DS数模 大神带你来建模

2021年国赛B题思路

国赛B题涉及到化学的知识较多，因此比较适合化学类专业的同学做，同时其中所涉及到的一些统计化学相关的知识，所以在这里，我们将为大家讲解所使用的相关方法与知识点

第一问：

第一问主要需要我们分析附件中每种催化剂之间的组合，来研究乙醇转化率、C4烯烃的选择性与温度之间的关系

学过化学的同学都知道，温度越高，乙醇挥发越快，在羟基磷灰石的催化作用下，能够将乙醇转化为乙烯、C4烯烃、乙醛、碳数为 4-12 脂肪醇，但是，做题的时候我们不能直接使用这个结论，需要我们利用统计学相关的知识。

第一问需要用到Pearson相关系数

如果两组数据 $X:\{X_1,X_2,\cdots,X_n\}$ 和 $Y:\{Y_1,Y_2,\cdots,Y_n\}$ 是总体数据（例如普查结果），

那么总体均值：$E(X)=\frac{\sum_{i=1}^{n}X_i}{n}$，$E(Y)=\frac{\sum_{i=1}^{n}Y_i}{n}$

总体协方差：$\mathrm{Cov}(X,Y)=\frac{\sum_{i=1}^{n}(X_i-E(X))(Y_i-E(Y))}{n}$

直观理解协方差： 如果X、Y变化方向相同，即当X大于（小于）其均值时，Y也大于（小于）其均值，在这两种情况下，乘积为正。如果X、Y的变化方向一直保持相同，则协方差为正；同理，如果X、Y变化方向一直相反，则协方差为负；如果X、Y变化方向之间相互无规律，即分子中有的项为正，有的项为负，那么累加后正负抵消。

注意： 协方差的大小和两个变量的量纲有关，因此不适合做比较。

回顾《概率论与数理统计》中的数理统计部分：

如果两组数据 $X:\{X_1,X_2,\cdots,X_n\}$ 和 $Y:\{Y_1,Y_2,\cdots,Y_n\}$ 是总体数据（例如普查结果），

那么总体均值：$E(X)=\frac{\sum_{i=1}^{n}X_i}{n}$，$E(Y)=\frac{\sum_{i=1}^{n}Y_i}{n}$

总体协方差：$\mathrm{Cov}(X,Y)=\frac{\sum_{i=1}^{n}(X_i-E(X))(Y_i-E(Y))}{n}$

总体Pearson相关系数：$\rho_{XY}=\frac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{\sum_{i=1}^{n}\frac{(X_i-E(X))}{\sigma_X}\frac{(Y_i-E(Y))}{\sigma_Y}}{n}$

$\sigma_X(sigma\ X)$ 是 X 的标准差，$\sigma_X=\sqrt{\frac{\sum_{i=1}^{n}(X_i-E(X))^2}{n}}$，$\sigma_Y=\sqrt{\frac{\sum_{i=1}^{n}(Y_i-E(Y))^2}{n}}$

可以证明，$|\rho_{XY}|\le 1$，且当 $Y=aX+b$ 时，$\rho_{XY}=\begin{cases}1, & a>0\\ -1, & a<0\end{cases}$

Pearson相关系数能够分析多个变量之间而定影响，因此非常适合于本题的相关性分析。

然后需要我们对附件二进行分析，需要我们重复五次上述操作，最后得到结果的变量只是与时间有关

350度时给定的某种催化剂组合的测试数据

时间（min）	乙醇转化率(%)	选择性(%) 乙烯选择性	C4烯烃选择性	乙醛选择性	碳数为4-12脂肪醇	甲基… 甲基…
20	43.5	4.23	39.9	5.17	39.7	2
70	37.8	4.28	38.55	5.6	37.36	4
110	36.6	4.46	36.72	6.37	32.39	4
163	32.7	4.63	39.53	7.82	31.29	4
197	31.7	4.62	38.96	8.19	31.49	4
240	29.9	4.76	40.32	8.42	32.36	4
273	29.9	4.68	39.04	8.79	30.86	3

大体上的思路就是乙醇在催化剂的作用下经过了多少时间的，转化率发生了多少变化。

最简单的分析就可以写：在350度的环境下，随着时间的变化乙醇转化率下降，乙烯选择性较为稳定，C4烯烃的选择性较为稳定，乙醛的转化率随时间变化不断升高，脂肪醇的选择性随时间变化下降。

第二问：

不同催化剂对于C4烯烃选择性的影响，需要从附件1中寻找数据，不同温度对于C4烯烃选择性的影响需要从附件2中寻找数据。　　因此在这里，需要我们使用统计学上相关的典型相关分析，典型相关分析是用来分析两个变量之间的相关性关系的方法，

典型相关分析由Hotelling提出，其基本思想和主成分分析非常相似。

首先在每组变量中找出变量的线性组合，使得两组的线性组合之间具有最大的相关系数；

然后选取和最初挑选的这对线性组合不相关的线性组合，使其配对，并选取相关系数最大的一对；

如此继续下去，直到两组变量之间的相关性被提取完毕为止。

被选出的线性组合配对称为**典型变量**，它们的相关系数称为**典型相关系数**。典型相关系数度量了这两组变量之间联系的强度。

典型相关分析的建模思路：

假设两组变量分别为：$X^{(1)}=(X_1^{(1)},X_2^{(1)},\cdots,X_p^{(1)}),X^{(2)}=(X_1^{(2)},X_2^{(2)},\cdots,X_q^{(2)})$

分别在两组变量中选取若干有代表性的综合变量 U_i、V_i，

使得每一个综合变量是原变量的线性组合，即

$$U_i=a_1^{(i)}X_1^{(1)}+a_2^{(i)}X_2^{(1)}+\cdots+a_p^{(i)}X_p^{(1)}\triangleq \mathbf{a}^{(i)\prime}\mathbf{X}^{(1)}$$

$$V_i=b_1^{(i)}X_1^{(2)}+b_2^{(i)}X_2^{(2)}+\cdots+b_q^{(i)}X_q^{(2)}\triangleq \mathbf{b}^{(i)\prime}\mathbf{X}^{(2)}$$

注意：综合变量的组数是不确定的，如果第一组就能代表原样本数据大部分的信息，那么一组就足够了。假设第一组反映的信息不够，我们就需要找第二组了。并且为了让第二组的信息更有效，需要保证两组的信息不相关。

不相关：$cov(U_1,U_2)=cov(V_1,V_2)=0$

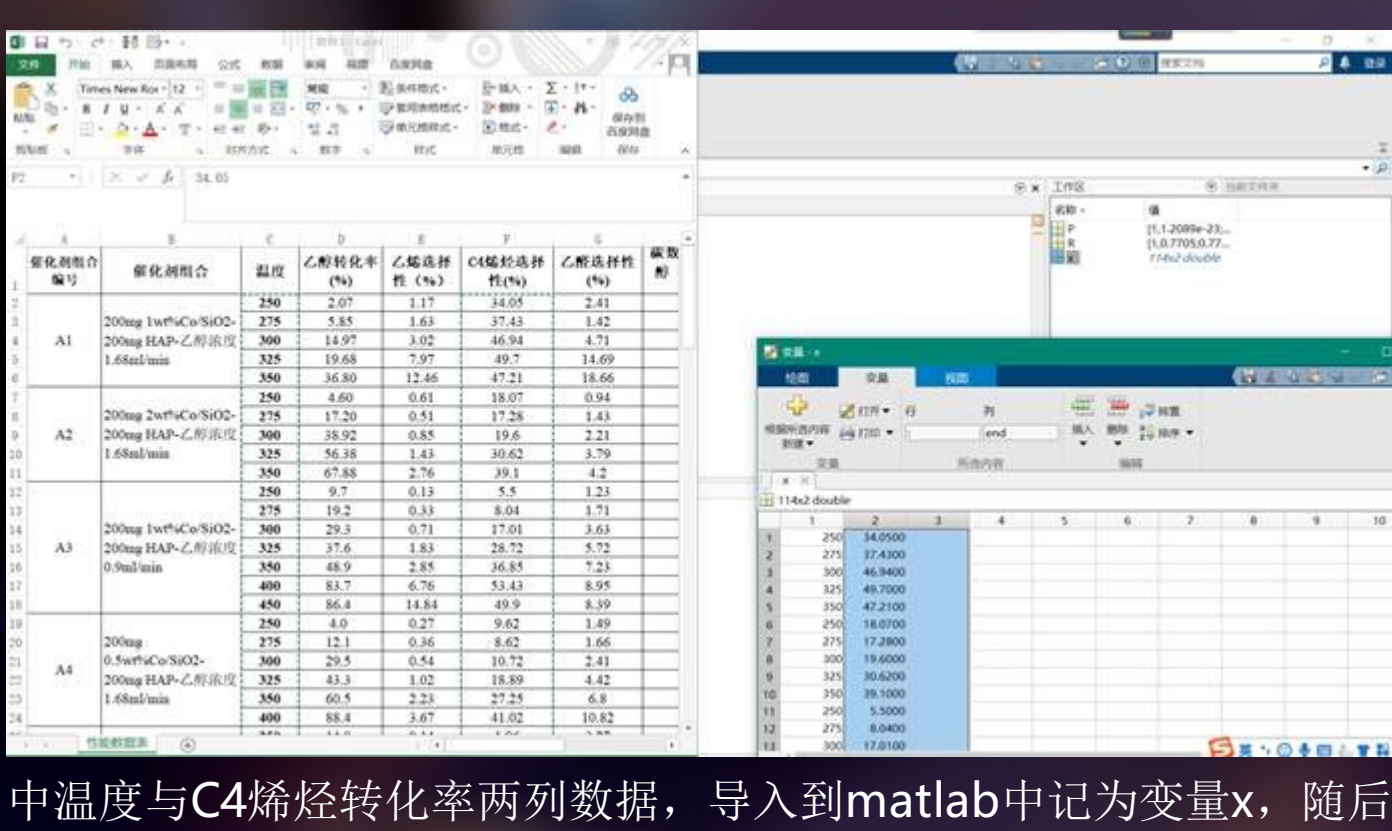
第一组要满足的条件：

在 $var(U_1)=var(V_1)=1$ 满足的条件下，找到 $\mathbf{a}^{(1)}$ 和 $\mathbf{b}^{(1)}$ 两组系数，使得 $\rho(U_1,V_1)$ 最大

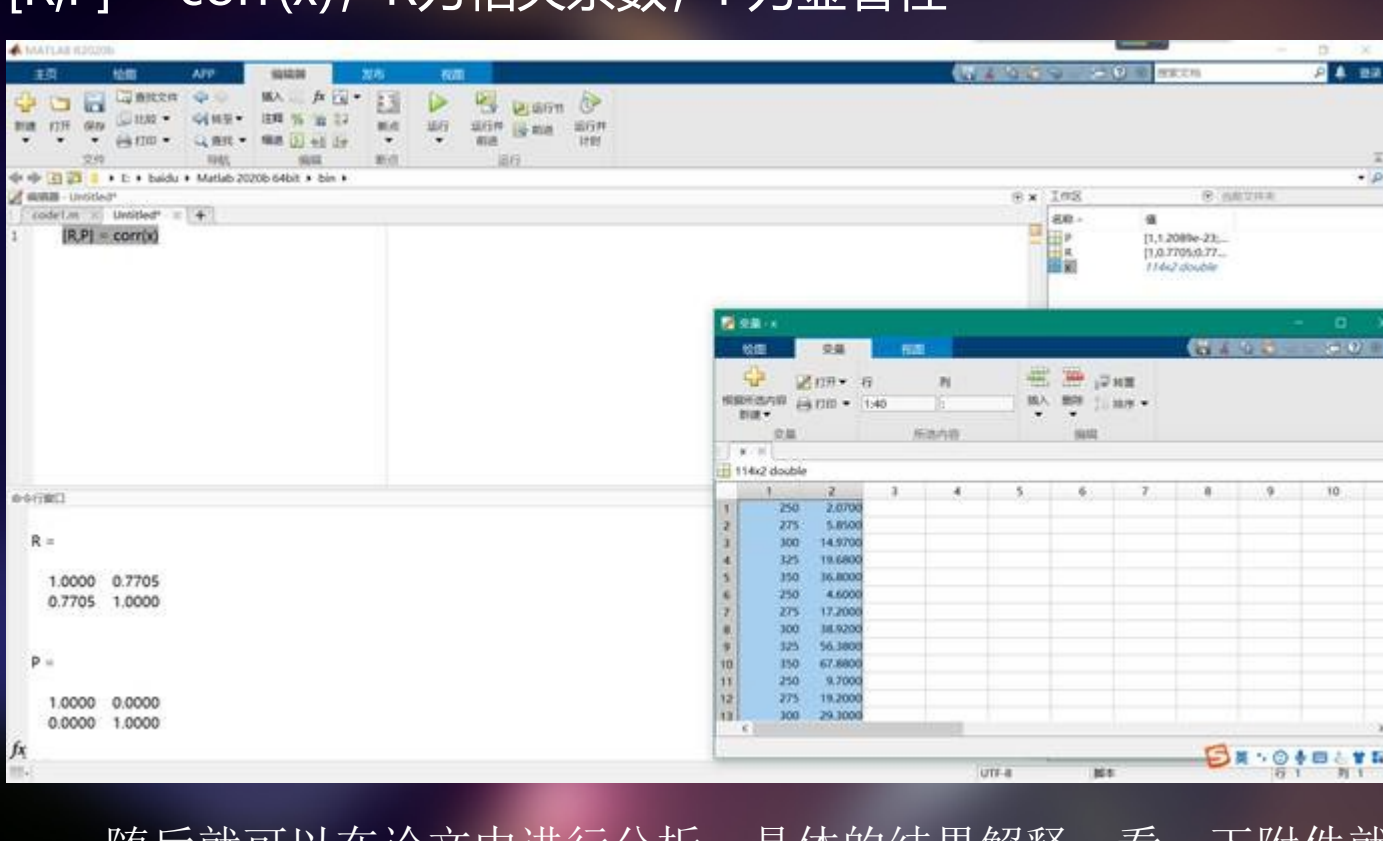
（为什么要固定这个条件：因为相关系数与量纲无关：$\rho(U_1,V_1)=\rho(aU_1,bV_1)$）

如果大家想要深入了解典型相关分析，我整理了厦门大学的典型相关分析课件，免费赠送给大家

接下来我为大家在Matlab中进行一下实操：



首先选中温度与C4烯烃转化率两列数据，导入到matlab中记为变量x，随后输入代码[R,P] = corr(x)，R为相关系数，P为显著性



随后就可以在论文中进行分析，具体的结果解释，看一下附件就都会明白了。

第三问：

第三问是要求我们计算催化剂与温度，对于C4烯烃选择性的影响，在做这一问题时，如果观察原始数据图，我们会发现基本上是看不出来什么规律的

催化剂组合编号	催化剂组合	温度	乙醇转化率(%)	乙烯选择性（%）	C4烯烃选择性(%)	乙醛选择性(%)
A1	200mg 1wt%Co/SiO2-200mg HAP-乙醇浓度1.68ml/min	250	2.07	1.17	34.05	2.41
		275	5.85	1.63	37.43	1.42
		300	14.97	3.02	46.94	4.71
		325	19.68	7.97	49.7	14.69
		350	36.80	12.46	47.21	18.66
A2	200mg 2wt%Co/SiO2-200mg HAP-乙醇浓度1.68ml/min	250	4.60	0.61	18.07	0.94
		275	17.20	0.51	17.28	1.43
		300	38.92	0.85	19.6	2.21
		325	56.38	1.43	30.62	3.79
		350	67.88	2.76	39.1	4.2
A3	200mg 1wt%Co/SiO2-200mg HAP-乙醇浓度0.9ml/min	250	9.7	0.13	5.5	1.23
		275	19.2	0.33	8.04	1.71
		300	29.3	0.71	17.01	3.63
		325	37.6	1.83	28.72	5.72
		350	48.9	2.85	36.85	7.23
		400	83.7	6.76	53.43	8.95
		450	86.4	14.84	49.9	8.39
		250	4.0	0.27	9.62	1.49

所以我们需要转变一下思路，将数据重新整理一下来看：

温度	A1	A2	A3	A4	A5	A6	A7	A8	A9	A10
250	34.05	18.07	5.5	9.62	1.96	3.3	5.75	5.63	5.4	2.19
275	37.43	17.28	8.04	8.62	6.65	7.1	6.56	8.52	9.68	1.65
300	46.94	19.6	17.01	10.72	10.12	7.18	8.84	13.82	16.1	2.17
325	49.7	30.62	28.72	18.89	13.86	10.65	18.64	25.89	31.04	3.3
350	47.21	39.1	36.85	27.25	18.75	37.33	33.25	41.42	42.04	10.29
400			53.43	41.02	38.23					
450			49.9							
	2.632	4.206	6.27	3.526	3.358	6.806	5.5	7.158	7.328	1.62

我们将表格整理成为如下的形式来进行分析：我们可以通过数值的增长率来分析其中所蕴含的规律，最后一列是我们计算出来的数值增长率。

1、通过观察我们发现A1与A2的变化是CO/Si的浓度，观察数据，我们发现，CO/Si的浓度可以提高反应的速度，因为A2的增长率是4而A1的增长率为2，明显加快。

催化剂组合编号	催化剂组合
A1	200mg 1wt%Co/SiO2-200mg HAP-乙醇浓度1.68ml/min
A2	200mg 2wt%Co/SiO2-200mg HAP-乙醇浓度1.68ml/min

2、再分析A1与A3，我们可以发现A1与A3不同的是乙醇的浓度，A1的浓度是明显大于A3的，再观察增长率，我们发现，A3的增长率是要大于A1的，所以就有乙醇的浓度越高，C4烯烃的选择性越高

A1	200mg 1wt%Co/SiO2-200mg HAP-乙醇浓度1.68ml/min

A3	200mg 1wt%Co/SiO2-200mg HAP-乙醇浓度0.9ml/min

再根据这样的结论以此类推，就可以得出最后的结论。

第四问：

第四问需要重新设计五次实验，并给出合适的理由，在一二三问，只出现了乙醇转化率，C4烯烃的选择性、温度，所以，我们建议在设计实验时，优先考虑乙醇的选择性和脂肪醇，这部分之前的题目并没有涉及，但是出现在了附件中，肯定是有用的，所以我们要在这里对它进行涉及。

而对于实验的参照物，我们在这里给大家举五组例子，为大家做参考：

1、通过分析B1和B2，我们发现，这两组实验，是探究催化剂和乙醇浓度的剂量对于实验的影响，但是，它所涉及到的实验，塑化剂的量和乙醇的量都不同，无法进行对照实验，所以我们就可以说：

需要我们使用50mgCo/Si、100mgHAP与50mgCo/Si、50mgHAP，这两组实验进行对照。

B1	50mg 1wt%Co/SiO2-50mg HAP-乙醇浓度1.68ml/min
B2	100mg 1wt%Co/SiO2-100mg HAP-乙醇浓度1.68ml/min

2、通过分析B3-B7我们可以发现，它想要进行的实验时不同剂量的营销，但是从B5开始乙醇的浓度发生了改变，所以我们就可以再设计几组对照实验，分别对应

与B3进行对照的实验：10mgCo/Si、10mgHAP浓度2.1

与B4进行对照的实验：25mgCo/Si、25mgHAP浓度2.1

与B5进行对照的实验：50mgCo/Si、50mgHAP浓度1.68

B3	10mg 1wt%Co/SiO2-10mg HAP-乙醇浓度1.68ml/min
B4	25mg 1wt%Co/SiO2-25mg HAP-乙醇浓度1.68ml/min
B5	50mg 1wt%Co/SiO2-50mg HAP-乙醇浓度2.1ml/min
B6	75mg 1wt%Co/SiO2-75mg HAP-乙醇浓度1.68ml/min