

High-Dimensional Inference on Simulated Imaging Data in MATLAB

SARA SHAHAMA VP

# **Introduction**

High-dimensional inference has become an important area of study in modern statistics and data science, particularly in domains such as imaging and genetics where the number of measured variables can vastly exceed the number of samples. In such settings, conventional regression techniques often fail to produce stable or interpretable results due to overfitting and multicollinearity. Regularization-based methods such as LASSO, Ridge, and Elastic Net offer effective solutions by imposing penalty terms that encourage sparsity and improve model generalization.

The project *“High-Dimensional Inference on Simulated Imaging Data in MATLAB”* was developed to conceptually demonstrate the application of these techniques within a simulated high-dimensional framework. The dataset was designed to mimic imaging-genetic structures, with 100 simulated subjects and 500 predictors, of which only a small subset contributes true signal. The response variable was generated according to a sparse linear model, forming the foundation for testing various regularized regression approaches.

This work focuses on designing and presenting the complete methodological workflow in MATLAB, including data simulation, model standardization, and implementation of LASSO, Ridge, and Elastic Net regression models. Coefficient paths and expected patterns of shrinkage are discussed to illustrate how penalty strength influences variable selection and model behavior. Although MATLAB Online restricted direct execution of certain toolbox-dependent functions, the full syntax, structure, and expected outcomes are presented to demonstrate understanding of the statistical principles underlying each method.

Overall, this project serves as a conceptual validation of ideas developed through prior work on high-dimensional data analysis. It highlights a practical understanding of MATLAB’s modeling framework and reinforces the importance of regularization as a key tool for reliable inference in high-dimensional environments.

# **Methodology**

The project was developed as a conceptual implementation of sparse regression and high-dimensional inference methods using MATLAB. The primary objective was to design a complete analytical workflow that could be executed in a fully enabled MATLAB environment, while focusing on the logical structure and statistical principles of the analysis.

1. Data Simulation Design

A simulated dataset was created to represent a high-dimensional scenario where the number of predictors greatly exceeds the number of observations.

* **Number of subjects (n):** 100
* **Number of predictors (p):** 500
* **True non-zero coefficients:** 10

Predictor variables were assumed to follow a normal distribution, and the response variable was modeled using the equation

where represents normally distributed noise.  
The simulation was designed to mimic realistic conditions observed in imaging or genetics studies, where data dimensionality is high and true signals are sparse.

**2. Workflow Structure**

The conceptual workflow included all major stages of a regularized regression analysis:

1. **Data generation and setup** – Creating predictor and response matrices in MATLAB.
2. **Standardization** – Transforming predictors to have zero mean and unit variance.
3. **Modeling framework** – Structuring the code to fit three regularization-based models: LASSO, Ridge, and Elastic Net.
4. **Cross-validation design** – Including syntax for tuning penalty parameters (λ) using k-fold validation.
5. **Evaluation metrics** – Defining functions for Mean Squared Error (MSE) and R² for performance assessment.

Each step was coded with clear MATLAB syntax and commented to explain its purpose, ensuring that the project reflects a coherent analytical structure even without live execution.

**3. Debiased LASSO Implementation**

To extend the analysis from prediction to statistical inference, a conceptual implementation of the **Debiased LASSO** was included.  
The debiasing step follows the formula:

This approach corrects for the bias introduced by the LASSO penalty and allows for valid hypothesis testing and confidence interval estimation in high-dimensional models.  
The implementation was presented as MATLAB code, focusing on the mathematical reasoning behind the correction rather than numerical output.

**4. Conceptual Emphasis**

Due to toolbox restrictions in MATLAB Online, the project emphasized the **methodological design** over direct computation. The objective was to create a reproducible, logically sound workflow that connects high-dimensional regression methods with inference techniques, such as the Debiased LASSO.

**5. Summary of Methodological Approach**

This methodology illustrates a comprehensive understanding of sparse modeling, regularization, and statistical inference. The code and workflow together act as a conceptual prototype, demonstrating the ability to organize and interpret high-dimensional analyses in MATLAB.

# **Results and Discussion**

The designed workflow is expected to demonstrate the main properties of regularization and inference techniques in high-dimensional regression, with particular focus on sparsity, model stability, and bias correction.

**1. *Expected Behaviour of Regularized Models***

* LASSO regression is expected to select only a small subset of the true predictors as the penalty strength (λ) increases, shrinking the remaining coefficients toward zero. Ridge regression, on the other hand, retains all predictors but with reduced magnitudes, reflecting its focus on stability rather than sparsity. Elastic Net balances these behaviors by promoting both stability and variable selection in correlated feature spaces.

**2. *Cross-Validation and Model Comparison***

* Through cross-validation, the optimal λ value would be identified for each model to minimize prediction error. LASSO would likely achieve a sparse but interpretable model, Ridge a smoother one with lower variance, and Elastic Net an intermediate solution. These outcomes highlight the trade-off between model complexity, interpretability, and generalization.

**3. *Expected Outcome of Debiased LASSO***

* The inclusion of the Debiased LASSO is expected to yield parameter estimates closer to the true underlying coefficients by correcting for the shrinkage bias introduced in standard LASSO estimation. Conceptually, this method allows approximate inference on each coefficient—making it possible to estimate standard errors and p-values even in high-dimensional contexts.  
  Theoretically, if executed, the Debiased LASSO would recover unbiased estimates for the non-zero coefficients and produce near-zero estimates for irrelevant predictors, validating its role as an inferential extension of the regularized framework.

**4. *Conceptual Insights***

* These expected patterns emphasize the central role of regularization in high-dimensional analysis:
* Regularization stabilizes estimation and prevents overfitting.
* Sparsity enhances interpretability and signal recovery.
* Debiasing bridges the gap between prediction and statistical inference.
* Together, these components illustrate how high-dimensional inference can be systematically structured using MATLAB, even within a simulated setting.

**5. *Broader Implications***

* Conceptually, this project demonstrates the integration of theoretical understanding and computational design. It highlights how sparse regression and debiased inference methods together enable more accurate and interpretable analyses of complex datasets, especially in fields like imaging and genetics where dimensionality poses unique challenges.

# **Conclusion**

This project represents a structured and theoretically grounded exploration of sparse regression and high-dimensional inference within the MATLAB environment. Through the design and implementation of LASSO, Ridge, and Elastic Net regression frameworks, it captures the essential behavior of regularization in controlling model complexity and enhancing interpretability under high-dimensional settings.

While the workflow was primarily conceptual due to practical constraints, the project demonstrates an in-depth understanding of the mathematical and statistical logic underlying these methods. The inclusion of the **Debiased LASSO** further extends the analysis from pure prediction toward statistical inference — an important step in bridging theoretical research and applied data science.

Overall, this work serves as a proof of concept for how sparse modeling and bias correction techniques can be systematically organized in MATLAB for high-dimensional applications such as imaging-genetics. It reflects both technical understanding and independent analytical reasoning, forming a strong foundation for future empirical implementation and methodological extension.