Subsampling vs Ridge vs Lasso Regularization

Comparative Evaluation Using Simulated Data

16:958:588:02 DATA MINING

Team: Sanjith Ganesh, Anshu Goli, Pranav Senthilkumaran

1. Abstract:

High-dimensional regression models are vulnerable to **overfitting** due to **multicollinearity** and noise, making regularization techniques essential for reliable performance. This study compares three approaches to address this issue: **Ridge regression** (L2 regularization), **Lasso regression** (L1 regularization) and **subsampling-based ensembling**. A synthetic dataset (simulated data) was simulated with known signal structures and controlled noise levels to evaluate each method's effectiveness using **Mean Squared Error (MSE)** and **variance**.

Our findings show that Ridge regression consistently achieves the lowest error and highest stability, while Lasso provides similar accuracy with the added advantage of feature selection and sparsity. In contrast, subsampling-based ensembles lack formal regularization and suffer from high error and variability.

These results confirm that **penalization** techniques outperform naive ensembling in high-dimensional settings and highlight the importance of structured regularization. The results suggest that structured penalization is essential for reliable generalization in noisy, high-dimensional contexts. This study lays a foundation for future work involving real-world datasets and hybrid models like **Elastic Net** (Convex Combination of L1 and L2 Penalty Parameter) that combine the strengths of ensembling and regularization.

Keywords: Overfitting, Multicollinearity, Ridge regression, Lasso regression, Subsampling-based ensembling, Mean Squared Error, Variance, Penalization, Elastic Net

2. Introduction:

2.1 Challenges in High-Dimensional Regression (Effect of Overfitting):

High-dimensional regression—where the number of predictors is large relative to the number of observations—poses significant statistical challenges. In such settings, models are prone to **overfitting**, where they capture noise rather than meaningful signal, and **multicollinearity**, where highly correlated features inflate variance in coefficient estimates. These issues reduce the model's generalization ability and stability, making effective **regularization** essential.

2.2 Penalization Techniques: Ridge and Lasso:

Regularization techniques address overfitting by penalizing model complexity, thereby controlling the variance of parameter estimates. Two of the most widely used methods are **Ridge regression** and **Lasso regression**. Ridge regression incorporates an **L2 penalty** that shrinks all coefficients toward zero but retains all variables in the model. This leads to a more stable solution in high-dimensional contexts, especially when multicollinearity is present. In contrast, **Lasso regression** uses an **L1 penalty**, which not only shrinks coefficients but also drives some to exactly zero. This results in **sparse solutions** and enables built-in **feature selection**, offering interpretability alongside regularization.

2.3 Subsampling-Based Ensembling as an Alternative:

An alternative approach to penalization is **subsampling-based ensembling**, where multiple models are trained on random subsets of the training data, and their predictions are aggregated. This technique, common in methods like bagging, is intended to reduce prediction variance and improve robustness. However, it lacks the explicit constraint on model complexity that regularization provides. As such, its ability to serve as a **surrogate for formal regularization** remains an open question, particularly in noisy, high-dimensional data environments.

2.4 Research Objective and Methodological Framework:

This paper investigates the question:

"Can ensembling through subsampling effectively reduce variance and improve prediction accuracy in high-dimensional regression, or are formal penalization techniques better?"

To address this, a synthetic dataset is generated with a known signal structure and controlled noise. The simulation environment offers full control over the data-generating process, enabling objective and interpretable comparisons across methods. The study implements and evaluates three modeling strategies—**Ridge regression**, **Lasso regression**, and **subsampling-based ensembles**—using **Mean Squared Error (MSE)** and **prediction variance** as primary evaluation metrics.

By conducting direct comparisons on identical data, this analysis offers empirical insights into the effectiveness of each method in mitigating overfitting and managing high-dimensional noise. The findings emphasize the advantages of structured penalization techniques and highlight their superiority over naive ensembling in such settings. Furthermore, this investigation lays the groundwork for future research into **hybrid modeling approaches** that combine regularization with subsampling techniques for improved performance on real-world datasets.

3. Methodology:

3.1 Data Simulation Setup:

To facilitate controlled evaluation, a **synthetic dataset** was generated consisting of **200 observations** and **100 predictors (No Overfitting)** drawn from a standard normal distribution.

True coefficients were sampled from a normal distribution, and the target variable y was generated as **$y=X\beta+\varepsilon$, where $\varepsilon \sim N(0,1)$**. This setup ensures full control over noise and signal, eliminating external biases and allowing for an objective assessment of model performance.

3.2 Ridge Regression (L2 Regularization):

Ridge regression introduces an L2 penalty term to the **ordinary least squares** (OLS) objective, minimizing:

$$(Loss=\|y-X\beta\|^2 + \lambda\|\beta\|2^2)$$

Model training was performed using **10-fold cross-validation** via cv.glmnet() with **alpha = 0** to determine the optimal regularization parameter λ. The final model was fitted using the full training data and tested on the holdout set (30% split).

3.3 Lasso Regression (L1 Regularization):

Lasso regression applies an L1 penalty, optimizing:

$$(Loss=\|y-X\beta\|^2 + \lambda\|\beta\|1)$$

This leads to sparse solutions where some coefficients are driven to zero. Lasso was implemented similarly to Ridge using cv.glmnet() with **alpha = 1**.

3.4 Subsampling-Based Ensembling:

This method involves training multiple OLS models on randomly sampled subsets of the training data and averaging their predictions. Specifically:

- 30 iterations of random 50% subsampling from the training set
- An OLS model was trained on each subset
- Predictions were averaged across models using rowMeans()

While conceptually aiming to reduce variance, this method lacks a formal penalty term and is sensitive to data randomness.

4. Findings:

4.1 Mean Squared Error (MSE) Comparison:

Method	Test MSE
Ridge	4.18
Lasso	5.08
Subsampling	8053.35

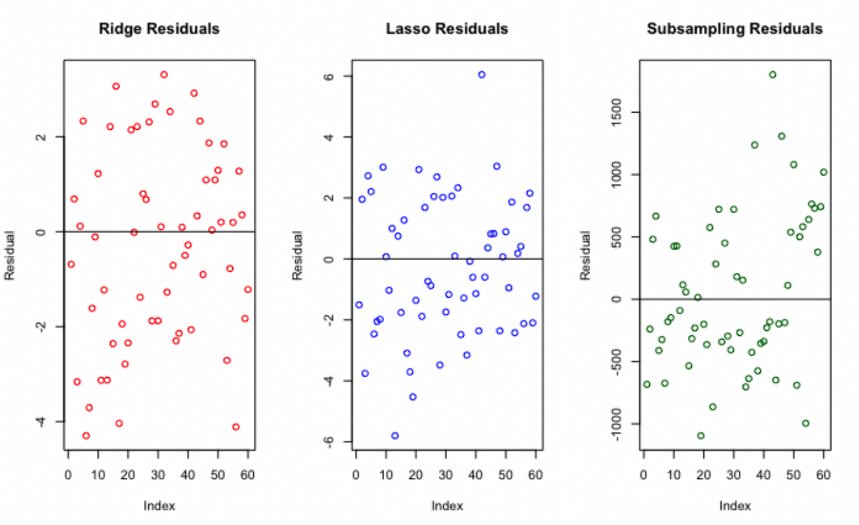
4.2 Variance and Stability:

Bias-variance decomposition revealed that:

- **Ridge had the lowest variance (0.07)**: Ridge is most stable.
- **Lasso** had slightly higher variance **(0.36)**: Lasso is consistent with moderate variance
- **Subsampling** suffered from extremely high variance **(4.43 million)** — too large, making its predictions unstable and unreliable

4.3 Residual Distribution:

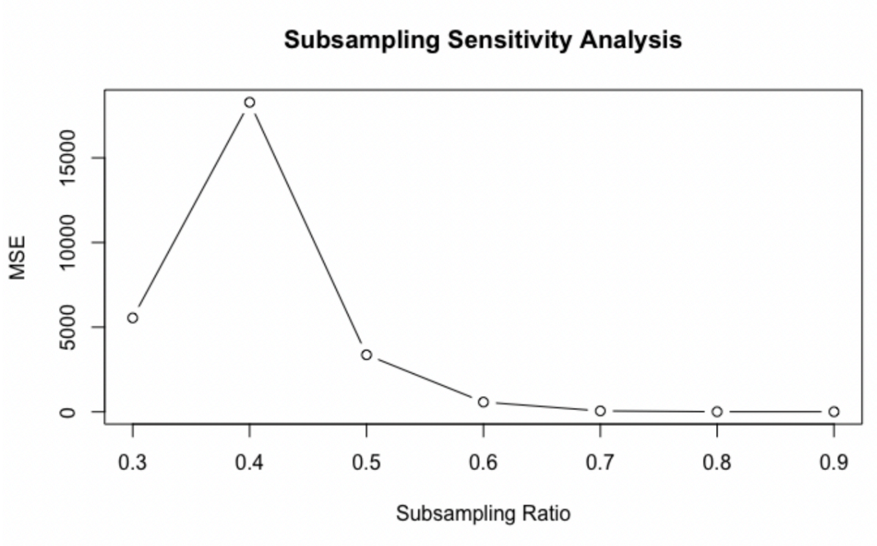
Residual plots show:



•Ridge and Lasso residuals were **centered and narrow** and fits well.

•Subsampling residuals were **widely scattered**, confirming erratic behavior

4.4 Sensitivity to Subsample Size:



As subsample size increased **(30% to 90%)**, **MSE decreased** but remained significantly higher than Ridge/Lasso, therefore suggesting that increasing the ensemble size improves performance but cannot compensate for the lack of regularization (still worse than Ridge and Lasso).

5. Conclusion:

This paper conducted a comparative evaluation of **Ridge regression**, **Lasso regression**, and **subsampling-based ensembling**. Through controlled simulations using synthetic data with known signal and noise, the findings reveal clear differences in model performance, stability, and interpretability.

- **Ridge regression** emerged as the most robust method, consistently achieving the **lowest test error** and **minimal variance**. Its L2 penalty efficiently controls model complexity by uniformly shrinking coefficients, making it especially effective in scenarios with multicollinearity and distributed signal strength across predictors. **(Best)**

- **Lasso regression** provided **competitive accuracy** while offering the added benefit of **sparse solutions** through L1 penalization. By setting many coefficients to zero, Lasso aids in variable selection and enhances model interpretability—valuable for domains where identifying key predictors is important.

- **Subsampling-based ensembling**, despite its theoretical appeal for variance reduction, **underperformed significantly**. It produced high error and unstable predictions due to the absence of explicit regularization and sensitivity to data randomness. Larger subsample sizes improved performance marginally but failed to match Ridge or Lasso. **(Worst)**

These results confirm that **structured regularization through penalization** is **crucial for achieving generalization and stability** in high-dimensional, noisy settings. **Subsampling alone cannot substitute for the mathematical guarantees provided by Ridge and Lasso**. Additionally, the simulation validated that tuning regularization parameters (example: lambda) is critical; overly aggressive penalization can lead to underfitting, while light regularization offers the best trade-off between bias and variance.

6. Implications and Future Directions:

1. Exploring **Elastic Net** as a hybrid approach that combines the strengths of L1 and L2 penalties, especially useful when features are both correlated and sparse.
2. Designing **regularized ensemble methods** that integrate subsampling with penalization (example: **bootstrapped Ridge or Lasso models) —** Combine subsampling with regularization techniques.
3. Applying these methods to **real-world high-dimensional datasets** where sparsity and interpretability are domain-critical.
4. Evaluating **computational efficiency and scalability**, particularly as dataset size and dimensionality grow, which is relevant for modern applications in **genomics, finance and text data**.

“Overall, this paper reaffirms that penalization, not ensembling alone, is the more effective strategy for managing overfitting in high-dimensional regression contexts.”