INDIAN STATISTICAL INSTITUTE

First Semester EXAMINATION: 2019

Course Name: M. TECH (QROR) II

Subject Name: OR II

Date: 27/11/2019 **Maximum Marks: 100** **Duration: 3 hours**

Answer 5. and any three from 1. to 4.

1. a) Define convex, pseudo-convex and quasi-convex function.
   b) Let $f(x_1, x_2) = 2x_1 - 5x_2 + 2x_1^2 - 2x_1x_2 + x_2^2$. Find the Hessian matrix $H(x)$ and show that $H(x) \in$ PSD.
   c) Suppose A is an $m \times n$ matrix and c is an n vector. Then, show that exactly one of the following two systems has a solution:

$$System\,1 \quad Ax \le 0 \text{ and } c'x > 0 \quad \text{for some } x \in R^n$$

$$System\,2 \quad A'y = c \text{ and } y \ge 0 \quad \text{for some } y \in R^m.$$

[6+6 + 8 = 20]

2. a) Define epigraph and sub-gradient of a function.
   b) Let S be a nonempty convex set in R^n and let $f : S \to R$. Then show that f is convex if and only if $epi\ f$ is a convex set.
   c) Let S be a nonempty closed convex set in R^n and $y \notin S$. Then show that there exists a nonzero vector p and a scalar α such that $p'y > \alpha$ and $p'x \le \alpha$ for each $x \in S$.

[6+8 + 6 = 20]

3. a) Formulate the general model for optimization problem. State KKT necessary and sufficient condition for optimality.

   b) Write the differences between Fritz John necessary optimality condition and KKT necessary optimality condition.
   c) Consider the problem

$$\text{Minimize} \quad (x_1 - 3)^2 + (x_2 - 2)^2$$

$$\text{Subject to} \quad x_1^2 + x_2^2 \le 5$$

$$x_1 + 2x_2 \leq 4$$

$$x_1, x_2 \geq 0.$$

Show that Fritz John condition does not hold at $x' = (0, 0)$.

[8+12 = 20]

4. a) Justify that Linear Complementarity Problem is a unified approach for mathematical programming problem.
   b) Define principal pivot transform.
   c) Consider the following LCP (q, M) where

$$M = \begin{bmatrix} 2 & 2 & 0 & 1 \\ 5 & 1 & 6 & 3 \\ 4 & 2 & 1 & 0 \\ 0 & 0 & 3 & 7 \end{bmatrix} \qquad q = \begin{bmatrix} -1 \\ -4 \\ -3 \\ -8 \end{bmatrix}$$

Solve this LCP (q, M) by using Lemke's algorithm.

[4+8+8=20]

5. a) Define copositive, copositive-plus matrix and copositive star matrix.
   b) Consider the matrix

$$A = \begin{bmatrix} 1 & 2 & 4 & 0 \\ 1 & 3 & 1 & 0 \\ 0 & 0 & 3 & 1 \\ 0 & 1 & 2 & 4 \end{bmatrix}$$

Is the matrix A copositive-star?

   c) State a method to solve linear fractional programming problem with illustration.

Discuss critically the advantage of separable programming problem. Give an illustration.

[9+6+15=30]

6. Assignment [10]

/

INDIAN STATISTICAL INSTITUTE

First Semestral Examination : 2019-20

M. Tech. (QR & OR)-II

Industrial Experimentation

Date: 21/11/2019 Maximum Marks: 100 Duration 3 hours

NOTE: (i) This paper carries 114 marks. Answer as much as you can but the maximum you can score is 100. The marks are indicated in [ ] on the right margin.

(ii) The symbols and notations have the usual meaning as introduced in your class.

(iii) Only simple scientific calculators are allowed in the examination hall.

1. An experimenter has run two replicates of a 2^4 design. The following effect estimates have been obtained:

$A = 76.95$	$AB = -51.32$	$BD = 14.74$	$ACD = 10.20$
$B = -67.52$	$AC = 11.69$	$CD = 1.27$	$BCD = -7.98$
$C = -7.84$	$AD = 9.78$	$ABC = -2.82$	$ABCD = -6.25$
$D = -18.73$	$BC = 20.78$	$ABD = -6.50$	

If the raw sum of squares of the data and the sum of the observations are 1055926 and 5441 respectively then test for significance of the effects and write an appropriate regression model based on the results of your test.

[10+6 = 16]

2. What is partial confounding? Illustrate the concepts with a 2^3 factorial confounded in two blocks and, has three replications. Write the ANOVA table showing only the Sources of Variation and their Degrees of Freedom.

[3+9+5 = 17]

3. Consider the following design given in Table 1 involving factors A, B, C and D.

Table 1: The Treatment Combinations (A / B / C / D) in that Order

0010	1002	2021
0101	1120	2112
0222	1211	2200

Identify the design. Find its defining relation and give justification for your answer. Write the aliases of C. What is its resolution and why?

[2+13+5+2 = 22]

4. A structural engineer is studying the strength of aluminium alloy purchased from three vendors (A). Each vendor submits the alloy in *standard-sized bars* (C) of 1.0, 1.5, or 2.0 inches. The processing of different sizes of bar stock from a common ingot involves different forging techniques, and so this factor may be important. Furthermore, the bar stock is forged from ingots made in different heats (B). Each vendor submits two tests

specimens of each size bar stock from the three heats. The resulting strength data are collected from each test specimen.

Identify the design. Obtain the expected mean squares for the design, assuming that vendors and bar size are fixed and heats are random. How do you propose to test the significance of different effects?

[4+12+4 = 20]

6. An experiment was performed to investigate the capability of a measurement system. Ten parts were randomly selected, and two randomly selected operator measured outer diameter of each part three times. The tests were made in random order and the data on outer diameter, in mm, so obtained is given in Table 2.

Table 2: Outer Diameter of Different Parts in millimetre

Part No.	Operator 1				Operator 2		
	Measurements				Measurements		
	1	2	3		1	2	3
1	50	49	50		51	49	52
2	55	55	54		55	55	55
3	53	50	50		55	53	52
4	49	51	50		49	51	52
5	38	39	38		39	40	39
6	52	50	50		53	51	51
7	61	61	61		62	61	61
8	52	50	49		54	49	51
9	50	51	50		52	49	50
10	47	46	49		46	47	48

Analyse the data from this experiment. Suppose the parts have a specification of 50 ± 15 mm, do you consider the measurement system to be adequate for the purpose and why?

[15+4 = 19]

7. a) Starting from the orthogonal matrix B, given below, construct a 3-dimensional simplex for fitting a first-order response surface model in three variables.

$$B = \begin{pmatrix} 1 & -3 & 1 & -1 \\ 1 & -1 & -1 & 3 \\ 1 & 1 & -1 & -3 \\ 1 & 3 & 1 & 1 \end{pmatrix}$$

b) The region of experimentation for three factors are time ($40 \leq T_1 \leq 80$ min), temperature ($200 \leq T_2 \leq 300\ °C$), and pressure ($20 \leq P \leq 50$ psig). A first-order model in coded variables has been fit to yield data. The fitted model is

$$\hat{y} = 30 + 5x_1 + 2.5x_2 + 3.5x_3$$

Do the points (i) $T_1 = 85, T_2 = 325, P = 60$ and (ii) $T_1 = 100, T_2 = 300, P = 56$ lie on the path of steepest ascent?

c) The fitted second-order response surface is

$$\hat{y} = 60 + 5x_1 + 3x_2 + 5x_1^2 + 5x_2^2 + 6x_1x_2.$$

Find the stationary point and characterize it.

[6+6+(5+3) = 20]

Table 3: *F* distribution (5%) Table

$F_{0.05,\, v1,\, v2}$

/	df$_1$=1	2	3	4	5	6	7	8	9	10	12	24
df$_2$=2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.45
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.64
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.77
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.53
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.84
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.41
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.12
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	2.90
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.74
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.61
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.51
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.42
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.35
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.29
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.24
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.19
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.15
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.11
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.08
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.05
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.03
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.01
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	1.98
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	1.96
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	1.89
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.52

Table 4: Percentage Points if the T Distribution

Tail Probabilities

One Tail		0.10	0.05	0.025	0.01	0.005	0.001	0.0005	
Two Tails		0.20	0.10	0.05	0.02	0.01	0.002	0.001	
D	1	3.078	6.314	12.71	31.82	63.66	318.3	637	1
E	2	1.886	2.920	4.303	6.965	9.925	22.330	31.6	2
G	3	1.638	2.353	3.182	4.541	5.841	10.210	12.92	3
R	4	1.533	2.132	2.776	3.747	4.604	7.173	8.610	4
E	5	1.476	2.015	2.571	3.365	4.032	5.893	6.869	5
E	6	1.440	1.943	2.447	3.143	3.707	5.208	5.959	6
S	7	1.415	1.895	2.365	2.998	3.499	4.785	5.408	7
	8	1.397	1.860	2.306	2.896	3.355	4.501	5.041	8
O	9	1.383	1.833	2.262	2.821	3.250	4.297	4.781	9
F	10	1.372	1.812	2.228	2.764	3.169	4.144	4.587	10
	11	1.363	1.796	2.201	2.718	3.106	4.025	4.437	11
F	12	1.356	1.782	2.179	2.681	3.055	3.930	4.318	12
R	13	1.350	1.771	2.160	2.650	3.012	3.852	4.221	13
E	14	1.345	1.761	2.145	2.624	2.977	3.787	4.140	14
E	15	1.341	1.753	2.131	2.602	2.947	3.733	4.073	15
D	16	1.337	1.746	2.120	2.583	2.921	3.686	4.015	16
O	17	1.333	1.740	2.110	2.567	2.898	3.646	3.965	17
M	18	1.330	1.734	2.101	2.552	2.878	3.610	3.922	18
	19	1.328	1.729	2.093	2.539	2.861	3.579	3.883	19
	20	1.325	1.725	2.086	2.528	2.845	3.552	3.850	20
	21	1.323	1.721	2.080	2.518	2.831	3.527	3.819	21
	22	1.321	1.717	2.074	2.508	2.819	3.505	3.792	22
	23	1.319	1.714	2.069	2.500	2.807	3.485	3.768	23
	24	1.318	1.711	2.064	2.492	2.797	3.467	3.745	24
	25	1.316	1.708	2.060	2.485	2.787	3.450	3.725	25
	26	1.315	1.706	2.056	2.479	2.779	3.435	3.707	26
	27	1.314	1.703	2.052	2.473	2.771	3.421	3.690	27
	28	1.313	1.701	2.048	2.467	2.763	3.408	3.674	28
	29	1.311	1.699	2.045	2.462	2.756	3.396	3.659	29
	30	1.310	1.697	2.042	2.457	2.750	3.385	3.646	30
	32	1.309	1.694	2.037	2.449	2.738	3.365	3.622	32
	34	1.307	1.691	2.032	2.441	2.728	3.348	3.601	34
	36	1.306	1.688	2.028	2.434	2.719	3.333	3.582	36
	38	1.304	1.686	2.024	2.429	2.712	3.319	3.566	38
	40	1.303	1.684	2.021	2.423	2.704	3.307	3.551	40
	42	1.302	1.682	2.018	2.418	2.698	3.296	3.538	42
	44	1.301	1.680	2.015	2.414	2.692	3.286	3.526	44
	46	1.300	1.679	2.013	2.410	2.687	3.277	3.515	46
	48	1.299	1.677	2.011	2.407	2.682	3.269	3.505	48
	50	1.299	1.676	2.009	2.403	2.678	3.261	3.496	50
	55	1.297	1.673	2.004	2.396	2.668	3.245	3.476	55
	60	1.296	1.671	2.000	2.390	2.660	3.232	3.460	60
	65	1.295	1.669	1.997	2.385	2.654	3.220	3.447	65
	70	1.294	1.667	1.994	2.381	2.648	3.211	3.435	70
	80	1.292	1.664	1.990	2.374	2.639	3.195	3.416	80
	100	1.290	1.660	1.984	2.364	2.626	3.174	3.390	100
	150	1.287	1.655	1.976	2.351	2.609	3.145	3.357	150
	200	1.286	1.653	1.972	2.345	2.601	3.131	3.340	200
Two Tails		0.20	0.10	0.05	0.02	0.01	0.002	0.001	
One Tail		0.10	0.05	0.025	0.01	0.005	0.001	0.0005	

Tail Probabilities

INDIAN STATISTICAL INSTITUTE

Back Paper Examination : 2019-20

Course name	:	M. Tech. (QR & OR)-II	
Subject Name	:	Industrial Experimentation	
Date: /11/2019		Maximum Marks: 100	Duration 3 hours

NOTE: (i) This paper carries 100 marks. Answer all the questions. The markş are indicated in [ ] on the right margin.

(ii) The symbols and notations have the usual meaning as introduced in your class.

(iii) Only simple and scientific calculators are allowed in the examination hall

1. Write short notes on ***any three*** of the following:
   a) Basic Principles of Experimentation
   b) Duncan's multiple range test.
   c) Randomization in Latin square design.
   d) Partial Confounding
   e) Method of steepest ascent. (7 × 3) = [21]

2. List a fractional factorial design in six factors A, B, C, D, E and F having maximum possible resolution such that three two factor interaction effects AB, CD and EF along with all the main effects are estimable. What is the resolution of your design?

[10+1 = 11]

3. a) What is meant by response surface methodology? What is a response surface? What additional information do we obtain from central composite designs (CCD), relative to factorials or high-resolution fractional factorials?

   b) An experimenter plans to run a CCD in five factors, and want to save experimental effort. He wants to fit a full quadratic model with all main effects, all two-factor interactions and all quadratic terms. He is considering running a 2^{5-1} design for the factorial part of the CCD, instead of a full factorial. Does his approach make sense? Justify your answer.

   c) If the experimenter of part (b), wants to use a rotatable CCD what would be his choice of α? Further, if he decides to include six centre points then how many points would be there in his final design?

[(2+2+2)+4+(1+2) = 13]

INDIAN STATISTICAL INSTITUTE

Back Paper Examination : 2019-20

Course name	:	M. Tech. (QR & OR)-II	
Subject Name	:	Industrial Experimentation	
Date: /11/2019		Maximum Marks: 100	Duration 3 hours

NOTE: (i) This paper carries 100 marks. Answer all the questions. The markş are indicated in [ ] on the right margin.

(ii) The symbols and notations have the usual meaning as introduced in your class.

(iii) Only simple and scientific calculators are allowed in the examination hall

1. Write short notes on ***any three*** of the following:
   a) Basic Principles of Experimentation
   b) Duncan's multiple range test.
   c) Randomization in Latin square design.
   d) Partial Confounding
   e) Method of steepest ascent. (7 × 3) = [21]

2. List a fractional factorial design in six factors A, B, C, D, E and F having maximum possible resolution such that three two factor interaction effects AB, CD and EF along with all the main effects are estimable. What is the resolution of your design?

[10+1 = 11]

3. a) What is meant by response surface methodology? What is a response surface? What additional information do we obtain from central composite designs (CCD), relative to factorials or high-resolution fractional factorials?

   b) An experimenter plans to run a CCD in five factors, and want to save experimental effort. He wants to fit a full quadratic model with all main effects, all two-factor interactions and all quadratic terms. He is considering running a 2^{5-1} design for the factorial part of the CCD, instead of a full factorial. Does his approach make sense? Justify your answer.

   c) If the experimenter of part (b), wants to use a rotatable CCD what would be his choice of α? Further, if he decides to include six centre points then how many points would be there in his final design?

[(2+2+2)+4+(1+2) = 13]

4. Who introduced Robust Parameter Design (RPD) concepts? What was his approach based on? What is deemed crucial for a solution to a RPD problem? Discuss the Response Surface approach to such a problem and illustrate it using an example having two control, two noise factors and an appropriate first-order model. Why and how the transmission of error approach is applied in the response surface approach?

[2+4+2+7+6 = 21]

5. The pressure drop measured across an expansion value in a turbine is being studied. The design engineer considers the important variables that influence pressure drop reading to be gas temperature on the inlet side (A), operator (B), and the specific pressure gauge used by the operator (C). These three factors are arranged in a factorial design, with gas temperature fixed and operator and pressure gauge random. The coded data for two replicates are shown in the Table 1 below.

Table 1: Coded Pressure Drop Data for the Turbine Experiment

Pressure Gauge (C)	Gas Temperature (A)								
	Low (60°F)			Medium (75°F)			High (90°F)		
	Operator (B)			Operator (B)			Operator (B)		
	1	2	3	1	2	3	1	2	3
1	-1	-4	0	10	0	3	-2	-2	-4
	-2	-8	-7	6	2	0	-1	-1	-7
2	-6	-5	-8	12	8	6	-8	1	-8
	4	-1	-2	14	6	2	3	-7	-9
3	-2	-9	-8	14	6	1	-8	-2	-1
	-3	0	-1	14	0	2	-8	2	-2

Analyse the data based on an appropriate model involving all possible main effects and interactions, and test for significance of the effects. Based on the coded data, estimate the model parameters and total variability.

[24+10 = 34]

F distribution (5%) Table

$F_{0.05,\, v1,\, v2}$

	Degree of freedom for the Numerator (v_1)										
Degree of freedom for the Denominator (v_2)	1	2	3	4	5	6	7	8	10	12	24
2	18.5	19.0	19.2	19.2	9.3	19.3	19.4	19.4	19.4	19.4	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.79	8.74	8.64
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	5.96	5.91	5.77
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.74	4.68	4.53
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.06	4.00	3.84
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.64	3.57	3.41
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.35	3.28	3.12
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.14	3.07	2.90
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	2.98	2.91	2.74
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.85	2.79	2.61
12	4.75	3.88	3.49	3.26	3.11	3.00	2.91	2.85	2.75	2.69	2.51
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.67	2.60	2.42
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.60	2.53	2.35
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.54	2.48	2.29
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.49	2.42	2.24
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.45	2.38	2.19
18	4.41	3.55	3.16	2.93	2.77	3.66	2.58	2.51	2.41	2.34	2.15
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.38	2.31	2.11
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.35	2.28	2.08

Indian Statistical Institute
First Semester Examination: 2019-20
M.Tech (QR & OR): II year
Reliability II

Please answer all questions. Marks allotted to each question is given in [ ].

Maximum Marks = 100 Time = 3 hrs. Date : 25 November 2019.

1) Define the following with respect to life distributions:
   a) IFR
   b) DFRA
   c) NBU
   d) NWUE

$[2\,{}^{1}\!/_{2}\; X\, 4 = \mathbf{10}]$

2) a) Show that if f is the density of a DFR distribution, then f(x)> 0 for x≥ 0.
   b) Prove that if F is IFR, then F is IFRA. Construct a counter example to show that the converse is not true.

$[7 + (5 + 8) = \mathbf{20}]$

3) Suppose each of the independent components of a coherent system has an IFRA life distribution. Then prove that the system itself has an IFRA life distribution.

[25]

4) Explain in brief the software quality characteristics. Justify why reliability of a software is the most important quality characteristic?

$[8 + 2 = \mathbf{10}]$

5) a) Discuss in brief the different warranty policies associated with both new and used products.
   b) Describe in brief the different kinds of maintenance done for repairable products.

$[8 + 7 = \mathbf{15}]$

6) An extensive accelerated life time experiment is conducted by subjecting a device to temperatures of 100°C, 200°C and 300°C. The average failure times at these temperatures are 9287 hrs, 5244 hrs and 1295 hrs, respectively and the failure time distribution at each temperature and at normal operating condition is exponential. Determine the mean time to failure of the device at 30°C.

[20]

INDIAN STATISTICAL INSTITUTE

First Semestral Examination: 2019-20

Programme: M. Tech. (QR OR); II Year

Course: Advanced Multivariate Analysis

Date: 18-12-2019 Maximum Marks: 100 Duration: 3½ hours

Note:
1. This paper carries115 marks. Answer all questions but the maximum you can score is 100.
2. All notations have their usual meanings.
3. Give to the point answers.

1) Write Agree or Disagree and briefly Justify

a) Single, complete and average linkage methods would always lead to the same cluster solution.

b) Principal component analysis is an interdependence technique.

c) Discriminant function analysis is an interdependence technique.

d) A model developed by multiple linear regression method represents the underlying causal model.

e) One may carry out ANOVA for each variable instead of a MANOVA.

f) Factor analysis is a dependence technique.

g) Principal Component Regression is an interdependence technique

[2 x 7 = 14]

2) Fill in the gaps:

i) The factors in *factor analysis* are ________ _________ and not the directly observed variables.

ii) ______________ ______________ are defined as linear combinations of the original variables, whereas in ________ _________ the original variables are expressed as linear combinations of the (background) factors

iii) ______________ ______________ is an exploratory data analysis technique that attempts to discover hitherto unknown groups of objects in a data set.

iv) Objects in a *cluster* will have ________ ___________ similarity and those belonging to different *clusters* will have ________ __________ similarity

[1+ 2 + 1 + 2 = 6]

INDIAN STATISTICAL INSTITUTE

First Semestral Examination: 2019-20

Programme: M. Tech. (QR OR); II Year

Course: Advanced Multivariate Analysis

Date: 18-12-2019 Maximum Marks: 100 Duration: 3½ hours

Note:
1. This paper carries115 marks. Answer all questions but the maximum you can score is 100.
2. All notations have their usual meanings.
3. Give to the point answers.

1) Write Agree or Disagree and briefly Justify

a) Single, complete and average linkage methods would always lead to the same cluster solution.

b) Principal component analysis is an interdependence technique.

c) Discriminant function analysis is an interdependence technique.

d) A model developed by multiple linear regression method represents the underlying causal model.

e) One may carry out ANOVA for each variable instead of a MANOVA.

f) Factor analysis is a dependence technique.

g) Principal Component Regression is an interdependence technique

[2 x 7 = 14]

2) Fill in the gaps:

i) The factors in *factor analysis* are ________ _________ and not the directly observed variables.

ii) _____________ _____________ are defined as linear combinations of the original variables, whereas in ________ _________ the original variables are expressed as linear combinations of the (background) factors

iii) _____________ _____________ is an exploratory data analysis technique that attempts to discover hitherto unknown groups of objects in a data set.

iv) Objects in a *cluster* will have ________ __________ similarity and those belonging to different *clusters* will have ________ __________ similarity

[1+ 2 + 1 + 2 = 6]

3) a) What are the purposes of principal component analysis?

b) Let X be a random vector of p variables with dispersion matrix Σ. Let Σ have eigenvalue and eigenvector pairs $(\lambda_1, e_1), (\lambda_2, e_2), \dots (\lambda_p, e_p)$, $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p > 0$. Also let $Y_1, Y_2, \dots Y_p$ be the associated principal components, then show that

   i) $Var\ (Y_i) = \lambda_i$ for all i = 1,...,p,

   ii) Cov (Y_i, Y_j) = 0 for all i ≠ j

   iii) The proportion of the total population variance accounted for by the i^{th} principal component is given by $\frac{\lambda_i}{\sum_{j=1}^{p} \lambda_j}$

c) If some of the eigenvalues are very close to zero, what conclusion can be made regarding collinearity of the variables?

d) If the dispersion matrix Σ is a diagonal matrix will there be any gain in extracting the principal components? – Justify.

e) If some the λ_i's are equal, what can we say about the uniqueness of the corresponding principal component?

[3 + (7 + 1 + 2) + 2+2+3] = 20]

4) a) What do you understand by multicollinearity? How does it affect estimates of regression coefficients?

b) What is Variance Inflation Factor? How does it help to detect the multicollinearity?

c) What are the methods for dealing with multicollinearity?

[(3 + 3) + (1 + 3) + 5 = 15]

5) a) Distinguish between a classification problem and a Clustering problem.

b) Consider a classification problem with two multivariate populations G_1 and G_2. Let $f_1(x)$ and $f_2(x)$ be the pdfs corresponding to G_1 and G_2 respectively. Show that the classification rule for allocating a new object to one of the two populations obtained by minimizing the expected cost of misclassification (ECM) depends on the density ratio, cost ratio and probability ratio.

c) In a particular case misclassification costs were estimated as C(2 / I) = INR.1600/- and C(1 /2) = INR 3400/-. It is also known that population G_1 is twice as large as population G_2. Measurements on a new object yield the density values as: $f_1(x)$ = 0.25 and $f_2(x)$ = 0.5. Assign the object to G_1 or G_2.

d) What are OER, AER and APER? What is the disadvantage of APER?

[3 + 8 + 3 + (4 + 2) = 20]

6) a) What are the similarities and dissimilarities between Principal Component Analysis and Factor Analysis?

b) Write down the principal component estimates of factor loadings. Explain the notations used. Show that contribution of the jth factor to the total sample variance is θ_j, where $\theta_1, \theta_2, \ldots\ldots \theta_p$ are the eigenvalues of S, the sample variance covariance matrix.

c) In an exploratory factor analysis study, the following factor solutions were obtained from the correlation matrix:

Variables	Principal Component Loadings		Varimax Rotated Loadings	
	F_1	F_2	F_1^*	F_2^*
Y_1	.817	-.157	**.732**	.395
Y_2	.838	-.336	**.861**	.271
Y_3	.874	.288	.494	**.776**
Y_4	.838	-.308	**.844**	.292
Y_5	.762	.547	.244	**.905**

i) Compute communalities for each variables from both Principal Component and Varimax Rotated loadings.

ii) Why the communalities computed before and after the rotation remained unchanged? State and prove the result in general.

iii) Calculate proportion to total variance due to each factor after rotation.

[5+ 3 + (5 + 4 + 3) = 20]

7) a) What are the differences between hierarchical and partitioning method of clustering?

b) Describe the K – means algorithm using a flowchart.

c) Measurements on two variables Y_1 and Y_2 for four objects A, B, C and D are given below. Using K-means algorithm, group the objects into K = 2 clusters. Initial centroids are given as (0,0) and (1,1)

Objects	Observations	
	Y_1	Y_2
A	4	3
B	0	-3
C	-2	0
D	2	0

d) Why is initial seed selection important in K-means Clustering?

[4 + 4 + 10 + 2 = 20]

INDIAN STATISTICAL INSTITUTE

First Semester Examination: 2019 – 20

Course Name: M Tech (QROR), 2nd Year
Subject Name: Business Analytics

Date: 29 November 2019 **Maximum Marks: 100** **Duration: 3 Hours**

Notes: Answer question 1 and any 4 from the rest. The total marks is 105. The maximum you can score is 100.

1. Answer the following.
   a. Provide a sketch of typical training error and test error curves on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. Explain how this curve is used to assess over fitting. [6 + 3 = 9]
   b. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred? [6 + 2 + 2 = 10]
   c. Consider the Gini index, classification error, and cross-entropy in a simple classification setting with two classes. Create a single plot that displays each of these quantities as a function of the estimated class probabilities. The x axis should display the estimated class probability, ranging from 0 to 1, and the y-axis should display the value of the Gini index, classification error, and entropy. [3 X 2 = 6]
2. Answer the following
   a. Explain briefly the concept of cost complexity pruning in the context of a decision tree used for value estimation. What happens when the tuning parameter is zero or very large? [5 + 3 = 8]
   b. What are the different parameters used in boosting a tree? Which of the parameters are likely to impact the chance of over fitting? Write the boosting algorithm in the contest of regression tree. [3 + 3 + 6 = 12]
3. Answer the following
   a. Someone claims that automatic detection of fraudulant credit card transaction is a supervised analytics problem whereas a classification technique may be used but a medical insurance fraud carried out jointly by the medical practitioner and the patient cannnot be addressed using a supervised classification technique. Do you agree? Explain. [4]
   b. State the different assumptions of LDA. [3]
   c. Suppose you have 10 explanatory variables. How many parameters will you have to estimate if you need to fit a QDA? [3]
   d. Suppose you are trying to fit a model to estimate the probability of death given the severity of accident. Suppose the severity may assume 5 possible values – 1, 2, 3, 4, 5. The number of accidents and deaths for these 5 different severity classes are (30,0), (16,2), (6,4), (9,6) and (13,10) respectively, i.e. 22 deaths were observed in 74 accidents. Do you think it would be wise to fit a logistic regression model to this data? Explain. [4]
   e. Give one real-life example each of situations where classification and value estimation might be useful. Describe the response, as well as the predictors. Is the goal of each application explanation or prediction? [3 + 3 = 6]

4. Answer the following
    a. Explain the K-fold cross validation method. [5]
    b. What are advantages and disadvantages of the K-fold cross validation approach relative to
        i. Validation set approach
        ii. LOOCV [4 + 4 = 8]
    c. Suppose we are interested in predicting the % change in the US dollar in relation to the weekly changes in the US, British and German stock markets. We collect weekly data for all of 2018. For each week we record the % change in the dollar, the % change in the US market, the % change in the British market, and the % change in the German market. We plan to use a multiple linear regression model. Do you think this approach may have some problems? If yes, why? [4]
    d. Suppose we have collected a set of data on the top 500 firms in India. For each firm we record profit, number of employees, industry segment and the CEO salary. We are interested in understanding which factors impact CEO salary.
        i. What are the explanatory and response variables?
        ii. Is this a value estimation or a classification problem?
        iii. Is it an explanatory or predictive problem? [3 X 1 = 3]
5. Answer the following
    a. Consider 24 observations from a time series – 106, 107, 98, 98, 101, 99,102, 104, 97, 103, 107, 105, 106, 98, 99, 96, 95, 99, 100, 102, 108, 106, 104, 98. Do the series appear to be stationary? [6]
    b. Examine the following EACF matrix and identify the possible ARIMA / ARMA models. [6]

p/	Q					
	0	1	2	3	4	5
0	X	X	X	X	X	X
1	X	X	X	X	X	X
2	X	0	0	0	0	0
3	X	X	0	0	0	0
4	X	X	X	0	0	0

    c. State the stationarity and invertibility conditions for AR(2) / ARMA(2,q) and MA(2) / ARMA(p,2) process. How will the conditions change for ARMA(p,q) processes with p / q > 2? [4 + 4 = 8]
6. Let k be the bias parameter of ridge regression.
    a. What are the estimating equations for the ridge regression coefficients? [4]
    b. What would be estimated coefficients if k increases indefinitely? [2]
    c. Suppose we are fitting a ridge regression model with three explanatory variables. Examine the ridge trace table given below and answer the following. (Note: P_1, P_2 and P_3 are the estimated parameters. SSE(k) gives the sum of squared residuals.)
        i. What are the OLS estimates of the parameters? [2]
        ii. Suggest some 'good' values of k with a brief explanation. Can you draw the ridge trace? [7 + 3 = 10]

iii. Do you expect the ridge regression with the suggested value of k to have a higher R^2 compared to the model fitted using OLS? [2]

K	P_1	P_2	P_3	SSE(k)
0.000	-0.339	0.213	1.303	0.0810
0.001	-0.117	0.215	1.080	0.0837
0.003	0.092	0.217	0.870	0.0911
0.005	0.192	0.217	0.768	0.0964
0.007	0.251	0.217	0.709	0.1001
0.009	0.290	0.217	0.669	0.1027
0.010	0.304	0.217	0.654	0.1038
0.012	0.328	0.217	0.630	0.1056
0.014	0.345	0.217	0.611	0.1070
0.016	0.359	0.217	0.597	0.1082
0.018	0.370	0.216	0.585	0.1093
0.020	0.379	0.216	0.575	0.1102
0.022	0.386	0.216	0.567	0.1111
0.024	0.392	0.215	0.560	0.1118
0.026	0.398	0.215	0.553	0.1126
0.028	0.402	0.215	0.548	0.1132
0.030	0.406	0.214	0.543	0.1139
0.040	0.420	0.213	0.525	0.1170
0.050	0.427	0.211	0.513	0.1201
0.060	0.432	0.209	0.504	0.1234
0.070	0.434	0.207	0.497	0.1271
0.080	0.436	0.206	0.491	0.1310
0.090	0.436	0.204	0.486	0.1353
0.100	0.436	0.202	0.481	0.1400