Mathematical Foundations

Dr. Juliho Castillo

Tecnológico de Monterrey

2025-09-29

Table of contents

Índice

Appendix: Mathematical Foundations

Matrix Algebra Foundations

Essential Matrix Operations for Multivariate Analysis

Matrix Transpose Properties:

- $(\boldsymbol{A}^\top)^\top = \boldsymbol{A}$
- $(\boldsymbol{A} + \boldsymbol{B})^\top = \boldsymbol{A}^\top + \boldsymbol{B}^\top$
- $(\boldsymbol{AB})^\top = \boldsymbol{B}^\top \boldsymbol{A}^\top$
- $(c\boldsymbol{A})^\top = c\boldsymbol{A}^\top$ for scalar c

Matrix Multiplication Rules:

- $(\boldsymbol{AB})\boldsymbol{C} = \boldsymbol{A}(\boldsymbol{BC})$ (associativity)

- $\boldsymbol{A}(\boldsymbol{B}+\boldsymbol{C}) = \boldsymbol{AB}+\boldsymbol{AC}$ (distributivity)
- Generally: $\boldsymbol{AB} \neq \boldsymbol{BA}$ (not commutative)

Trace Properties:

- $\mathrm{tr}(\boldsymbol{A}) = \sum_{i=1}^{n} a_{ii}$ (sum of diagonal elements)
- $\mathrm{tr}(\boldsymbol{A}+\boldsymbol{B}) = \mathrm{tr}(\boldsymbol{A}) + \mathrm{tr}(\boldsymbol{B})$
- $\mathrm{tr}(\boldsymbol{AB}) = \mathrm{tr}(\boldsymbol{BA})$ (cyclic property)
- $\mathrm{tr}(c\boldsymbol{A}) = c\ \mathrm{tr}(\boldsymbol{A})$ for scalar c

Important for PCA/FA: These properties ensure that variance decompositions are mathematically consistent.

Eigenvalues and Eigenvectors: Deep Dive

Definition: For square matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, scalar λ and vector $\boldsymbol{v} \neq \mathbf{0}$ satisfy: $\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}$

Characteristic Equation: $\det(\boldsymbol{A} - \lambda\boldsymbol{I}) = 0$

Fundamental Properties:

- $n \times n$ matrix has exactly n eigenvalues (counting multiplicities)
- Eigenvalues can be real or complex conjugate pairs
- Sum of eigenvalues equals trace: $\sum_{i=1}^{n} \lambda_i = \mathrm{tr}(\boldsymbol{A})$
- Product of eigenvalues equals determinant: $\prod_{i=1}^{n} \lambda_i = \det(\boldsymbol{A})$

Geometric Interpretation:

- Eigenvector: direction unchanged by transformation
- Eigenvalue: scaling factor in that direction
- Eigendecomposition reveals the «natural axes» of the transformation

For Symmetric Matrices (Covariance/Correlation):

- All eigenvalues are real
- Eigenvectors are orthogonal: $\boldsymbol{v}_i^\top \boldsymbol{v}_j = 0$ for $i \neq j$
- Can be orthonormalized: $\boldsymbol{v}_i^\top \boldsymbol{v}_i = 1$

Spectral Decomposition Theorem

For Symmetric Matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$:

Theorem: Every real symmetric matrix can be diagonalized by an orthogonal matrix: $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$

where:

- $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, ..., \lambda_n)$ (eigenvalues on diagonal)
- $\boldsymbol{Q} = [\boldsymbol{q}_1 \mid \boldsymbol{q}_2 \mid ... \mid \boldsymbol{q}_n]$ (orthonormal eigenvectors)
- $\boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{I}$ (orthogonal matrix)

Equivalent Outer Product Form: $\boldsymbol{A} = \sum_{i=1}^{n} \lambda_i \boldsymbol{q}_i \boldsymbol{q}_i^\top$

Significance for Multivariate Statistics:

- Covariance matrices are symmetric $\rightarrow$ always diagonalizable
- Eigenvalues represent variance in principal directions
- Eigenvectors define the principal directions
- Enables dimension reduction by truncating small eigenvalues
- Foundation for both PCA and Factor Analysis

Positive Semidefinite Property: For covariance matrices: $\lambda_i \geq 0$ for all i (variances cannot be negative)

Matrix Norms and Distance Measures

Frobenius Norm (Essential for Factor Analysis): $\|\boldsymbol{A}\|_F = \sqrt{\mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{A})} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij}^2}$

Properties:

- $\|\boldsymbol{A}\|_F^2 = \sum_{i=1}^{n} \sigma_i^2$ where σ_i are singular values
- $\|\boldsymbol{A}\boldsymbol{B}\|_F \leq \|\boldsymbol{A}\|_F \|\boldsymbol{B}\|_2$ (submultiplicativity)
- $\|\boldsymbol{A} + \boldsymbol{B}\|_F \leq \|\boldsymbol{A}\|_F + \|\boldsymbol{B}\|_F$ (triangle inequality)

Spectral Norm (2-norm): $\|\boldsymbol{A}\|_2 = \sigma_{\max(\boldsymbol{A})} = \sqrt{\lambda_{\max(\boldsymbol{A}^\top \boldsymbol{A})}}$

Mahalanobis Distance: $d^2(\boldsymbol{x}, \boldsymbol{\mu}) = (\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})$

Applications in Multivariate Analysis:

- Frobenius norm: measuring fit in factor analysis
- Spectral norm: condition n umber analysis
- Mahalanobis distance: outlier detection, multivariate normality tests
- All preserve relationships under orthogonal transformations

Matrix Calculus for Optimization

Scalar Functions of Matrices:

Trace Derivatives:

- $\frac{\partial}{\partial \boldsymbol{A}} \operatorname{tr}(\boldsymbol{A}) = \boldsymbol{I}$
- $\frac{\partial}{\partial \boldsymbol{A}} \operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{B}) = \boldsymbol{B}$
- $\frac{\partial}{\partial \boldsymbol{A}} \operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{A}) = 2\boldsymbol{A}$

Quadratic Form Derivatives:

- $\frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x} = (\boldsymbol{A} + \boldsymbol{A}^\top)\boldsymbol{x}$
- For symmetric $\boldsymbol{A}$: $\frac{\partial}{\partial \boldsymbol{x}} \boldsymbol{x}^\top \boldsymbol{A} \boldsymbol{x} = 2\boldsymbol{A}\boldsymbol{x}$

Determinant and Inverse:

- $\frac{\partial}{\partial \boldsymbol{A}} \log\det(\boldsymbol{A}) = \left(\boldsymbol{A}^{-1}\right)^\top$
- $\frac{\partial}{\partial \boldsymbol{A}} \boldsymbol{A}^{-1} = -\boldsymbol{A}^{-1} \frac{\partial \boldsymbol{A}}{\partial \boldsymbol{A}} \boldsymbol{A}^{-1}$

Applications:

- PCA: Maximizing variance subject to orthogonality constraints
- Factor Analysis: Maximum likelihood estimation
- Lagrange multipliers for constrained optimization
- Newton-Raphson methods for iterative algorithms

Singular Value Decomposition (SVD)

Universal Matrix Decomposition: For any matrix $\boldsymbol{A} \in \mathbb{R}^{m \times n}$: $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$

where:

- $\boldsymbol{U} \in \mathbb{R}^{m \times m}$: left singular vectors (orthogonal)
- $\boldsymbol{\Sigma} \in \mathbb{R}^{m \times n}$: diagonal matrix of singular values $\sigma_i \geq 0$
- $\boldsymbol{V} \in \mathbb{R}^{n \times n}$: right singular vectors (orthogonal)

Relationship to Eigendecomposition:

- $\boldsymbol{A}^\top \boldsymbol{A} = \boldsymbol{V}\boldsymbol{\Sigma}^\top \boldsymbol{\Sigma}\boldsymbol{V}^\top$ (eigendecomposition)

- $\boldsymbol{A}\boldsymbol{A}^\top = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^\top\boldsymbol{U}^\top$ (eigendecomposition)
- Singular values: $\sigma_i = \sqrt{\lambda_i(\boldsymbol{A}^\top\boldsymbol{A})}$

Properties:

- Rank of $\boldsymbol{A}$ = n umber of non-zero singular values
- $\|\boldsymbol{A}\|_2 = \sigma_1$ (largest singular value)
- $\|\boldsymbol{A}\|_F = \sqrt{\sum_{i=1}^r \sigma_i^2}$ where $r = \mathrm{rank}(\boldsymbol{A})$

Applications in Multivariate Statistics:

- Principal Component Analysis (PCA of data matrix)
- Pseudoinverse computation: $\boldsymbol{A}^+ = \boldsymbol{V}\boldsymbol{\Sigma}^+\boldsymbol{U}^\top$
- Low-rank matrix approximation
- Numerical stability in factor analysis algorithms

Statistical Foundations

Multivariate Normal Distribution

Definition: Random vector $\boldsymbol{X} \in \mathbb{R}^p$ follows multivariate normal distribution: $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$

Probability Density Function: $f(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{p}{2}} \, |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\!\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$

Parameters:

- $\boldsymbol{\mu} \in \mathbb{R}^p$: mean vector
- $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$: covariance matrix (positive definite)

Key Properties:

- Linear combinations remain normal: $\boldsymbol{a}^\top \boldsymbol{X} \sim N(\boldsymbol{a}^\top \boldsymbol{\mu}, \boldsymbol{a}^\top \boldsymbol{\Sigma} \boldsymbol{a})$
- Marginal distributions are normal: $X_i \sim N(\mu_i, \sigma_{ii})$
- Conditional distributions are normal (given subset of variables)

Standardization: If $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $(\boldsymbol{X} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{X} - \boldsymbol{\mu}) \sim \chi^2_p$

Importance for Factor Analysis:

- Maximum likelihood estimation assumes multivariate normality
- Goodness-of-fit tests based on chi-square distribution
- Confidence intervals and hypothesis tests for factor loadings

Sample Statistics and Their Properties

Sample Mean Vector: $\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i = \frac{1}{n}\boldsymbol{X}^\top \mathbf{1}_n$

Properties:

- $E[\overline{\boldsymbol{x}}] = \boldsymbol{\mu}$ (unbiased)
- $\mathrm{Cov}(\overline{\boldsymbol{x}}) = \frac{1}{n}\boldsymbol{\Sigma}$ (scales with sample size)
- $\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}) \to N(\mathbf{0}, \boldsymbol{\Sigma})$ (asymptotic normality)

Sample Covariance Matrix: $S = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top$

Matrix Form: $S = \frac{1}{n-1}\left(\boldsymbol{X} - \mathbf{1}_n\overline{\boldsymbol{x}}^\top\right)^\top\left(\boldsymbol{X} - \mathbf{1}_n\overline{\boldsymbol{x}}^\top\right)$

Properties:

- $E[\boldsymbol{S}] = \boldsymbol{\Sigma}$ (unbiased)
- $(n-1)\boldsymbol{S} \sim W_{p(n-1,\boldsymbol{\Sigma})}$ (Wishart distribution)
- Positive semidefinite with probability 1 if $n \geq p$

Sample Correlation Matrix: $\boldsymbol{R} = \boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{S}\boldsymbol{D}^{-\frac{1}{2}}$ where $\boldsymbol{D} = \mathrm{diag}\left(s_{\{11\}}, s_{\{22\}}, \ldots, s_{\{pp\}}\right)$

Maximum Likelihood Estimation

General Principle: Find parameters that maximize the likelihood of observing the sample data.

For Multivariate Normal Data: $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{\frac{p}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\right)$

Log-Likelihood: $\ell(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{np}{2}\log(2\pi) - \frac{n}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})$

ML Estimators:

- $\hat{\boldsymbol{\mu}} = \overline{\boldsymbol{x}}$ (sample mean)
- $\hat{\boldsymbol{\Sigma}} = \texttt{frac}\{1\}\{n\} \sum_{i=1}^{n} (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\top$ (ML covariance)

Properties:

- Consistent: $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$ as $n \to \infty$
- Asymptotically normal: $\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right) \to N(\mathbf{0}, \boldsymbol{I}^{-1}(\boldsymbol{\theta}))$
- Asymptotically efficient: achieves Cramér-Rao lower bound

Application to Factor Analysis:

- Estimate factor loadings $\boldsymbol{\Lambda}$ and unique variances $\boldsymbol{\Psi}$
- Iterative algorithms (EM algorithm)
- Model comparison via likelihood ratio tests

Hypothesis Testing in Multivariate Analysis

Likelihood Ratio Test Principle: $\Lambda = \texttt{frac}\left\{L\left(\hat{\boldsymbol{\theta}}_0\right)\right\}\left\{L\left(\hat{\boldsymbol{\theta}}\right)\right\} = \texttt{frac}\{\text{max likelihood under } H_0\}\{\text{max likelihood unrestricted}\}$

Test Statistic: $-2\log\Lambda \sim \chi^2_n u$ (asymptotically) where nu = difference in n umber of parameters

Bartlett's Test of Sphericity:

- H_0: $\boldsymbol{\Sigma} = \sigma^2\boldsymbol{I}$ (variables uncorrelated, equal variances)

- Test statistic: $\chi^2 = -(n - 1 - \texttt{frac}\{2p + 5\}\{6\}) \log|\boldsymbol{R}|$
- Degrees of freedom: $\texttt{frac}\{p(p - 1)\}\{2\}$

Goodness-of-Fit in Factor Analysis:

- H_0: $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}^\top + \mathbf{\Psi}$ (model fits)
- Test statistic: $\chi^2 = (n - 1)\left[\log|\hat{\mathbf{\Sigma}}| - \log|\boldsymbol{S}| + \operatorname{tr}\left(\boldsymbol{S}\hat{\mathbf{\Sigma}}^{-1}\right) - p\right]$
- Degrees of freedom: $\texttt{frac}\{p(p - 1)\}\{2\} - pk - p$ where k = n umber of factors

Multiple Testing Considerations:

- Bonferroni correction for multiple comparisons
- False Discovery Rate (FDR) control
- Family-wise error rate control

Asymptotic Theory and Large Sample Properties

Central Limit Theorem for Multivariate Data: If $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_n$ are i.i.d. with $E[\boldsymbol{X}_i] = \boldsymbol{\mu}$ and $\mathrm{Cov}(\boldsymbol{X}_i) = \boldsymbol{\Sigma}$: $\sqrt{n}\left(\overline{\boldsymbol{X}} - \boldsymbol{\mu}\right) \xrightarrow{d} N(\mathbf{0}, \boldsymbol{\Sigma})$

Delta Method: For differentiable function g: $\sqrt{n}\left(g\left(\overline{\boldsymbol{X}}\right) - g(\boldsymbol{\mu})\right) \xrightarrow{d} N\left(\mathbf{0}, \boldsymbol{D}\Sigma\boldsymbol{D}^{\top}\right)$ where $\boldsymbol{D} = \nabla g(\boldsymbol{\mu})$ (gradient matrix)

Slutsky's Theorem: If $\boldsymbol{X}_n \xrightarrow{d} \boldsymbol{X}$ and $\boldsymbol{Y}_n \xrightarrow{p} \boldsymbol{c}$:

- $\boldsymbol{X}_n + \boldsymbol{Y}_n \xrightarrow{d} \boldsymbol{X} + \boldsymbol{c}$

- $\boldsymbol{Y}_n^\top \boldsymbol{X}_n \xrightarrow{d} \boldsymbol{c}^\top \boldsymbol{X}$

Contin uous Mapping Theorem: If $\boldsymbol{X}_n \xrightarrow{d} \boldsymbol{X}$ and g is contin uous: $g(\boldsymbol{X}\}_{\boldsymbol{n}} \xrightarrow{d} g(\boldsymbol{X}\}$

Applications:

- Confidence intervals for eigenvalues and eigenvectors
- Standard errors for factor loadings
- Bootstrap methods for complex statistics
- Robustness analysis under model misspecification

Information Theory and Model Selection

Kullback-Leibler Divergence: $D_{\mathrm{KL}(P \parallel Q)} = \int p(\boldsymbol{x}) \log \mathtt{frac}\{p(\boldsymbol{x})\}\{q(\boldsymbol{x})\} d\boldsymbol{x}$

Measures «distance» between probability distributions

- $D_{\mathrm{KL}(P \parallel Q)} \geq 0$ with equality if and only if $P = Q$
- Not symmetric: $D_{\mathrm{KL}(P \parallel Q)} \neq D_{\mathrm{KL}(Q \parallel P)}$

Akaike Information Criterion (AIC): $\mathrm{AIC} = -2\ell\left(\hat{\boldsymbol{\theta}}\right) + 2k$ where k = n umber of parameters

Bayesian Information Criterion (BIC): $\mathrm{BIC} = -2\ell\left(\hat{\boldsymbol{\theta}}\right) + k \log(n)$

Model Selection Strategy:

1. Fit multiple models with different n umbers of factors
2. Compare AIC/BIC values
3. Choose model with lowest criterion value
4. Validate with cross-validation or hold-out data

Factor Analysis Application:

- Compare models with $k = 1, 2, 3, ...$ factors
- Balance model fit vs. complexity
- BIC typically favors more parsimonious models than AIC
- Consider interpretability alongside statistical criteria

Information Matrix: $\boldsymbol{I}(\boldsymbol{\theta}) = -E\left[\texttt{frac}\{\partial^2 \ell\}\{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top\}\right]$

- Provides asymptotic covariance: $\mathrm{Cov}\left(\hat{\boldsymbol{\theta}}\right) \approx \boldsymbol{I}^{-1}(\boldsymbol{\theta})$

- Used for standard errors and confidence intervals

Computational Considerations and Numerical Stability

Condition Number: $\kappa(\boldsymbol{A}) = \texttt{frac}\{\sigma_{\max}\}\{\sigma_{\min}\}$ (ratio of largest to smallest singular value)

Numerical Issues:

- $\kappa(\boldsymbol{A}) > 10^{\{12\}}$ indicates near-singularity
- Small eigenvalues lead to unstable inverse calculations
- Rounding errors accumulate in iterative algorithms

Stable Algorithms:

1. **QR Decomposition** instead of normal equations: $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$
2. **SVD** for pseudoinverses: $\boldsymbol{A}^{+} = \boldsymbol{V}\boldsymbol{\Sigma}^{+}\boldsymbol{U}^{\top}$
3. **Cholsky Decomposition** for positive definite matrices: $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{L}^{\top}$

Regularization Techniques:

- Ridge regularization: $\boldsymbol{A} + \lambda \boldsymbol{I}$ for small $\lambda > 0$
- Shrinkage estimators for covariance matrices
- Robust correlation estimation (e.g., Spearman rank correlation)

Convergence Criteria:

- Relative change in parameters: $\texttt{frac}\left\{\|\boldsymbol{\theta}_{\{k+1\}} - \boldsymbol{\theta}_k\|\right\}\{\|\boldsymbol{\theta}_k\|\} < \varepsilon$
- Change in log-likelihood: $|\ell_{\{k+1\}} - \ell_k| < \varepsilon$
- Gradient norm: $\|\nabla \ell\| < \varepsilon$

Practical Guidelines:

- Use double precision arithmetic
- Monitor condition n umbers
- Check convergence from multiple starting points
- Validate results with alternative algorithms

Summary: Mathematical Foundations for Factor Analysis

Essential Matrix Theory:

- Spectral decomposition enables eigenvalue methods
- SVD provides n umerical stability and generality
- Matrix norms quantify approximation quality

Statistical Theory:

- Multivariate normality justifies ML estimation
- Asymptotic theory provides inference framework

- Information criteria guide model selection

Computational Aspects:

- Numerical stability requires careful algorithm design
- Condition n umbers indicate potential problems
- Regularization prevents overfitting and instability

Integration in Practice:

- Theory guides algorithm development
- Computation enables practical applications
- Statistics provides inference and validation

Key Insight: The mathematical foundations ensure that Factor Analysis results are not just empirically useful, but theoretically sound and computationally reliable. Understanding these foundations helps practitioners make informed decisions about model specification, estimation methods, and result interpretation.

For Further Study:

- Matrix Analysis (Horn & Johnson)
- Multivariate Statistical Analysis (Anderson)
- Numerical Linear Algebra (Trefethen & Bau)
- Factor Analysis (Harman)