成分 $y=[y_1, y_2, \ldots, y_r]^T$ 的协方差阵为对角阵)
3. G4 G6 (对于样品之间的距离，原则总是“距离越短进行合并”，最长距离、最短距离是针对**类间距离**而言)
4. $y_1=\boldsymbol{l}_1^T\boldsymbol{x}$
5. $\boldsymbol{Y}=\boldsymbol{L}^T\boldsymbol{X}$； $\boldsymbol{y}_k^T=\left(\boldsymbol{l}_k^T\boldsymbol{X}\right)^T$

三、计算题

1.(1) 采用 $1-\cos\theta_{ij}$ 作为距离度量，满足值越小表示样品距离越近

$D_{(0)}$	G1	G2	G3	G4	G5	G6
G1	\					
G2	1.71	\				
G3	0.55	1.95	\			
G4	0.68	1.89	0.01	\		
G5	1.71	1.00	0.68	0.55	\	
G6	1.00	1.71	0.11	**0.05**	0.29	\

(2) 系统聚类过程如下：

- **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	\

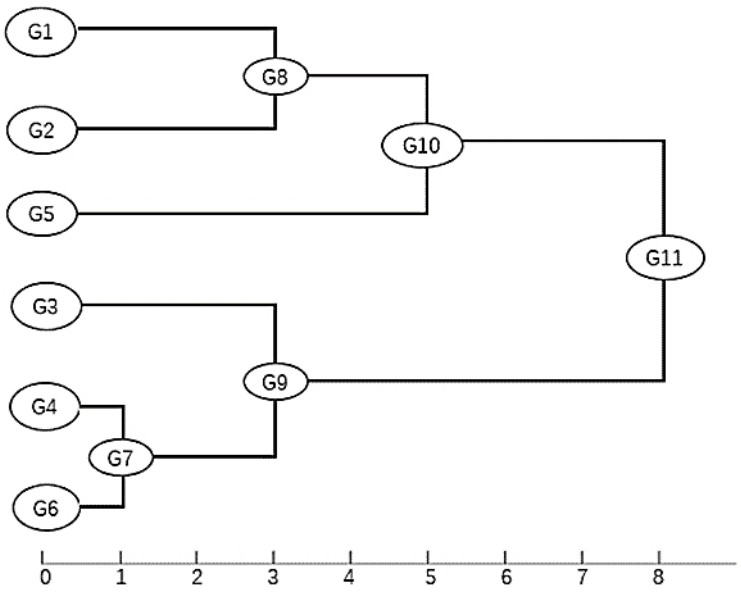
- **Step 4：G1 与 G2 合并为 G8， G3 与 G7 合并为 G9；**

$D_{(2)}$	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)

2. 提示：考察贝叶斯公式和错判损失，计算 $\min\left\{p_i f_i L\left(g\mid h\right)\right\}\left(i=1,2,3\right)$

3. 解答过程如下：

- 协方差阵$\boldsymbol{S}=(\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}})/n$ (设$\boldsymbol{X}$已是距平资料阵)的阶数为m，　如果m非常大，求解大矩阵对内存容量有较高要求，不方便.

◆ 先求出$\boldsymbol{S}_{\mathrm{Q}}=(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})/m$的某特征值为$\lambda_Q$，相应的单位化的特征向量记为$\boldsymbol{l}_{\mathrm{Q}}$

$$(\frac{1}{m}\boldsymbol{X}^T\boldsymbol{X})\boldsymbol{l}_Q=\lambda_Q\boldsymbol{l}_Q \quad \text{其中，}\boldsymbol{l}_Q\text{已单位化：}\boldsymbol{l}_Q^{\mathrm{T}}\boldsymbol{l}_Q=1$$

◆ 将上式左乘 $\frac{1}{n}\boldsymbol{X}\longrightarrow(\frac{1}{n}\boldsymbol{X}\boldsymbol{X}^T)\boldsymbol{X}\boldsymbol{l}_Q=\frac{m}{n}\lambda_Q\boldsymbol{X}\boldsymbol{l}_Q$

因此，$\boldsymbol{S}$的特征值$\lambda=\frac{m}{n}\lambda_Q$,对应的特征向量为$\boldsymbol{X}\boldsymbol{l}_{\boldsymbol{Q}}$

$\boldsymbol{X}\boldsymbol{l}_Q$模的平方等于：$(\boldsymbol{X}\boldsymbol{l}_Q)^{\mathrm{T}}(\boldsymbol{X}\boldsymbol{l}_Q)=\boldsymbol{l}_Q^T\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{l}_Q=m\lambda_Q\boldsymbol{l}_Q^T\boldsymbol{l}_Q=m\lambda_Q$

所以，$\boldsymbol{S}$的单位化的特征向量$\boldsymbol{l}=\frac{\boldsymbol{X}\boldsymbol{l}_Q}{\sqrt{m\lambda_Q}}$

习题三

出题人：蒋斌

涉及章节：第八章《时间序列分析》；第九章《波谱分析》

一、选择题

1. 关于随机过程，下列说法正确的是_____.

A. 已知两个随机过程的均值函数和方差函数都完全一样，则两个随机过程的特点完全一样

B. 平稳随机过程的均值函数与时间无关，其自协方差函数仅与起止时间有关

C. 平稳随机过程具有“各态历经性”是指：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

D. 白噪声过程是一种非平稳随机过程，因为其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

2. 下列有关说法错误的是_____.

A. 对于平稳随机序列 $\boldsymbol{x}$，其时滞为 τ 的自协方差函数可以写为

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

B. 符合一阶自回归模型的平稳随机过程称为红噪声过程

C. p 阶自回归模型需要利用前期 p (t-1, t-2, …, t-p)个时刻的值作为因子变量

D. p 阶自回归方程的检验与多元回归方程的检验用到的统计量均为

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

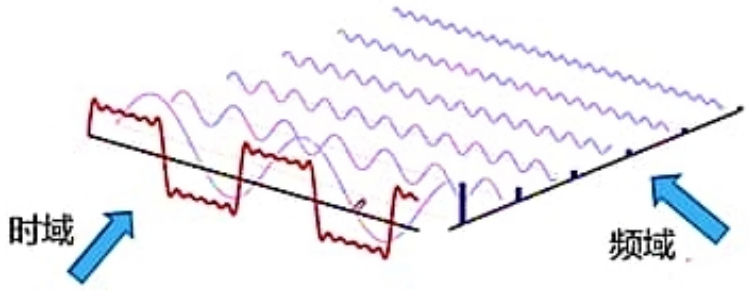
3. [多选]根据下图，关于傅里叶变换与逆变换理解，正确的有_____.

A. 在时域方向看，只能看到一个随时间变化的时域信号

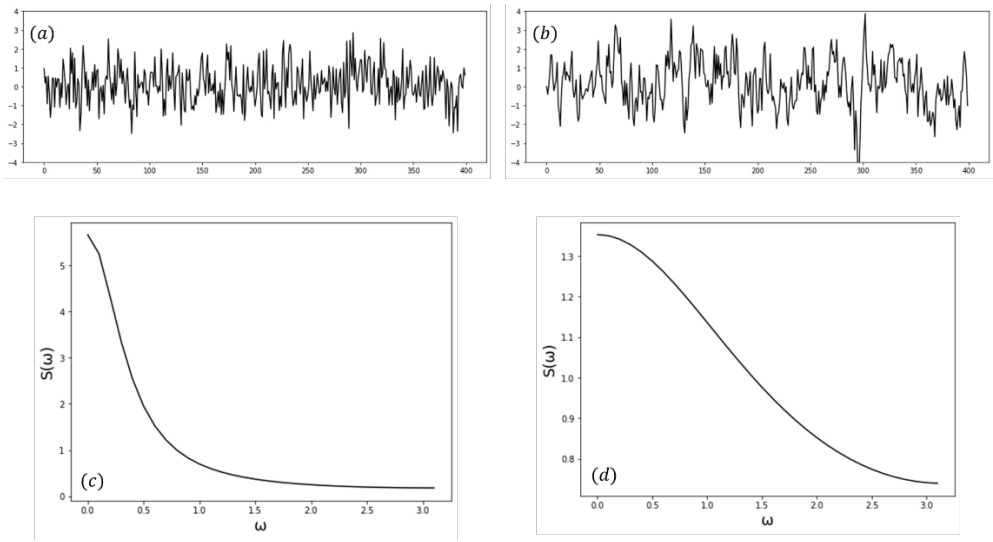
B. 从频域方向看，可以看出这个信号究竟包含哪些频率分量

C. 从频域方向看，可以看到每一个频率分量的幅值是多少

D. 从频域方向看，可以确定每一个频率分量的初始相位



4. [多选]已知两个一阶自回归方程分别为：$x_t = 0.15x_{t-1} + a_t, y_t = 0.7y_{t-1} + b_t$，其中$a_t, b_t$均表示白噪声。图(a)和图(b)表示不同一阶自回归方程的时间序列图，图(c)和图(d)表示不同一阶自回归方程的功率谱，下列说法正确的是_____.



A. 图(*a*)表示 *x* 的时间序列，图(*b*) 表示 *y* 的时间序列

B. 图(*a*)表示 *y* 的时间序列，图(*b*) 表示 *x* 的时间序列

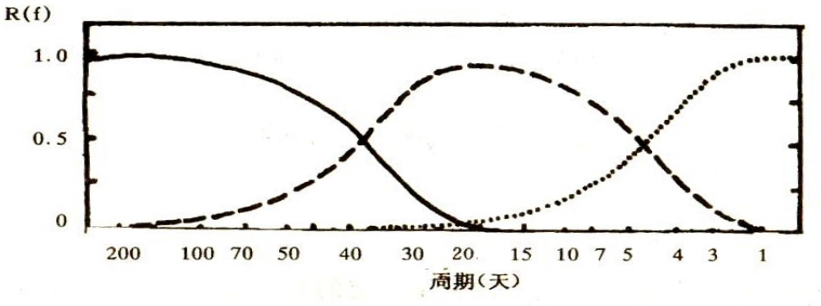
C. 图(*c*)表示 *x* 的功率谱，图(*d*) 表示 *y* 的功率谱

D. 图(*c*)表示 *y* 的功率谱，图(*d*) 表示 *x* 的功率谱

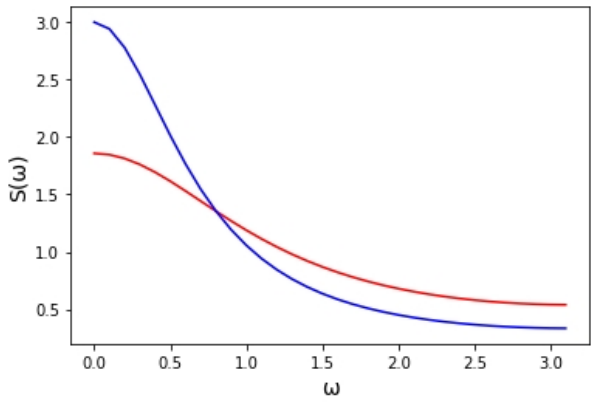
二、填空题

1. 平稳时间序列$a(1), a(2), ..., a(N)$和$b(1), b(2), ..., b(N)$。已知 *a* 序列的方差为 *A*，*b* 序列的方差为 *B*，两时间序列的均值均为 0。则 *b* 超前 *a* 2 个时间单位的自相关函数的表达式为_____.
2. 傅里叶变换的表达式为_____.(离散与连续的形式均需要写出)

3. 奈奎斯特采样定理表明，当采样频率应与信号频率满足_____时，才能获得准确的信号周期，否则会产生混叠效应(虚假周期)。
4. 维纳辛勤定理表明，对于平稳随机过程，_____和_____是一对傅氏变换对。
5. 下图的实线(“—”)表示_____滤波(选填“低通”“高通”“带通”)。



6. 下图分别表示两种不同红噪声过程的功率谱图。则红线对应的时滞为 1 的自相关系数与蓝线相比，哪一个更大？_____.



7. 已知某滤波方式的响应函数为$H(f)=\frac{\sin \pi fm}{m\sin \pi f}$，则该滤波方式为_____；某滤波方式的响应函数为$H(f)=\cos^m(\pi f)$，则该滤波方式为_____；某滤波方式的响应函数为$\left|H\left(f\right)\right|=\left(2\left|\sin \pi f\right|\right)^q$，则该滤波方式为_____。(选填“低通滤波”“高通滤波”“带通滤波”)
8. 五点二项式系数滑动平均的权重系数为_____。

三、推导证明题

1. 对于一阶自回归模型 $x_t=\beta x_{t-1}+a_t$，其中，a 为白噪声。时滞 (落后间隔)为 1 个采样间隔的自相关记为 ρ_1，时滞为τ个采样间隔的自相关记为ρ_τ。证明：$\hat{\beta}=\rho_1$，且 $\rho_\tau=\rho_1^\tau$
2. 已知某时间序列为 $f\left(t\right)$，其频域函数为 $F\left(\omega\right)$，采用一阶差分法对该时间序

列进行高通滤波，得到滤波后的时间序列为$g(t)$，频域函数为$G(\omega)$，该过程的频响函数为$H(\omega)$，试推导$H(\omega)$的表达式。

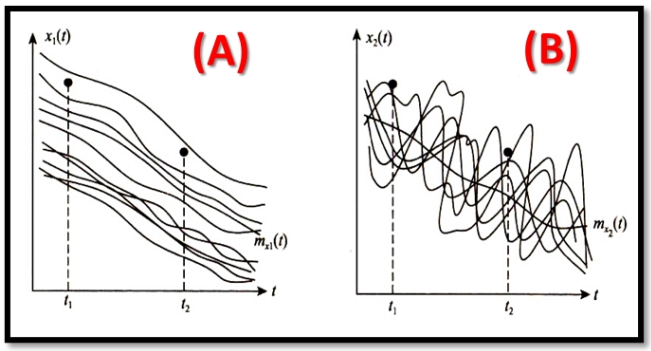
四、计算题

已知三阶自回归方程：$X_t = 0.88X_{t-1} + 0.37X_{t-2} - 0.56X_{t-3} + a_t$，试建立其二阶自回归方程。

答案

一、选择题

1. C(A 项，即使两个随机过程的均值函数和方差函数都完全一样，那么他们仍可能具有完全不同的特点，如下图，(A)与(B)的均值函数和方差函数都相同，但是(A)在 t_1 和 t_2 时刻具有相关性，而(B)在 t_1 和 t_2 时刻不具相关性。B 项，平稳随机过程的协方差函数只与时间间隔有关，而与时间的起止无关。D.白噪声过程属于平稳随机过程的一种)



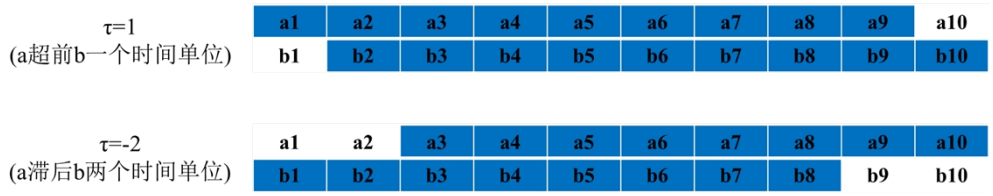
2. A(正确的表达式为：$\hat{K}(\tau) = \frac{1}{n-\tau}\sum_{t=1}^{n-\tau}(x_t - \bar{x})(x_{t+\tau} - \bar{x})$或

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

3. ABCD
4. AD

二、填空题

1. 1. $\frac{\sum_{t=1}^{N-2} b(t)a(t+2)}{(N-2)\sqrt{AB}}$(关于计算两个时间序列的超前滞后的协方差如下图所示，填色部分为计算使用的序列)



2. 离散：$C_k=\int_{-\frac{T}{2}}^{\frac{T}{2}} f(t)e^{-i\omega t}dt$；连续：$F(\omega)=\int_{-\infty}^{\infty} f(t)e^{-i\omega t}dt$
3. 采样频率≥信号频率的 2 倍
4. 功率谱密度　自相关函数
5. 低通
6. 蓝线 (红噪声功率谱的表达式为$S(\omega)=\frac{1-\rho_1^2}{1\text{-}2\rho_1\cos\omega+\rho_1^2}$，取$\omega=0$，则红噪声功率谱可写为$S(\omega)=\frac{1-\rho_1^2}{(1-\rho_1)^2}=\frac{1+\rho_1}{1-\rho_1}$，可见时滞为 1 的自相关系数越大，对应的功率谱值越大)
7. 低通滤波　低通滤波 高通滤波
8. 1/16, 4/16, 6/16, 4/16, 1/16

三、推导证明题

1. 证明过程如下：

对于白噪声过程，其(协)方差满足关系 $K_a(t,t+\tau)=\begin{cases}\sigma_a^2 & (\tau=0)\\ 0 & (\tau\neq 0)\end{cases}$

对于一阶自回归模型 $x_t=\beta x_{t-1}+a_t$

用前一时刻的x_{t-1}乘以上式两边，然后取数学期望，得

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

一般对数据进行中心化处理(求距平)，有$E(x_t)=0$，而$E(x_{t-1}a_t)$，故有 $\hat{\beta}=\frac{E(x_t x_{t-1})}{E(x_{t-1}x_{t-1})}=\rho_1$

一阶自回归模型实际上又是一个递推的公式，对于时滞为τ的情况，有回归模型

$$x_t=\rho_1^{\tau}x_{t-\tau}+A\left(A=\sum_{k=0}^{\tau-1}\rho_1^k a_{t-k}\right)$$

用前τ时刻的$x_{t-\tau}$乘以上式两边，然后取数学期望，同一阶自回归模型推导过程，得

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

2. 由题意可得关系式：$G(\omega)=H(\omega)F(\omega)$，而

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

类比可得，$H(\omega)=1-e^{-i\omega}$，它的模为：

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

四、计算题

解：三阶自回归模型的尤拉-沃克方程组为：

$$\begin{bmatrix}1&\rho_1&\rho_2\\\rho_1&1&\rho_1\\\rho_2&\rho_1&1\end{bmatrix}\begin{bmatrix}\varphi_1\\\varphi_2\\\varphi_3\end{bmatrix}=\begin{bmatrix}\rho_1\\\rho_2\\\rho_3\end{bmatrix}$$

已知$\varphi_1=0.98,\varphi_2=1.01,\varphi_3=-0.56$，于是有方程组：

$$\begin{cases}0.88+0.37\rho_1-0.56\rho_2=\rho_1\\0.88\rho_1+0.37-0.56\rho_1=\rho_2\\0.88\rho_2+0.37\rho_1-0.56=\rho_3\end{cases}$$

根据第一式和第二式$\begin{cases}0.63\rho_1+0.56\rho_2=0.88\\0.32\rho_1-\rho_2=-0.37\end{cases}$，可以解得$\rho_1=0.83,\rho_2=0.64$

建立二阶尤拉-沃克方程组为：$\begin{bmatrix}1&\rho_1\\\rho_1&1\end{bmatrix}\begin{bmatrix}\varphi_1'\\\varphi_2'\end{bmatrix}=\begin{bmatrix}\rho_1\\\rho_2\end{bmatrix}$

解得$\varphi_1'=0.96,\varphi_2'=-0.16$，所以二阶自回归方程为$X_t=0.96X_{t-1}-0.16X_{t-2}$