High Dimensional Analysis: *Analysis of Golub's Leukemia Dataset Using Regularization, Dimensionality Reduction, and Classification Techniques*

Research Proposal

Cindy Tieu ; Jonatan Blank
Lund School of Economics and Management

Background and Motivation

The Golub's Leukemia dataset presents a high-dimensional problem, with 7,129 **gene expression features** and only 72 samples classified into **two groups**: acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML).

This structure is ideal for testing feature selection and dimensionality reduction methods, which address overfitting and improve interpretability. Regularization techniques like Lasso and Ridge regression, combined with dimensional reduction (PCA), provide robust methodologies for classification and exploratory analysis of such datasets.

Dataset can be found at: Golub Data

Objectives

1. Perform an **exploratory analysis** of the dataset, understanding its structure and identifying key patterns.
2. Apply and compare **classification techniques** using:

- Best subset selection.
- Lasso regression.
- Ridge regression.
- Dimensional reduction (PCA) with classification.
- Log-likelihood.

3. Evaluate the **performance** of each methodology in terms of prediction accuracy, feature interpretability, and computational efficiency.
4. Statistical inference (if we do bootstrap): p-values etc

Methodology

The methodology involves analyzing the Golub's Leukemia dataset to classify leukemia types (ALL vs. AML) through feature reduction, logistic regression, and robust evaluation. First, the dataset is explored to summarize its structure, assess feature distributions, and visualize patterns (e.g., using PCA or heatmaps). Next, four feature reduction methods are applied:

1. **Best Subset Selection** identifies the optimal subset of features based on criteria like AIC or BIC.
2. **Lasso Regression** selects key variables by shrinking irrelevant ones to zero using L1-regularization.
3. **Ridge Regression** stabilizes coefficients through L2-regularization without eliminating features.
4. **PCA** reduces dimensions by transforming data into orthogonal components that explains the variance.

Logistic regression models are then fitted using features or components from each method. These results will be reported and compared.

To ensure robustness, bootstrapping (1,000 resamples) and train-test splits are used to evaluate the stability of feature selection and parameter distributions, focusing on Lasso and Ridge. Results are analyzed to compare feature reduction, classification performance, and parameter stability, providing insights into trade-offs between interpretability, predictive power, and robustness.

Additional Notes

- Best-subset selection modification - Source: Best subset binary prediction - ScienceDirect