**INDIAN STATISTICAL INSTITUTE**

MS in QMS

TEST ON MULTIVARIATE DATA ANALYSIS

Date: 04 May 2026 Time: 3 hours Maximum Marks: 50

*Answer as many questions as you can. The maximum you can score is 50*

---

1. Explain how multicollinearity in regression models leads to inflated variances of the estimated coefficients. Define ridge regression and lasso regression, highlighting their mathematical formulations and underlying principles. How do these methods differ in terms of regularization and coefficient shrinkage? Describe how model coefficients are computed in ridge regression.

[12]

2. Explain discriminant analysis. Explain the role of Bayes theorem in discriminant analysis. Compare discriminant analysis and logistic regression. Differentiate between linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). What assumptions about covariance structures distinguish them, and under what circumstances is each method most appropriate?

[10]

3.
   a. State and prove the variance decomposition theorem as applied in principal component analysis (PCA). How does this theorem establish the relationship between total variance and the variance explained by principal components?
   b. State and prove the theorem that the first principal component maximizes variance among all possible linear combinations of the original variables. Why is this property fundamental to the construction of PCA?

[10]

4.
   a. Compare principal component analysis (PCA) with factor analysis. Discuss their objectives, assumptions, and methodological differences?
   b. State and prove the spectral (eigenvalue) decomposition theorem as applied in factor analysis. How does this theorem establish the relationship between covariance/correlation matrices and latent factor structures?
   c. State and prove the theorem of rotational indeterminacy in factor analysis. Why does the factor solution remain unchanged under orthogonal rotation?

[12]

5.

a. State and prove the variance decomposition theorem as applied in cluster analysis. How does this theorem establish the relationship between total variance, within-cluster variance, and between-cluster variance?
b. State and prove the theorem that the cluster mean as centroid minimizes the sum of squared distances from centroid within a cluster. Why is this property fundamental to the optimality of the k-means clustering algorithm?
c. State and prove the convergence theorem for k-means clustering. Explain why the iterative process of reassigning points and updating centroids is guaranteed to converge to a local optimum, and discuss the implications of this result.

[11]