

Statistical Inference for High-Dimensional Models with Applications to Imaging Genetics

SARA SHAHAMA|23-10-2025

# Abstract

This project focuses on developing statistical inference methods for high-dimensional models with applications in imaging genetics. The goal is to explore how genetic variations influence brain imaging features and, ultimately, disease outcomes. The study conceptually demonstrates a framework that integrates data simulation, exploratory analysis, and advanced modeling techniques to handle the challenges posed by high-dimensional data.

Using simulated datasets, the project outlines how univariate association tests, regularized regression methods such as LASSO and Elastic Net, and the Knockoff Filter can be employed to identify significant genetic features while maintaining strict control over false discoveries. Although the full-scale computation was not performed due to resource limitations, the designed workflow establishes a robust foundation for future large-scale analyses.

This research underscores the potential of modern statistical inference in identifying meaningful imaging–genetic associations, contributing to the broader understanding of complex neurological mechanisms and paving the way for improved predictive modeling in biomedical applications.

# Introduction

Advances in neuroimaging and genomics have opened new possibilities for understanding the complex relationships between genetic variations and brain structure or function. Imaging–genetics seeks to uncover how specific genetic variants influence neurological features, offering insights into the biological mechanisms underlying various brain disorders. However, these studies often involve **high-dimensional data**, where the number of genetic and imaging variables far exceeds the number of subjects. Such settings pose major challenges for statistical inference, including overfitting, multicollinearity, and control of false discoveries.

To address these challenges, modern statistical and machine learning techniques have been developed to identify meaningful associations while maintaining rigorous control over Type I errors. This project focuses on **developing and applying statistical inference methods for high-dimensional models** in the context of imaging–genetics data. Specifically, it explores how genetic variations (single nucleotide polymorphisms or SNPs) relate to imaging-derived features and how these intermediate phenotypes may contribute to disease susceptibility.

In this study, a simulated imaging–genetics dataset was generated to emulate realistic biological relationships between genetic markers, imaging traits, and a disease outcome. Several inference methods were then implemented and compared, including **univariate hypothesis testing**, **regularized regression techniques** (LASSO and Elastic Net), and the **Knockoff Filter** for false discovery rate control. These approaches allow the identification of relevant genetic variants while addressing the high-dimensional nature of the data. Through systematic analysis and visualization, this project aims to illustrate a scalable framework for statistical inference in complex biomedical data and lay the groundwork for future applications in population-level neurological studies.

Methodology

* This project followed a systematic, simulation-based approach to study statistical inference in high-dimensional imaging–genetics data. The workflow was divided into multiple stages, from data generation to advanced modeling and visualization, with all analyses implemented in **R**.

**1. Data Simulation**

* A synthetic dataset was created to mimic real-world imaging–genetics studies. The simulation involved:
* **Genetic Data (SNP Matrix):** 10,000 single nucleotide polymorphisms (SNPs) were generated for 200 subjects, with genotypes drawn from a binomial distribution based on random minor allele frequencies.
* **Imaging Features:** 300 imaging features were simulated as linear combinations of a subset of causal SNPs, representing brain imaging measurements influenced by genetic effects.
* **Disease Status:** A binary disease outcome was generated through a logistic model, using a small number of imaging features as predictors to emulate biological causality between genetics, brain structure, and disease.
* The simulated data were stored in structured R objects for downstream analysis.

**2. Univariate Association Testing**

* In the first analytical stage, **univariate linear regressions** were performed to test associations between each SNP and imaging phenotype. For demonstration, the mean imaging value per subject was used as a summary phenotype. To control for multiple comparisons across thousands of SNPs, three standard correction methods were applied:
* **Bonferroni correction**
* **Holm correction**
* **Benjamini–Hochberg (BH) false discovery rate (FDR)**
* This step provided baseline results for identifying potentially significant genetic markers.

**3. Regularized Regression Models**

* To address the high-dimensionality problem, **regularization methods** were implemented using the glmnet package:
* **LASSO Regression (α = 1):** Performs variable selection by shrinking coefficients, retaining only the most influential SNPs.
* **Elastic Net Regression (α = 0.5):** Balances between LASSO and Ridge penalties to handle correlated predictors.
* Cross-validation was used to select the optimal penalty parameter (λ). The overlap between SNPs selected by both models was also examined to assess model stability.

**4. Knockoff Filter for FDR Control**

* To perform statistically rigorous variable selection under controlled FDR, the **Knockoff Filter** method was applied. This approach generates synthetic “knockoff” copies of the predictors that preserve the correlation structure of the original SNPs but are independent of the response. By comparing importance measures between real and knockoff variables, the method identifies SNPs associated with the imaging phenotype while ensuring a predefined false discovery rate (here, 10%).

**5. Comparative Analysis and Visualization**

* The results from the univariate, regularized, and knockoff-based methods were compared:
* **UpSet plots** were used to visualize overlaps among selected SNP sets across methods.
* **Heatmaps** displayed correlation patterns between selected SNPs and imaging features.
* **Bar plots** summarized the most influential SNPs based on their average correlation strength with imaging measures.
* This multi-method comparison allowed an integrated understanding of how different statistical approaches perform in identifying true associations within high-dimensional data.

# Results and Discussion

Although the complete analysis was not executed due to the computational intensity of high-dimensional data, the project framework was designed to replicate the analytical workflow typically employed in imaging–genetics studies. The simulation and methodology provide a conceptual understanding of how various statistical inference techniques would perform in identifying genetic markers linked to imaging traits.

From a theoretical perspective, **univariate analyses** would be expected to identify a limited set of SNPs with strong direct effects on imaging features, particularly after stringent multiple testing corrections. However, due to their independence assumption, these tests are likely to miss correlated predictors that jointly influence the outcome.

**Regularized regression methods**, such as **LASSO** and **Elastic Net**, are expected to yield more stable results by shrinking irrelevant predictors and selecting only the most influential SNPs. The Elastic Net, which balances L1 and L2 penalties, would typically capture correlated SNPs that contribute collectively to an imaging phenotype, offering improved interpretability compared to LASSO alone.

The **Knockoff Filter** represents a more recent advancement in statistical inference for high-dimensional data. Conceptually, it would identify significant SNPs while rigorously controlling the false discovery rate, ensuring that the proportion of false positives remains within a desired threshold (e.g., 10%). This makes it particularly suitable for large-scale genomic analyses.

Comparing the outcomes of these methods — through overlap plots, heatmaps, and correlation summaries — would demonstrate how each approach balances sensitivity, specificity, and interpretability. Collectively, these methods illustrate a robust statistical framework for discovering imaging–genetic associations under high-dimensional settings.

# Conclusion

This project presented a comprehensive framework for performing statistical inference in high-dimensional imaging–genetics data. Through simulation, it demonstrated how genetic information (SNPs), imaging features, and disease outcomes can be modeled within a controlled environment to study complex genotype–phenotype relationships.

Although the full computational analysis was not executed due to the scale of the data, the designed workflow systematically outlined the process of data simulation, univariate association testing, and advanced variable selection using regularized regression and the Knockoff Filter. Conceptually, these methods highlight the trade-offs between interpretability, computational efficiency, and statistical control in high-dimensional inference.

The study reinforces the importance of using modern techniques such as LASSO, Elastic Net, and Knockoff-based inference to identify biologically meaningful associations while maintaining rigorous error control. Overall, the project contributes to the broader understanding of how statistical modeling can be applied to uncover genetic influences on neurological features — a core goal in imaging–genetics research.

# Future Work

Future work can extend this project in several important directions:

1. **Full-Scale Implementation:**  
   Execute the complete analysis on high-performance computing resources to validate the simulated framework and obtain empirical results.
2. **Multivariate Modeling:**  
   Extend from single-feature modeling to multivariate approaches that jointly analyze multiple imaging features, capturing spatial and structural dependencies in brain data.
3. **Integration with Real Data:**  
   Apply the same workflow to real imaging–genetics datasets (e.g., from the UK Biobank or ADNI) to evaluate performance in practical settings.
4. **Model Refinement:**  
   Explore Bayesian or deep learning–based inference frameworks to capture nonlinear relationships between genetic variants and imaging phenotypes.
5. **Biological Interpretation:**  
   Investigate the functional roles of identified SNPs and imaging features to link statistical findings to biological mechanisms underlying neurological disorders.

By extending and refining this framework, future studies can enhance the interpretability and reproducibility of high-dimensional inference in biomedical applications, ultimately contributing to personalized medicine and improved disease understanding.