Applied Multivariate Data Analysis – Multivariate Data

- Characterizing and Displaying Multivariate Data
- Testing Assumptions
- Multivariate Normal Distribution

I. Characterizing and Displaying Multivariate Data

1. Mean and Variance of a Univariate Random Variable

- Mean (Expected Value):
  The "center" of the distribution of a variable X.

$$\mu_X = E[X] = \sum_i x_i p(x_i) \quad \text{(discrete)}$$

$$\mu_X = \int_{-\infty}^{\infty} x f_X(x) dx \quad \text{(continuous)}$$

- Sample mean:

$$\bar{x} = \tfrac{1}{n} \sum_{i=1}^{n} x_i$$

- Variance:
  Measures spread or dispersion around the mean:

$$\sigma_X^2 = Var(X) = E[(X - \mu_X)^2]$$

- Sample variance:

$$s^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

- Notes:
  - Standard deviation $\sigma = \sqrt{\sigma^2}$
  - Units are same as data for mean, squared units for variance.

2. Covariance and Correlation of Bivariate Random Variables

- Covariance: Measures joint variability between X and Y

$$Cov(X,Y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

  - Positive covariance → variables increase together
  - Negative covariance → one increases, other decreases
- Correlation coefficient (Pearson's r): Standardized covariance

$$r_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

  - Ranges from -1 to 1
  - r = 1 → perfect positive linear relationship
  - r = -1 → perfect negative linear relationship
  - r = 0 → no linear association

3. Scatter Plots of Bivariate Samples

- Purpose: Visualize relationships between two variables.
- Features to observe:
  - Linear/non-linear trends
  - Outliers
  - Clustering

4. Graphical Displays for Multivariate Samples

- Scatterplot matrix: Shows all pairwise relationships.
- Boxplots: Identify outliers per variable.
- Histograms: Show univariate distributions.
- Heatmaps of correlations: Quick visualization of relationships between all variables.

Tip: Color-code by group for categorical distinction.

5. Mean Vectors

- Definition: Vector of means for p variables:

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix}$$

- Interpretation: Represents the “center” in p-dimensional space.

6. Covariance Matrices

- Definition: Square matrix showing covariance between each pair of variables

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{bmatrix}$$

- Properties:
  - Symmetric ($\sigma_{ij} = \sigma_{ji}$)
  - Positive semi-definite (all eigenvalues nonnegative)
- Use: Measures overall joint variability among variables.

7. Correlation Matrices

- Definition: Standardized covariance matrix

$$R = \begin{bmatrix} 1 & r_{12} & \ldots & r_{1p} \\ r_{21} & 1 & \ldots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \ldots & 1 \end{bmatrix}$$

- Purpose: Compare relationships on the same scale [-1,1].

8. Linear Combinations of Variables

- Definition: Weighted sum of variables

$$Y = a_1X_1 + a_2X_2 + \cdots + a_pX_p$$

- Mean:

$$\mu_Y = \sum a_i \mu_i$$

- Variance:

$$Var(Y) = \mathbf{a}^T \Sigma \mathbf{a}$$

- Use: Principal Component Analysis, portfolio optimization, factor scores.

9. Measures of Overall Variability

$$\sum_{i=1}^{p} \sigma_i^2 = \text{trace}(\Sigma)$$

- Total variance:
- Generalized variance: |Σ| (determinant of covariance matrix)
- Interpretation: Overall "spread" in p-dimensional space.

---

10. Estimation of Missing Values

- Methods:
  1. Mean imputation: Replace missing with mean
  2. Regression imputation: Predict using other variables
  3. Multiple imputation: Generate several plausible values, combine results
  4. KNN imputation: Use closest neighbors in p-dimensional space

Notes: Imputation preserves sample size but may reduce variance.

---

11. Distance Between Vectors

- Euclidean distance:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$

- Mahalanobis distance: Accounts for variable correlations

$$D^2 = (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

- Use: Outlier detection, cluster analysis, classification.

II. Testing the Assumptions of Multivariate Analysis – Detailed Explanation

Multivariate analysis methods (like MANOVA, discriminant analysis, PCA) rely on certain assumptions to ensure valid results. Violating these assumptions can lead to misleading conclusions.

1. Assessing Individual Variables vs Variate

Check each variable for unusual patterns or distributional problems before combining into multivariate analysis.
How:

- Visual tools:
  - Histograms → check shape, skewness, kurtosis
  - Boxplots → detect outliers
  - Density plots → compare variable distributions
- Numerical summaries:
  - Mean, median, variance, skewness, kurtosis

- Multivariate methods assume reasonably "well-behaved" data for each variable
- Extreme outliers or highly skewed variables can distort covariance, correlations, and multivariate distances

Example: A dataset with one variable heavily skewed may bias the Mahalanobis distance and inflate Type I error in MANOVA.

2. Normality

Each variable (univariate) and the combination of variables (multivariate) should follow a normal distribution.

- Many multivariate methods assume normality of residuals or the underlying distribution to ensure:

    - Accurate significance testing
    - Valid estimation of confidence regions
    - Proper performance of linear discriminant functions

How:

Univariate normality:

1. Shapiro-Wilk test
    - Null hypothesis H0: Data is normally distributed
    - p-value > 0.05 → fail to reject → normal
2. Kolmogorov-Smirnov test
    - Compares sample distribution with theoretical normal
3. Q-Q plot (Quantile-Quantile plot)
    - Points follow the 45° line → approximately normal

Multivariate normality:

1. Mardia's test
    - Measures multivariate skewness and kurtosis
    - Skewness near 0, kurtosis near $p(p+2)$ → normal
2. Henze-Zirkler test
    - Global test for deviation from multivariate normality
3. Royston's test
    - Extension of Shapiro-Wilk to multiple variables

Notes:

- Minor deviations from normality are often acceptable
- Severe non-normality → consider transformations (log, square root)

3. Homoscedasticity (Equal Variances)

Assumes that the variances and covariances of variables are equal across groups.

- Ensures that group comparisons are valid
- Unequal covariances can inflate Type I or II errors in MANOVA and discriminant analysis

How:

1. Levene's test – tests equality of variances for each variable
    - $p > 0.05$ → equal variance assumed
2. Box's M test – tests equality of covariance matrices across groups

   - Sensitive to departures from normality
3. Visual inspection – residual plots for homogeneity

Notes:

- If assumption violated: consider robust methods or data transformations
- Standardizing variables can help if scale differences are the issue

4. Linearity

What: Assumes linear relationships between variables.

- Methods like PCA, MANOVA, and discriminant analysis rely on linear combinations of variables
- Non-linear relationships may distort covariance, eigenvectors, and discriminant functions

How:

- Scatterplot matrices: Look for straight-line patterns
- Partial regression plots: Show linearity of variable against others
- Residual plots: Non-random patterns indicate nonlinearity

What to do if violated:

- Apply non-linear transformations (log, square root, polynomial)
- Consider non-linear multivariate methods

III. The Multivariate Normal Distribution

1. Multivariate Normal Density Function

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$$

- Interpretation: Generalizes the bell curve to p dimensions

2. Generalized Population Variance

- Determinant of covariance matrix |Σ|
- Large |Σ| → more spread in multivariate space
- Small |Σ| → less dispersion

3. Diversity of Applications

- Finance: Portfolio risk assessment
- Psychology: Test batteries with multiple traits
- Marketing: Customer segmentation
- Biology: Multivariate trait analysis

4. Properties of Multivariate Normal Random Variables

1. Any linear combination of MVN is normal
2. Marginal distributions of MVN are univariate normal
3. Conditional distributions are also normal
4. Covariance matrix Σ must be positive semi-definite

5. Estimation in the Multivariate Normal

$$\hat{\boldsymbol{\mu}} = \bar{\mathbf{X}}$$

- Sample mean vector:
- Sample covariance matrix:

$$\hat{\Sigma} = \frac{1}{n-1} \sum (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})^T$$

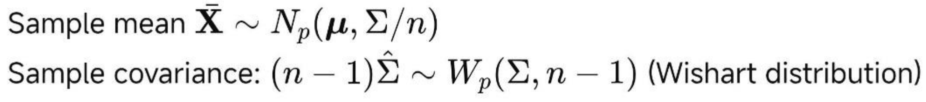
6. Maximum Likelihood Estimation

- Likelihood function:

$$L(\boldsymbol{\mu}, \Sigma) = \prod_{i=1}^{n} f_{\mathbf{X}_i}(\mathbf{x}_i)$$

- Solve for μ,Σ that maximize L

7. Distribution of Sample Mean and Covariance



Sample mean $\bar{\mathbf{X}} \sim N_p(\boldsymbol{\mu}, \Sigma/n)$

Sample covariance: $(n-1)\hat{\Sigma} \sim W_p(\Sigma, n-1)$ (Wishart distribution)

8. Assessing Multivariate Normality

- Graphical:
  - Q-Q plots of Mahalanobis distances vs Chi-square quantiles
  - Scatterplot matrix for linear relationships
- Statistical:
  - Mardia's skewness and kurtosis test
  - Henze-Zirkler test
  - Royston's test

References

1. Johnson, R. A., & Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis* (6th ed.).
2. Rencher, A. C., & Christensen, W. F. (2012). *Methods of Multivariate Analysis*.
3. Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate Analysis*.