INDIAN STATISTICAL INSTITUTE

M-Stat I, First Semester, 2019-20

MULTIVARIATE ANALYSIS

Date: 29/11/2019 Semester Examination Time: 3 hours

[Total points 53. Maximum you can score is 50. The use of a calculator is allowed.]

1. Suppose $z_0, z_1, \ldots, z_p$ are independently and identically distributed with mean 0 and variance $\sigma^2$. Let $X_i = z_0 + z_i$, $i = 1, 2, \ldots, p$. Verify that there is a principal component of $X = (X_1, \ldots, X_p)$ which is proportional to $\bar{X}$. Argue that this is the principal component with the maximum variance, i.e. the first principal component. [8 points]

2. Let $\boldsymbol{X} = (X_1, X_2)$ represent the head length and head breadth measurements of the first son in the family. Similarly let $\boldsymbol{Y} = (Y_1, Y_2)$ represent the head length and head breadth of the second son in the family. In a sample of 25 families, the matrix of correlations of $(\boldsymbol{X}, \boldsymbol{Y})$ is given by

$$\boldsymbol{R} = \left( \begin{array}{cc|cc} 1.0000 & 0.7346 & 0.7108 & 0.7040 \\ 0.7346 & 1.0000 & 0.7470 & 0.7086 \\ \hline 0.7108 & 0.7470 & 1.0000 & 0.8392 \\ 0.7040 & 0.7086 & 0.8392 & 1.0000 \end{array} \right).$$

Find the two canonical correlations between $\boldsymbol{X}$ and $\boldsymbol{Y}$, and the canonical variable pairs. Can you argue from the form of the correlation matrix that the second canonical correlation should be close to zero? [7+1=8 points]

3. A study was performed on market value profitability measures. With 8 variables, the rotated principal components estimates of factor loadings for a 3 factor model is obtained as:

| | Estimated factor loadings | | |
|---|---|---|---|
| Variable | $F_1$ | $F_1$ | $F_3$ |
| Historical return on assets | 0.433 | 0.612 | 0.499 |
| Historical return on equity | 0.125 | 0.892 | 0.234 |
| Historical return on sales | 0.296 | 0.238 | 0.887 |
| Replacement return on assets | 0.406 | 0.708 | 0.483 |
| Replacement return on equity | 0.198 | 0.895 | 0.283 |
| Replacement return on sales | 0.331 | 0.414 | 0.789 |
| Market Q ratio | 0.928 | 0.160 | 0.294 |
| Market relative excess value | 0.910 | 0.079 | 0.355 |

(a) Determine the specific variances, communalities, and the proportion of total variance explained by each factor. Does a 3 factor model appear to be appropriate to you?

(b) Assuming that estimated loadings below 0.4 are small, can you interpret the three factors? [6+2=8 points]

4. (a) State the orthogonal Factor Analysis Model with $p$ variables and $m$ factors ($m < p$). Show that under the usual assumptions the factor loading matrix $L$ is uniquely defined only up to orthogonal rotations of the coordinate system.

(b) Consider the following model with $p = 3$ and $m = 2$.

$$\begin{aligned} X_1 &= 0.5F_1 + 0.5F_2 + e_1 \\ X_2 &= 0.3F_1 + 0.3F_2 + e_2 \\ X_3 &= 0.5F_1 - 0.5F_2 + e_3 \end{aligned}$$

Suggest a factor rotation which allows better interpretation of this model.

[3+5=8 points]

5. Consider the general discriminant analysis problem with $g$ populations, $\pi_1, \dots \pi_g$. Let $p_i$ be the prior probability for $\pi_i$, and let $c(j|i)$ be the cost of allocating any item from $\pi_i$ to $\pi_j$. Prove that the classification regions that minimize the Expected Cost of Misclassification (ECM) are defined by: allocate an individual with observation $x$ to the population $\pi_k$, $k = 1, 2, \dots, g$, if the sum $\sum p_i f_i(x) c(j|i)$ is minimized at $j = k$, where the sum is over $i = 1, \dots, g$, $i \neq j$, and $f_i(x)$ is the density of $X$ under $\pi_i$. [7 points]

6. The following matrix represents dissimilarities between pairs of companies; the dissimilarity measure is the number of variables in which two companies have different scores out of a set of 32 binary variables.

| Company | A | B | C | D | E | F | G | H |
|---|---|---|---|---|---|---|---|---|
| A | - | 3 | 16 | 16 | 18 | 26 | 21 | 25 |
| B | | - | 15 | 15 | 19 | 26 | 24 | 22 |
| C | | | - | 6 | 22 | 24 | 23 | 19 |
| D | | | | - | 22 | 20 | 21 | 21 |
| E | | | | | - | 21 | 17 | 13 |
| F | | | | | | - | 13 | 15 |
| G | | | | | | | - | 8 |

Applying the Single Linkage and Complete Linkage methods, obtain dendrograms for clustering these companies, and comment on the form of the dendrograms. What similarities and differences do you see between the two dengrograms? [5+2=7 points]

7. Given an $n \times n$ dissimilarity matrix $D$, give a necessary and sufficient condition for the matrix to be Euclidean. Prove the necessity and sufficiency of the given condition.

[7 points]

Indian Statistical Institute

M.Stat. First year, First Semestral Exam: 2019-20

Regression Techniques

Maximum Marks: 80, Duration: 3 hours Date : 21.11.2019

**Answer all questions. Show your steps to get full credit.**

1. (a) Consider the standard multiple regression model (with $p$ predictors) $Y = X\beta + \epsilon$, where $\epsilon \sim N(0, \sigma^2 I_n)$. Find $\hat{\sigma}^2_M$, the MLE of $\sigma^2$. Suppose that $\frac{p}{n} \to c$, for $0 < c < 1$. Show that $\hat{\sigma}^2_M \xrightarrow{P} (1-c)\sigma^2$. [3+5]

   (b) In a biomedical study, data are collected on total medical expenditure and related covariates for a particular season of the year from $n$ individuals. In the data, $n_1 (< n)$ responses are found to be exactly zero and the non-zero (continuous) responses are denoted by $Y_1, Y_2 \ldots, Y_{n_2}$, where $n_1 + n_2 = n$. There are $p$ regressor variables $X_1, X_2, \ldots, X_p$ in the data. Suggest a suitable regression model for predicting the medical expenditure for the $(n+1)$-th individual given the relevant information on the predictors. [9]

2. Financial regulations require banks to report their daily risk measures called Value at Risk (VAR). Let $Y$ be the financial return of the bank and then for a given $\theta$ $(0 < \theta < 1)$, the VAR is the value $y^*$ satisfying $P(Y \leq y^*) = \theta$. The financial return depends on exchange rate $(x)$. Based on a sample of $n$ data points $(Y_i, x_i)$, develop a statistical approach for estimating VAR at $\theta$=0.80, and $\theta$=0.90. Specify explicitly:

   (i) statistical model under consideration, (ii) the method for estimating the regression coefficients, and (iii) limitations of the model (and method) used, if any. [5+8+5]

3. Explain briefly what is meant by each of the following:
   (i) Missing at Random (MAR), (ii) Mallow's Cp, (iii) Weighted False Discovery Rate (WFDR), (iv) Semi-parametric Regression Model, (v) Curse of Dimensionality. [15]

4. (a) What is the LASSO estimate of $\beta$ under the standard multiple regression model $Y = X\beta + \epsilon$, where $\epsilon \sim N(0, \sigma^2 I_n)$? Show that the LASSO estimate of $\beta$ is indeed the posterior mode with respect to a double exponential prior distribution for $\beta$. [7]

   (b) Let $W$ and $V$ be two orthogonal subspaces with projection matrices $P_w$ and $P_v$ respectively. Show that $||P_w y||^2$ and $||P_v y||^2$ are independently distributed, where $y \sim N(\mu, \sigma^2 I_n)$. [5]

(c) Discuss the importance of PRESS $R^2$ in regression diagnostics. [3]

5. (a) Define the "systematic component" of a generalized linear model (GLM). What is meant by "canonical link"? Find the canonical link function for a Poisson GLM. [5]

(b) Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables with a Poisson GLM with linear predictor (for the $i$-th subject) $\eta_i = x_i^T\beta$, and the link function $g$; where $x_i$ denotes the set of covariates for the $i$-th subject. Define $Z_i$=1, if $Y_i > 0$; and $Z_i$=0, otherwise. Show that $Z_1, Z_2, \ldots, Z_n$ follow a binary GLM. Find the link function of this GLM. [4]

(c) If $Z_1, Z_2, \ldots, Z_n$ have to follow a binary GLM with logit link, what will be the link function of the Poisson GLM?

[6]

Indian Statistical Institute
Backpaper Examination
First Semester, 2019-2020 Academic Year
M.Stat. First Year
Statistical Inference I

Date: 15.01.2020 Maximum Marks : 100 Duration: 4 Hours

Answer all questions

1. Consider a decision problem with $\Theta = \mathcal{A} = R$ and loss function $L(\theta, a) = k_1|\theta - a|I_{a \le \theta} + k_2|\theta - a|I_{a > \theta}$, where $k_1 > 0$ and $k_2 > 0$. Define $f(b) = EL(Z, b)$ for an arbitrary random variable $Z$ with finite first moment.

   (a) Show that $f(.)$ is minimized at $b$ which is a $p$-th quantile of the distribution of $Z$, for some $p$, where $p$ is some function of $k_1$ and $k_2$.

   (b) What will be the Bayes rules for this loss function ? [8+4=12]

2. Prove if a minimal complete class exists, it consists of exactly the admissible rules. [10]

3. Let the parameter space $\Theta$ be $R$ and suppose the risk $R(\theta, \delta)$ is a continuous function of $\theta$ for all decision rules $\delta$. Suppose that $\delta_0$ is a Bayes rule with respect to a prior distribution. Under appropriate conditions on the prior, show that $\delta_0$ is admissible. [12]

4. Show that in a finite parameter decision problem, the risk set is the convex hull of the risk set of the non-randomized rules. [10]

5. Stating appropriate conditions, prove the Minimax Theorem in a decision problem with finite parameter space. [20]

6. State and prove the Rao-Blackwell Theorem. [10]

7. Consider the testing problem of $H_0 : \theta_1 = \theta_0$ vs. $H_1 : \theta_1 \neq \theta_0$ for $\theta_0 \in R$ where the observation vector is obtained from a two-parameter exponential family with natural parameter $(\theta_1, \theta_2)$ and $\theta_1 \in R$. Derive the UMPU test of size $\alpha$ for this testing problem with full justification of all your steps. [16]

8. Let $X$ be a random observable following a distribution with a density involving a parameter $\theta$. Our problem is to test $H_0 : \theta = \theta_0$ against $H_1 : \theta = \theta_1$ where $\theta_0$ and $\theta_1$ are two specified parameter values. Consider the usual 0-1 loss function and find the risk of a randomized test given by a test function $\phi$. Show that the risk set $S$ is a convex subset of $R^2$. [10]

# Stochastic Processes

M. Stat 1st year, 1st semester
Final examination
Time: 3 hours
**Total points: 60**

November 27, 2019

Answer **all 4** questions. Each question carries **15 points**. Results done in the class may be quoted and used.

1. Suppose that $(X_t : t \geq 0)$ is an irreducible CTMC on a finite state space $\{1, \ldots, N\}$ with infinitesimal generator $Q$. If $P(1)$, which is the transition matrix at lag 1, is doubly stochastic, that is,
$$\sum_{i=1}^{N} P_{ij}(1) = 1\,, 1 \leq j \leq N\,,$$
then show that
$$\sum_{i=1}^{N} Q_{ij} = 0\,, 1 \leq j \leq N\,,$$
where $Q_{ij}$ and $P_{ij}(1)$ are $(i,j)$-th entries of $Q$ and $P(1)$, respectively.

2. Consider the usual age dependent branching process starting with 1 individual, where every individual lives for an Exponential $(\lambda)$ time, and at the time of death gives birth according to a progeny distribution of finite mean $m$. If $X_t$ denotes the number of individuals who have **died** by time $t$, calculate
$$\mathrm{E}(X_t)\,,$$
for every $t \geq 0$.

3. An urn has 1 white ball and 1 black ball. A ball is drawn from the urn at random. If the ball drawn is white, it is replaced by 2 black balls, and if the ball drawn is black, it is replaced by a random number of white balls chosen according to Geometric(1/2), that is, the distribution which puts mass $1/2, 1/4, 1/8, \ldots$ on $0, 1, 2, \ldots$, respectively. The process is repeated independently. Calculate the probability that urn will eventually become empty.
**Hint:** Think what would have happened if the urn had no black ball at the beginning.

4. Let $X_1, X_2, \ldots$ be i.i.d. from **Exponential(1)**, and suppose that $Y$ follows **Exponential(2)** independently of $(X_1, X_2, \ldots)$. Define

$$S_n = \sum_{i=1}^{n} X_i \,, n \geq 0 \,,$$

$$S'_n = Y + \sum_{i=1}^{n} X_i \,, n \geq 0 \,,$$

and let

$$N'(t) = \sum_{n=0}^{\infty} \mathbf{1}(S'_n \leq t) \,, t \geq 0 \,.$$

(a) **(5 points)** Show that for $t \geq 0$ and $n \geq 1$,

$$P(S'_n \leq t) = \int_0^t P\left(Y + s + S_{n-1} \leq t\right) e^{-s} ds \,.$$

(b) **(10 points)** Calculate

$$\mathrm{E}\left(N'(t)\right) \,,$$

for all $t \geq 0$.

# Indian Statistical Institute
## Final Examination

November 25, 2019

*Categorical Data Analysis, M1*

Total points: 40 Time: 3 hours

**Note:** This is a **closed notes/closed book** examination. Notations, if not explicitly explained, are to be interpreted as defined in class.

### 1. Another measure of predictive association [10]

Define a new measure of predictive association $\tau_{C|R}$ as the relative decrease in risk (with respect to the 0-1 loss) when one uses knowledge of $R$ in predicting $C$ versus when one does not, as per the following procedures:

(i) Without any knowledge of $R$, predict $C$ by drawing a random class from the pmf $(p_{\cdot j})$.

(ii) When $R$ is known, predict $C$ by taking a random class from the pmf $(p_{Rj}/p_{R\cdot})$.

Give an expression for $\tau_{C|R}$ in terms of the $p_{ij}$'s. What are the maximum and minimum values of $\tau_{C|R}$, and when are those achieved? When are the values 0 and 1 taken? Show that, in the special case when there are at least as many row categories as there are column categories and the column marginals are unifrom, $\tau_{C|R}$ equals the square of Cramer's $V$.

### 2. Logistic regression with a categorical predictor [3 + 3 + 4]

Suppose that you have two binary categorical variables $X$ and $Y$. Consider the logistic model

$$\text{logit}(\mathbb{P}(Y = 1|X)) = \beta_0 + \beta_1 X. \tag{1}$$

(a) Show that Model (1) is equivalent to a product-binomial model for the corresponding contingency table. Show that testing $\beta_1 = 0$ is the same as testing for homogeneity in the contingency table.

(b) Describe how one can perform a permutation test for testing $\beta_1 = 0$ in Model (1).

(c) Derive MLEs of $\beta_0$ and $\beta_1$ under Model (1). Also, derive an estimator of the asymptotic covariance matrix of the MLEs.

### 3. Correspondence analysis, SEM, causal inference [2 + 5 + 3]

(a) Suppose a contingency table $X$ has a block-diagonal structure, i.e.

$$X = \begin{pmatrix} X_1 & O \\ O & X_2 \end{pmatrix},$$

where $O$ denotes a matrix of all zeros of suitable dimension. After performing correspondence analysis on $X$, if we just take two factors, i.e. $r_1, r_2$ and $s_1, s_2$, and plot them on the plane, how would the plot look like qualitatively?

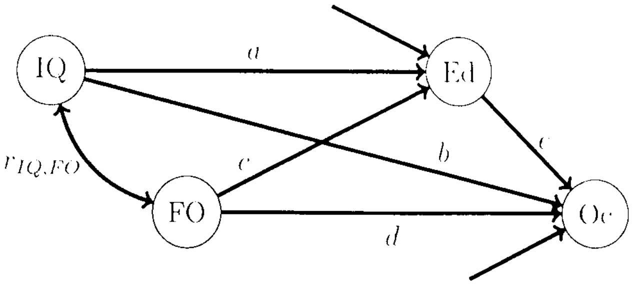


Figure 1: A path model for explaining Occupation (Oc) using explanatory variables Education (Ed), Father's Occupation (FO), and IQ.

(b) Consider the path diagram in Figure 1. Identify the endogenous and the exogenous variables, and write down the structural equations for this model. Using Wright's tracing rule or otherwise write down the correlation matrix of these variables in terms of the path coefficients and the exo-correlations. Argue for/against the validity of this model as a causal model.

(c) In an observational study, suppose we have binary treatment and response. Derive estimable upper and lower bounds on the causal effect $\theta$. Compute these bounds based on the data in Table 1. Compare with the true causal effect computed via oracle knowledge. Also compute the association $\alpha$ from this table (here, as usual, $X, Y$, and $C_i$ denote, respectively, the treatment variable, the response variable, and the potential outcomes).

| $X$ | $Y$ | $C_0$ | $C_1$ |
|---|---|---|---|
| 0 | 0 | 0 | 0 |
| 0 | 1 | 1 | 1 |
| 1 | 1 | 0 | 1 |
| 1 | 0 | 0 | 0 |
| 1 | 1 | 1 | 1 |

Table 1: Data from a fictitious "observational" study.

## 4. Graphical models

[3 + 3 + (1 + 1 + 2)]

(a) Consider three binary variables $X_1, X_2, X_3$ with joint pmf

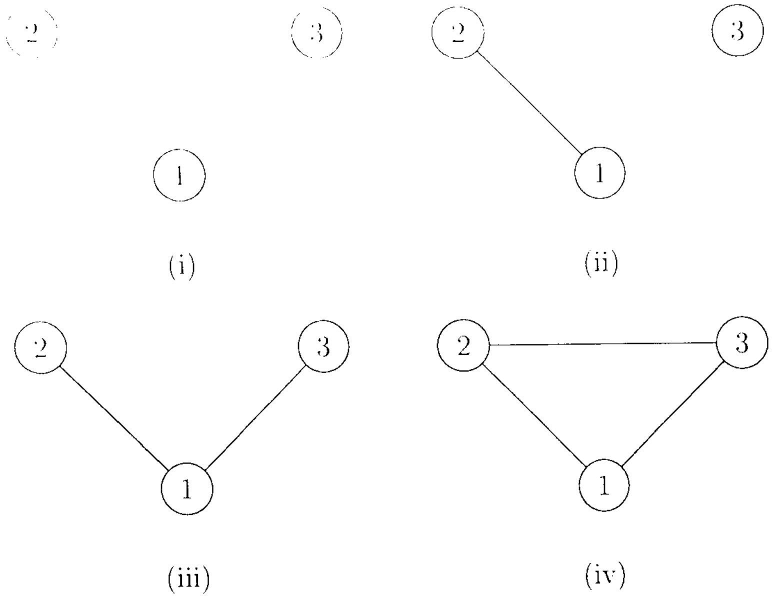
$$\begin{aligned} p(0,0,0) &= 1/3 \\ p(1,1,0) &= 1/6 \\ p(1,0,1) &= 1/6 \\ p(0,1,1) &= 1/6 \\ p(1,1,1) &= 1/6. \end{aligned}$$

Does this distribution factorize over the graphs in Figure 2? Does it satisfy any of the Markov properties with respect to these graphs?

(b) Suppose $(A, B, S)$ is a weak-decomposition of a graph $\mathcal{G}$. Suppose a probability distribution $\mathbb{P}$, with density $f$, is globally markov with respect to $\mathcal{G}$. Show that

$$f(x)f(x_S) = f(x_A, x_S)f(x_B, x_S).$$

Show also that $\mathbb{P}_{A\cup S}$ and $\mathbb{P}_{B\cup S}$ are globally Markov with respect to $\mathcal{G}_{A\cup S}$ and $\mathcal{G}_{B\cup S}$ respectively.



(i) (ii) (iii) (iv)

Figure 2: All non-isomorphic graphs on three vertices.

(c) Consider hierarchical log-linear models on three binary factors with generators $\mathcal{A}_0 = \{\{2,3\}\}$ and $\mathcal{A}_1 = \{\{1,2\},\{2,3\}\}$.

(i) Draw the incidence graphs corresponding to these models and check if any of these models are graphical.

(ii) What does the first model say about variable 1?

(iii) What would be the degrees of freedom of the limiting $\chi^2$ distribution for the likelihood ratio statistic for testing $H_0 : \mathcal{A} = \mathcal{A}_0$ vs $H_1 : \mathcal{A} = \mathcal{A}_1$?

# Indian Statistical Institute

## Backpaper Examination

January 16, 2020

*Categorical Data Analysis, M1*

Total points: 100 Time: 3 hours

**Note:** This is a **closed notes/closed book** examination. Notations, if not explicitly explained, are to be interpreted as defined in class.

### 1. Odds ratios

[6 + 14]

Consider an $r \times c$ product-multinomial table.

(a) Show that the odds ratios $\frac{p_{ij}p_{i'j'}}{p_{ij'}p_{i'j}}$. $1 \leq i.i' \leq r.1 \leq j.j' \leq c$, may be used for testing homogeneity.

(b) Find out the MLE of $\frac{p_{ij}p_{i'j'}}{p_{ij'}p_{i'j}}$ and derive its asymptotic distribution.

### 2. Measures of agreement

[12 + 2 + 2 + 3 + 1]

(a) What is Cohen's $\kappa$? Why is it considered a measure of agreement? What are its extremal values and when are these achieved? When is the value 0 achieved? Describe how one can compute the extremal values of $\kappa$ given fixed row and column marginals.

(b) Describe the notion of conditional agreement and propose a measure for it.

(c) Write down the maximum likelihood estimates of Cohen's $\kappa$ and your proposed measure.

(d) What are some drawbacks of Cohen's $\kappa$ as a measure of agreement and how to fix them?

(e) Is $1 - \kappa$ a good measure of disagreement?

### 3. Logistic regression

[10 + 4 + 6]

(a) Show that, for linearly separable data, MLE of logistic regression model parameters do not exist.

(b) Suppose, in the context of logistic regression, you want to test if the predictors have any influence on the outcomes. What is the Wald test for this hypothesis? What can you say about its power function when the predictors have very strong influence on the outcomes?

(c) What kind of logistic models are used when the outcome variable is ordinal?

### 4. Log-linear models

[4 + 3 + 5 + 3 + 5]

(a) What are log-linear models? Define (i) hierarchical, (ii) graphical, and (iii) decomposable log-linear models.

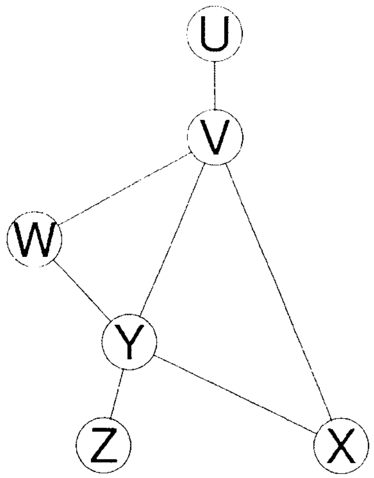


Figure 1: Incidence graph of a decomposable graphical model.

(b) Give an example of a hierarchical model which is not graphical, and an example of a graphical model which is not decomposable.

(c) What are the main benefits of fitting graphical models? Derive an explicit formula for the number of graphical models for a $d$-dimensional table.

(d) Describe the Iterative Proportional Scaling (IPS) algorithm for finding MLEs in hierarchical models.

(e) Consider the decomposable graphical model shown in Figure 1. Find out an RIP ordering of its cliques. Assuming that the variables $U, V, W, X, Y, Z$ are all binary, write down an expression for the MLE of $m_{01\bullet\bullet01}$.

## 5. Causal Inference, SEM

[8 + 2 + 4 + 6]

(a) Under the potential outcomes framework of Neyman-Rubin, define the causal effect $\theta$ and the association $\alpha$. For what kind of experiments do we have $\theta = \alpha$? Give an example where $\theta < 0$ but $\alpha > 0$. How can one estimate $\theta$?

(b) How are confounding variables taken care of in the potential outcomes framework?

(c) Show why Simpson's paradox disappears under the potential outcomes framework.

(d) What are Wright's tracing rules in the context of path analysis. State and prove the fundamental theorem of path analysis. Show how the tracing rules follow from this theorem.