VU LEARNING HUB

vulearninghub.xyz

HANDOUT

STA633

Multivariate Analysis

STA633 by VU Learning HUB

Prepared for VU students • Free study resources

https://vulearninghub.xyz

STA633

Multivariate Analysis

Analysis of Multivariate Data — Study Handout

This handout covers STA633 Multivariate Analysis: the extension of univariate statistics to vectors of variables — multivariate normal distribution, Hotelling's T^2, MANOVA, principal components, factor analysis, discriminant analysis and clustering. Matrix notation is used throughout: bold x is a p×1 vector, Σ (sigma) is the covariance matrix.

Notation: x■ = sample mean vector, S = sample covariance matrix, μ = population mean vector, Σ = population covariance matrix, |A| = determinant, tr(A) = trace. Work the small numerical examples by hand once — exams reward matrix-arithmetic fluency.

Table of Contents

Unit 1: Introduction: Multivariate Data and Matrix Tools (Lectures 1–5)
Unit 2: The Multivariate Normal Distribution (Lectures 6–9)
Unit 3: Hotelling's T^2: One-Sample and Two-Sample Tests (Lectures 10–14)
Unit 4: MANOVA: Comparing Several Mean Vectors (Lectures 15–18)
Unit 5: Principal Component Analysis (Lectures 19–23)
Unit 6: Factor Analysis (Lectures 24–27)
Unit 7: Discriminant Analysis and Classification (Lectures 28–31)
Unit 8: Cluster Analysis (Lectures 32–35)
Unit 9: Review: Choosing the Right Multivariate Method (Lectures 36–38)

Unit 1: Introduction: Multivariate Data and Matrix Tools

Lectures 1–5

Learning Objectives

- Explain why multivariate methods are needed
- Organize data in the n×p data matrix
- Compute the mean vector and covariance matrix
- Review essential matrix algebra: inverse, determinant, eigenvalues

Why Multivariate?

Real data comes in bundles: a patient's blood pressure, cholesterol and glucose; a customer's age, income and spending. Analyzing each variable separately **misses correlations** and inflates error rates (running 10 t-tests at 5% gives ~40% family-wise error). Multivariate methods analyze the **vector as a whole**, respecting how variables move together.

The Data Matrix and Summaries

Data matrix X (n×p): n observations (rows), p variables (columns). —

Mean vector x■ (p×1): the p variable means stacked. —

Sample covariance matrix S (p×p): S_jk = covariance of variables j and k; diagonal = variances. S is symmetric. —

Correlation matrix R: R_jk = S_jk / √(S_jj·S_kk); diagonal all 1s — scale-free, used when variables have different units.

Matrix Essentials

- **Determinant |A|:** for 2×2 [[a,b],[c,d]], |A| = ad − bc. |A| = 0 means singular (no inverse) — variables perfectly collinear.
- **Inverse:** A■¹A = I. For 2×2: (1/|A|)[[d,–b],[–c,a]].
- **Eigenvalues/eigenvectors:** solve |A − λI| = 0. Eigenvectors give the principal axes of the data cloud; eigenvalues their variances — the engine of PCA.
- **Trace:** sum of diagonal = sum of eigenvalues = total variance.

Worked: X = [[4,2],[2,4]] (covariance matrix). |X| = 16–4 = 12. Inverse = (1/12)[[4,–2],[–2,4]]. Eigenvalues: |X–λI| = (4–λ)² – 4 = λ² – 8λ + 12 = 0 → λ = 6, 2. Trace = 8 = 6+2 ✓. So total variance 8 splits into 6 along the first principal axis and 2 along the second.

Unit Summary

- Multivariate analysis treats correlated variables jointly; separate univariate tests miss correlation and inflate errors.
- Summarize with x■, S (covariances) and R (correlations).
- Determinant, inverse, eigenvalues/eigenvectors and trace are the working tools — PCA runs on eigen-decomposition.

vulearninghub.xyz | by VU Learning HUB

Exam Pointers

- ★ Compute S from a small data set and interpret one off-diagonal element.
- ★ Find eigenvalues, determinant and inverse of a 2×2 covariance matrix.
- ★ Why can't ten separate t-tests replace one multivariate test?

Unit 2: The Multivariate Normal Distribution

Lectures 6–9

Learning Objectives

- Write the multivariate normal density
- Explain the role of μ and Σ
- Describe properties: marginals, conditionals, linear combinations
- Use Mahalanobis distance

Density and Parameters

Density: f(x) = (2π)^(−p/2) |Σ|^(−1/2) exp(−½(x−μ)′Σ■¹(x−μ)). —

- μ **(p×1):** centers the distribution. Σ **(p×p):** shapes it — variances on the diagonal, covariances off-diagonal; must be positive definite.
- Contours of equal density are **ellipsoids** centered at μ, with axes along Σ's eigenvectors and lengths ∝ √eigenvalues.

Key Properties

- **Marginals** of a multivariate normal are normal; **conditionals** are normal with mean adjusted by regression on the known variables.
- **Linear combinations** a′x are univariate normal: mean a′μ, variance a′Σa.
- If Σ is diagonal (zero correlations), the variables are **independent** — for normal data, uncorrelated ■ independent.
- Sample x■ ~ N_p(μ, Σ/n); the scatter matrix (n−1)S ~ Wishart — the multivariate chi-square.

Mahalanobis Distance

D² = (x − μ)′Σ■¹(x − μ) — distance corrected for variances and correlations.

- Unlike Euclidean distance, it down-weights high-variance directions and accounts for correlation — the right metric for outlier detection and classification.

Worked: μ = (0,0), Σ = [[4,0],[0,1]]. Point (4, 1): Mahalanobis D² = 16/4 + 1/1 = 5. Euclidean distance² = 17. The point is 2 SDs out in x■ but only 1 in x■ — Mahalanobis correctly calls it less extreme than Euclidean suggests.

Unit Summary

- N_p(μ, Σ): μ centers, Σ shapes; density contours are ellipsoids.
- Marginals, conditionals and linear combinations stay normal; uncorrelated = independent.
- Mahalanobis D² = (x−μ)′Σ■¹(x−μ) is the correlation-aware distance.

Exam Pointers

★ Write the bivariate normal density and identify μ, Σ's roles.

★ Compute Mahalanobis distance for a given point and compare with Euclidean.

vulearninghub.xyz | by VU Learning HUB

★ State two properties that make the normal the default multivariate model.

Unit 3: Hotelling's T²: One-Sample and Two-Sample Tests

Lectures 10–14

Learning Objectives

- Construct Hotelling's T² for one sample
- Convert T² to an F statistic and test
- Extend to two independent samples
- Build simultaneous confidence intervals

One-Sample T²

T² = n(x■ – μ■)′S■⁻¹(x■ – μ■) — the multivariate analogue of the t-statistic squared.

Test: F = [(n–p) / (p(n–1))] × T² ~ F(p, n–p). Reject H■: μ = μ■ if F exceeds the critical value. —

- Intuition: squared Mahalanobis distance of x■ from μ■, scaled by n — big distance ■ reject.

Two-Sample T²

T² = [n■n■/(n■+n■)] (x■■ – x■■)′S_pooled■⁻¹(x■■ – x■■), where S_pooled = [(n■–1)S■ + (n■–1)S■]/(n■+n■–2). —

- Same F conversion with n = n■+n■. Assumes equal covariance matrices (test with Box's M).

Simultaneous Confidence Intervals

After rejecting, find **which** variables differ: for any linear combination a′μ, the interval a′x■ ± √[p(n–1)/(n–p) × F_α × a′Sa/n] holds **simultaneously** for all a at confidence 1–α. For single variables, set a to unit vectors — wider than ordinary t-intervals, honestly accounting for multiplicity.

Worked (p=2): n=20, x■=(5,8), μ■=(4,7), S=[[2,0.5],[0.5,3]]. d=(1,1). S■⁻¹ = (1/5.75)[[3,–0.5],[–0.5,2]]. d′S■⁻¹d = (1/5.75)(3–1+2) = 4/5.75 = 0.696. T² = 20×0.696 = 13.91. F = (18/(2×19))×13.91 = 6.59. F■.■■(2,18) = 3.55 → reject H■.

Unit Summary

- T² = n·(Mahalanobis distance of x■ from μ■)² — one test for the whole mean vector.
- Convert to F(p, n–p) for the p-value; two-sample version uses pooled S.
- Simultaneous CIs identify which variables drive rejection.

Exam Pointers

★ Compute one-sample T² and the F test from given x■, S, n, μ■.

★ Write the two-sample T² formula and state its assumption.

★ Why are simultaneous intervals wider than separate t-intervals?

Unit 4: MANOVA: Comparing Several Mean Vectors

Lectures 15–18

Learning Objectives

- Set up one-way MANOVA with H and E matrices
- Compute Wilks' lambda and test it
- Contrast MANOVA with separate ANOVAs
- Follow up significant results

H and E Matrices

One-way MANOVA with g groups: **H** (hypothesis SSCP) captures between-group variation; **E** (error SSCP) captures within-group variation — the multivariate analogues of SSB and SSW. H■: all group mean vectors equal.

Wilks' Lambda

Λ = |E| / |E + H|. Ranges 0→1; small Λ = groups differ a lot ■ reject. —

- Convert to F (Rao's approximation) or chi-square for large n: $-[n-1-(p+g)/2]\cdot \ln \Lambda \sim \chi^2(p(g-1))$.
- Alternatives: Pillai's trace, Hotelling–Lawley trace, Roy's largest root — Pillai is most robust to violations.

MANOVA vs Separate ANOVAs

- Separate ANOVAs inflate Type I error and ignore correlation between dependent variables.
- MANOVA can detect group differences in the **combination** of variables even when no single variable differs (the multivariate pattern).
- Assumptions: multivariate normality within groups, equal covariance matrices (Box's M test), independent observations.

Follow-up plan after significant MANOVA: 1) simultaneous CIs / discriminant analysis to find separating variables, 2) univariate ANOVAs with Bonferroni correction as a descriptive aid (not a replacement), 3) plot group means with confidence ellipses.

Unit Summary

- MANOVA: H (between) and E (within) SSCP matrices; Wilks' Λ = |E|/|E+H|, small = significant.
- Beats separate ANOVAs on error control and on detecting combined-variable effects.
- Assumptions: normality, equal covariances (Box's M), independence.

Exam Pointers

★ Define H and E and write Wilks' lambda.

★ Compute Λ from given |E|, |H+E| and interpret.

★ Give two reasons MANOVA is preferred over multiple ANOVAs.

Unit 5: Principal Component Analysis

Lectures 19–23

Learning Objectives

- Derive principal components from eigen-decomposition
- Decide how many components to retain
- Interpret loadings and scores
- Choose between covariance and correlation matrix

The Idea

PCA finds **new uncorrelated variables** (components) that are linear combinations of the originals, ordered by variance captured. PC1 = the direction of maximum variance (first eigenvector of S or R); PC2 = max variance orthogonal to PC1; and so on. If the first 2–3 PCs capture 80–90% of total variance, p variables collapse to 2–3 with little loss — the basis of dimension reduction and biplots.

Computation and Decisions

Solve |S − λI| = 0 → eigenvalues λ■ ≥ λ■ ≥ … ≥ λ_p; PC_i = e_i′x. Proportion of variance = λ_i / Σλ. —

- **How many to keep:** Kaiser's rule (λ > 1, for correlation matrix), scree plot elbow, cumulative 80–90% rule.
- **S or R?** Use R (correlation) when variables have different units/scales; S when units are commensurable and variance differences are meaningful.

Loadings and Scores

- **Loadings** (eigenvector elements, often scaled by √λ): correlation of variable with component — name the component from its high-loading variables.
- **Scores:** each observation's coordinates on the PCs — used for plots, outlier detection and as inputs to regression/clustering.

Worked: S = [[4,2],[2,4]] (from Unit 1): λ = 6, 2; PC1 explains 6/8 = 75%, PC2 25%. PC1 direction ∝ (1,1): the 'overall size' component; PC2 ∝ (1,–1): the contrast. With 75% in one component, a 2-D problem is nearly 1-D.

Unit Summary

- PCs = eigenvectors of S/R ordered by eigenvalue; each captures max remaining variance.
- Retain via Kaiser, scree elbow, or 80–90% cumulative variance.
- Loadings interpret components; scores locate observations; use R for mixed units.

Exam Pointers

★ Compute PCs, variance proportions for a 2×2 S.

★ State three retention rules.

★ When must PCA use the correlation matrix instead of covariance?

Unit 6: Factor Analysis

Lectures 24–27

Learning Objectives

- Distinguish factor analysis from PCA
- Write the orthogonal factor model
- Explain loadings, communalities and specific variances
- Describe rotation and factor interpretation

PCA vs Factor Analysis

PCA is a data-reduction trick: components are exact linear combos capturing max variance. **Factor analysis** is a **model**: observed variables are generated by a few unobserved (latent) factors plus unique noise. PCA explains variance; FA explains **covariance/correlation structure**. Use FA when you hypothesize latent traits (e.g., 'intelligence' behind test scores).

The Orthogonal Factor Model

$x = \mu + LF + \varepsilon$, where L (p×m) = loadings, F (m×1) = common factors, ε = specific factors. —

Covariance structure: $\Sigma = LL' + \Psi$, with Ψ diagonal (specific variances). —

- **Communality** h^2_j = sum of squared loadings of variable j = variance explained by common factors.
- **Specific variance** $\psi_j = 1 - h^2_j$ (using R) = unique + error variance.

Estimation and Rotation

- **Estimation:** principal-component method (eigen-decomposition of R) or maximum likelihood (assumes normality; gives goodness-of-fit test).
- **Rotation problem:** L is not unique — rotate (orthogonal: varimax; oblique: promax) to get simple structure where each variable loads highly on one factor only.
- Interpret factors from the rotated loading pattern; validate with the chi-square fit test (ML method).

Example: 6 exam scores → 2 factors: 'quantitative ability' (loads on math, physics) and 'verbal ability' (loads on Urdu, English), after varimax rotation. Communalities ~0.7 mean the two factors explain 70% of each test's variance; the rest is test-specific.

Unit Summary

- FA models $x = \mu + LF + \varepsilon$: latent factors explain correlations; $\Sigma = LL' + \Psi$.
- Communality = variance explained by factors; specific variance = the rest.
- Rotate (varimax) for interpretability; ML estimation adds a fit test.

Exam Pointers

★ Contrast PCA and factor analysis (purpose and model).

- ★ Write the factor model and define communality.
- ★ Why rotate? What is simple structure?

Unit 7: Discriminant Analysis and Classification

Lectures 28–31

Learning Objectives

- Build Fisher's linear discriminant function
- Classify observations via Mahalanobis distance
- Extend to multiple groups (canonical discriminants)
- Evaluate classification with confusion matrices

Fisher's Linear Discriminant (Two Groups)

Find the linear combination **y = a′x** that **maximizes between-group variance relative to within-group variance**: a = S_pooled■¹(x■■ – x■■). Project all data onto this line; classify by which group mean is nearer. This is the direction that best separates the groups — closely related to the two-sample T^2 and to MANOVA.

Classification Rule

Allocate x to group 1 if (x■■–x■■)′S_pooled■¹(x – (x■■+x■■)/2) > ln[(c■■p■)/(c■■p■)]-ish threshold; with equal costs and priors, the midpoint rule: nearer Mahalanobis distance wins. —

- With unequal priors or misclassification costs, shift the cutoff toward the cheaper/more likely group.
- **Quadratic discriminant analysis (QDA):** allow each group its own covariance — curved boundaries; needs more data.

Multiple Groups and Evaluation

- **Canonical discriminant functions:** up to min(g–1, p) uncorrelated directions separating g groups; plot the first two for a visual map.
- **Confusion matrix:** cross-tabulate true vs predicted group; hit rate = diagonal/total. Validate honestly: split-sample or **leave-one-out cross-validation** (resubstitution flatters).

Example: bank loan applicants (income, debt ratio) classified good/bad. Fisher's function weights income positively, debt ratio negatively; cutoff set from historical defaults. Confusion matrix on holdout data: 85% correct — but check the costly error: misclassifying a defaulter as good costs far more than the reverse, so adjust the threshold.

Unit Summary

- Fisher's rule: project on a = S_pooled■¹(x■■–x■■); classify by nearest (cost/prior-adjusted) center.
- QDA for unequal covariances; canonical functions for g > 2 groups.
- Report cross-validated confusion matrices, and weight errors by cost.

Exam Pointers

★ Derive/interpret Fisher's linear discriminant direction.

vulearninghub.xyz | by VU Learning HUB

- ★ Classify a new observation by Mahalanobis distance (numbers given).
- ★ Why is resubstitution accuracy optimistic? Name two honest alternatives.

Unit 8: Cluster Analysis

Lectures 32–35

Learning Objectives

- Distinguish hierarchical and non-hierarchical clustering
- Apply linkage methods and read a dendrogram
- Run k-means logic and choose k
- Validate and profile clusters

The Clustering Task

Unlike discriminant analysis, clustering is **unsupervised**: no group labels exist — we discover them. Steps: choose a **distance** (usually Euclidean on standardized variables, or Mahalanobis), choose an **algorithm**, decide **how many clusters**, then **profile** them (what makes each cluster distinctive?).

Hierarchical Clustering

- **Agglomerative:** start with n clusters, repeatedly merge the closest pair; draw the **dendrogram** and cut it at a chosen height.
- **Linkage rules** define cluster distance: single (nearest neighbors — can chain), complete (farthest — compact balls), average, Ward's (minimizes within-cluster variance increase — often best).

K-Means

Fix k; assign each point to the nearest centroid; recompute centroids; repeat until stable. Fast for large n. Weaknesses: must pre-specify k, sensitive to starting seeds and outliers, assumes spherical clusters. Choose k via the **elbow** in within-cluster sum of squares, silhouette scores, or practical interpretability.

- **Standardize first:** variables with big units dominate distances otherwise.

Example: segmenting 500 bank customers on (balance, transactions/month, loan usage). Standardize → Ward's hierarchical suggests 4 clusters → k-means refines → profiles: 'dormant savers', 'active transactors', 'borrowers', 'premium low-activity'. Marketing then targets each differently — the business payoff of clustering.

Unit Summary

- Clustering is unsupervised group discovery: distance + algorithm + k + profiling.
- Hierarchical: merge stepwise, read the dendrogram; linkage choice shapes clusters (Ward's often best).
- K-means: iterate assign/update; pick k by elbow/silhouette; always standardize first.

Exam Pointers

★ Contrast hierarchical and k-means clustering (4 points).

- ★ Explain single vs complete vs Ward's linkage.
- ★ Why standardize variables before clustering?

Unit 9: Review: Choosing the Right Multivariate Method

Lectures 36–38

Learning Objectives

- Map research questions to methods
- State the assumptions checklist for each method
- Avoid common multivariate pitfalls
- Plan a complete multivariate analysis

Method Selection Map

Compare one mean vector to a standard → Hotelling's one-sample T^2. —

Compare two groups' mean vectors → two-sample T^2. —

Compare 3+ groups' mean vectors → MANOVA (+ discriminant follow-up). —

Reduce dimensions / find latent structure → PCA (reduction) or factor analysis (latent model). —

Classify into known groups → discriminant analysis (QDA if covariances differ). —

Discover unknown groups → cluster analysis. —

Assumption Checklist

- Multivariate normality (check marginals + Mahalanobis D^2 chi-square plot).
- Equal covariance matrices for T^2/MANOVA/LDA (Box's M).
- No extreme multicollinearity (singular S breaks inverses) — drop or combine redundant variables.
- n comfortably exceeds p (rule of thumb: $n > 20p$ for stable covariance estimates).

Pitfalls and Workflow

- Running many univariate tests instead of one multivariate test (error inflation).
- Interpreting PCs/factors without rotation or validation.
- Reporting resubstitution accuracy for classifiers.
- Forgetting to standardize before PCA-on-R decisions, clustering or regularized methods.

Sensible workflow: 1) clean and screen data (outliers via D^2), 2) check assumptions, 3) pick method by question type, 4) fit, 5) validate (cross-validation, fit tests), 6) interpret in subject-matter terms.

Capstone mini-case: A bank has 8 financial ratios for 300 firms, half defaulted. Task: build an early-warning model. Plan: screen outliers (D^2), check normality/equal covariances, run LDA (or QDA if Box's M rejects), validate with leave-one-out, report the confusion matrix with cost-weighted errors, and interpret the discriminant weights as the risk factors. Every method from this course appears in one pipeline.

Unit Summary

- Match the question to the method: T^2 (1–2 groups), MANOVA (3+), PCA/FA (structure), LDA (classification), clustering (discovery).
- Check normality, equal covariances, no singularity, adequate n.
- Workflow: screen → assume-check → fit → validate → interpret.

Exam Pointers

★ Given five research scenarios, name the method for each.

★ List the assumption checks before running MANOVA.

★ Outline a full analysis plan for the bank default-prediction case.

Key Terms Glossary

All key terms from this handout in one place — use for quick revision.

Data matrix X (n×p): n observations (rows), p variables (columns). —

Mean vector x■ (p×1): the p variable means stacked. —

Sample covariance matrix S (p×p): S_jk = covariance of variables j and k; diagonal = variances. S is symmetric. —

Correlation matrix R: R_jk = S_jk / √(S_jj·S_kk); diagonal all 1s — scale-free, used when variables have different units.

Density: f(x) = (2π)^(−p/2) |Σ|^(−1/2) exp(−½(x−μ)′Σ■¹(x−μ)). —

D² = (x − μ**)′Σ■¹(x** − μ**)** — distance corrected for variances and correlations.

T² = n(x■ − μ**■)′S■¹(x■** − μ**■)** — the multivariate analogue of the t-statistic squared.

Test: F = [(n−p) / (p(n−1))] × T² ~ F(p, n−p). Reject H■: μ = μ■ if F exceeds the critical value. —

T² = [n■n■/(n■+n■)] (x■■ − x■■)′S_pooled■¹(x■■ − x■■), where S_pooled = [(n■−1)S■ + (n■−1)S■]/(n■+n■−2). —

Λ = |E| / |E + H|. Ranges 0→1; small Λ = groups differ a lot ■ reject. —

Solve |S − λ**I| = 0 → eigenvalues λ■ ≥ λ■ ≥ … ≥ λ_p; PC_i = e_i′x. Proportion of variance = λ_i / Σλ.** —

x = μ + LF + ε, where L (p×m) = loadings, F (m×1) = common factors, ε = specific factors. —

Covariance structure: Σ = LL′ + Ψ, with Ψ diagonal (specific variances). —

Allocate x to group 1 if (x■■−x■■)′S_pooled■¹(x − **(x■■+x■■)/2) > ln[(c■■p■)/(c■■p■)]-ish threshold; with equal costs and priors, the midpoint rule: nearer Mahalanobis distance wins.** —

Compare one mean vector to a standard → Hotelling's one-sample T². —

Compare two groups' mean vectors → two-sample T². —

Compare 3+ groups' mean vectors → MANOVA (+ discriminant follow-up). —

Reduce dimensions / find latent structure → PCA (reduction) or factor analysis (latent model). —

Classify into known groups → discriminant analysis (QDA if covariances differ). —

Discover unknown groups → cluster analysis. —