Milestone 1: Random Variables and Covariance Theory

Due: Should be finished by 2/11/2025

Overview

In this milestone, you will develop the theoretical foundations for understanding covariance matrices and their estimation from data. This work builds on our coverage of random vectors and their statistical properties.

Objectives

1. Derive the sample covariance estimator from first principles
2. Analyze how variable scaling affects covariance estimation
3. Understand the relationship between variance and feature importance

Required Deliverables

1. Theoretical Development (30%)

- Derive the sample covariance estimator starting from random variable principles:
  - Start with definition of covariance for two random variables
  - Extend to vector case using outer product notation
  - Show how sample estimates replace expectations
  - Prove why (n-1) denominator gives unbiased estimate
- Show why centering (mean subtraction) is necessary:
  - Consider what happens without centering
  - Use matrix algebra to show effect on eigenstructure
  - Connect to fundamental theorem through row space
- Analyze the statistical properties of your estimator:
  - Derive expected value to show unbiasedness
  - Consider consistency as n $\rightarrow \infty$

– Analyze effect of sample size on rank

2. Variable Importance Analysis (30%)

- Explain how variable scales affect covariance estimation:
  - Write out covariance matrix with explicit scales
  - Show how scaling matrix acts on left and right
  - Connect to similarity transformations
  - Demonstrate effect on eigenvalues
- Demonstrate how variable magnitude impacts principal directions:
  - Use perturbation theory for small scale changes
  - Show how large scale differences dominate directions
  - Connect to condition number of covariance matrix
- Connect variance to feature relevance:
  - Derive variance explained along any direction
  - Show optimality of principal directions
  - Relate to statistical signal-to-noise ratio

3. Documentation (20%)

- Clear mathematical notation and derivations:
  - Use consistent notation aligned with course conventions
  - Show intermediate steps in matrix manipulations
  - Connect to fundamental theorem notation:
    * Row space: $\mathcal{R}(A)$ and its role in variance
    * Column space: $\mathcal{C}(A)$ and sample projections
    * Null space: $\mathcal{N}(A)$ and invariant directions
    * Left null space: $\mathcal{N}(A^T)$ and error analysis
- Well-organized explanations:
  - Start with population quantities
  - Move to sample estimates
  - Show connections to linear transformations
  - Explain practical implications
- Professional presentation:
  - LaTeX for mathematical expressions
  - Clear section organization

   – Visual aids for geometric interpretations
   – Code snippets demonstrating key concepts

Evaluation Criteria

- Mathematical correctness and rigor
- Clarity of explanations
- Depth of analysis
- Quality of presentation

Resources

- Lessons 9-10 content
- Course notes on random vectors and covariance
- Recommended readings on estimation theory

Submission Guidelines

1. Submit a PDF document with your derivations and analysis
2. Include clear mathematical notation
3. Provide explanations for each step
4. Add relevant visualizations where helpful

Tips for Success

- Start with the basic definition of covariance
- Consider why we use sample estimates
- Think about the role of scaling in real data
- Connect theory to practical applications