A Comparative Study of LASSO and Ranked Sparsity Regularization

A Structured Framework for Feature Selection

Samuella Boadi

Advanced Computation Project Presentation

University of Iowa

Motivation for Ranked Sparsity Regularization

- In many modern regression problems, incorporating interactions and nonlinear effects causes the number of candidate predictors to grow rapidly.
- Standard LASSO treats all predictors equally, applying the same penalty regardless of whether a term is a main effect or a higher-order interaction.
- In practice, domain knowledge suggests that main effects are often more fundamental than interactions or polynomial terms.
- Ranked Sparsity Regularization (RSR) formalizes this intuition by introducing structure into the regularization process.

LASSO vs Ranked Sparsity Regularization (RSR)

Objective Functions

LASSO:

$$\min_{\beta} \ \frac{1}{2n}\|y - X\beta\|_2^2 \ + \ \lambda \sum_{j=1}^{p} |\beta_j|.$$

Ranked Sparsity Regularization (RSR):

$$\min_{\beta} \ \frac{1}{2n}\|y - X\beta\|_2^2 \ + \ \lambda \sum_{k=1}^{K} w_k \|\beta_{A_k}\|_1.$$

Notation:

- $X \in \mathbb{R}^{n \times p}$: design matrix with n observations and p predictors
- $y \in \mathbb{R}^{n}$: response vector
- $\beta \in \mathbb{R}^{p}$: regression coefficients
- $\lambda > 0$: regularization parameter controlling sparsity
- A_k: group of predictors of rank k (e.g., main effects, interactions, polynomials)

Ranked Sparsity Regularization: Key Components

Penalty Structure

- β_{A_k}: coefficient vector corresponding to group A_k
- w_k: penalty weight assigned to rank k
- A common choice of weights is

$$w_k = p_k^{1-2\gamma},$$

  where $p_k = |A_k|$ is the number of features in group A_k and $\gamma \in [0, 0.5]$

Interpretation

- Larger values of w_k impose stronger penalties on higher-rank features
- This discourages unnecessary interaction and polynomial terms

Key Differences Between LASSO and RSR

- **LASSO** applies the same ℓ_1 penalty to all coefficients, regardless of feature structure or complexity.
- As a result, LASSO does not distinguish between main effects and higher-order terms.
- **RSR** introduces structured regularization by assigning different penalties to different feature ranks.
- Higher-rank features (e.g., interactions and polynomials) receive stronger penalties through the weights w_k.

Why Ranked Sparsity Regularization (RSR)?

- Ranked Sparsity Regularization (RSR) encourages sparsity in a structured and principled manner.
- Main effects, which typically form smaller groups, receive weaker penalties.
- Interaction and higher-order terms receive stronger penalties.
- This structure improves interpretability and aligns with scientific intuition.
- By discouraging unnecessary complexity, RSR can also improve predictive performance.

Algorithmic Motivation: LASSO as a Baseline

- Before introducing the RSR algorithm, we review LASSO as a baseline.
- RSR is implemented using a similar coordinate descent framework.
- LASSO updates one coefficient at a time while holding others fixed.
- The same penalty parameter is applied to all predictors.

Algorithm 1: Coordinate Descent for LASSO

Algorithm 1 Coordinate Descent Algorithm for LASSO

Input: Design matrix $X \in \mathbb{R}^{n \times p}$, response vector $y \in \mathbb{R}^n$, penalty parameter $\lambda > 0$, tolerance $\epsilon > 0$ **Standardize** the columns of X and **center** y **Initialize** $\beta^{(0)} = \mathbf{0} \in \mathbb{R}^p$ $j = 1, \ldots, p$ Compute the partial residual:

$$r_j = y - \sum_{k \neq j} X_k \beta_k$$

Compute the partial gradient:

$$z_j = \frac{1}{n} X_j^\top r_j$$

Algorithm 1, cont

Update coefficient using soft-thresholding:

$$\beta_j \leftarrow \mathrm{sign}(z_j) \max\big(|z_j| - \lambda,\, 0\big)$$

$\|\beta^{(t)} - \beta^{(t-1)}\|_\infty < \epsilon$ **Output:** $\hat{\beta}$

Algorithm 2: Ranked Sparsity Regularization (RSR)

- Ranked Sparsity Regularization (RSR) extends LASSO by assigning rank-dependent penalty weights to groups of predictors.
- Lower-rank features (e.g., main effects) receive smaller penalties.
- Higher-rank features (e.g., interactions and polynomial terms) receive larger penalties.
- The optimization algorithm remains coordinate descent, but the soft-thresholding step now depends on feature rank.

Algorithm 2 Coordinate Descent Algorithm for Ranked Sparsity Regularization (RSR)

Input: Design matrix $X = [A_1, \ldots, A_K]$, response vector y, penalty parameter $\lambda > 0$, ranking parameter $\gamma > 0$, tolerance $\epsilon > 0$ **Standardize** the columns of X and **center** y **Initialize** coefficient vector $\beta^{(0)} = \mathbf{0}$ Compute group sizes $p_k = |A_k|$ for $k = 1, \ldots, K$ Compute group penalty weights:

$$w_k = p_k^{1-2\gamma}, \quad k = 1, \ldots, K$$

$k = 1, \ldots, K$ each coefficient $j \in A_k$ Compute the partial residual:

$$r_{kj} = y - \sum_{(g,h) \neq (k,j)} X_{gh} \beta_{gh}$$

Compute the partial gradient:

$$z_{kj} = \frac{1}{n} X_{kj}^\top r_{kj}$$

Frame Title

Update coefficient using rank-weighted soft-thresholding:

$$\beta_{kj} \leftarrow \mathrm{sign}(z_{kj})\max\big(|z_{kj}| - \lambda w_k,\, 0\big)$$

$\|\beta^{(t)} - \beta^{(t-1)}\|_\infty < \epsilon$ **Output:** $\hat{\beta}$

Data Description

The analysis is based on the `mtcars` dataset, which contains technical specifications for **32 automobiles** originally reported in *Motor Trend* magazine.

- **Sample size:** $n = 32$ cars
- **Response variable:**
  - `mpg`: miles per gallon (fuel efficiency)
- **Predictors:** The dataset includes **10 automotive characteristics**, such as:
  - `wt`: vehicle weight
  - `hp`: gross horsepower
  - `disp`: engine displacement
  - `cyl`: number of cylinders
  - `qsec`: quarter-mile time
  - `am`: transmission type (manual vs. automatic)
  - `gear`, `carb`, `vs`, `drat`

These variables capture a mix of **engine performance**, **vehicle design**, and **transmission features** that influence fuel efficiency.

Final LASSO Model

$$
\begin{aligned}
\widehat{\text{mpg}} = 36.03 &- 2.71\,\texttt{wt} - 0.89\,\texttt{cyl} - 0.012\,\texttt{hp} \\
&+ 0 \cdot \texttt{disp} + 0 \cdot \texttt{drat} + 0 \cdot \texttt{qsec} + 0 \cdot \texttt{vs} \\
&+ 0 \cdot \texttt{am} + 0 \cdot \texttt{gear} + 0 \cdot \texttt{carb}
\end{aligned}
$$

Interpretation: LASSO selects a small set of main effects, producing a sparse and interpretable model.

Final RSR Model

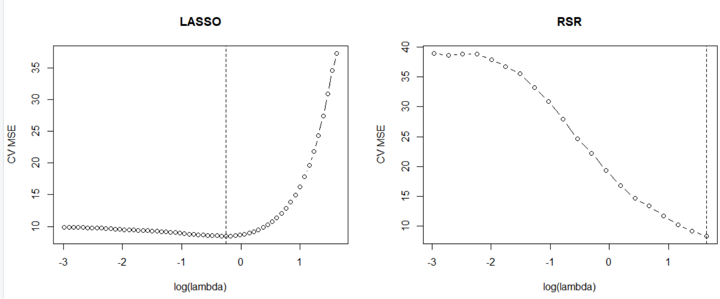
$$
\begin{aligned}
\widehat{\text{mpg}} = {} & 17.37 + 2.45\,(\texttt{vs} \times \texttt{am}) - 0.57\,(\texttt{wt} \times \texttt{gear}) - 0.25\,(\texttt{drat} \times \texttt{wt}) \\
& + 0.18\,(\texttt{qsec} \times \texttt{gear}) + 0.06\,(\texttt{drat} \times \texttt{qsec}) - 0.010\,(\texttt{hp} \times \texttt{vs}) \\
& - 0.0098\,(\texttt{qsec} \times \texttt{carb}) - 0.0085\,(\texttt{wt} \times \texttt{qsec}) - 0.0009\,(\texttt{hp} \times \texttt{qs} \\
& + \sum_{\text{all main effects}} 0 \cdot X_j + \sum_{\text{all quadratic terms}} 0 \cdot X_j
\end{aligned}
$$

Interpretation: RSR selects interaction effects while shrinking all main and quadratic terms to zero.

Some conclusions

- LASSO produces a sparse and interpretable model based on main effects.
- RSR allows interactions and nonlinear effects while controlling complexity.
- RSR achieves comparable or improved predictive performance.
- RSR is more expressive, but LASSO remains preferable when simplicity is required.

Prediction Accuracy: LASSO vs RSR

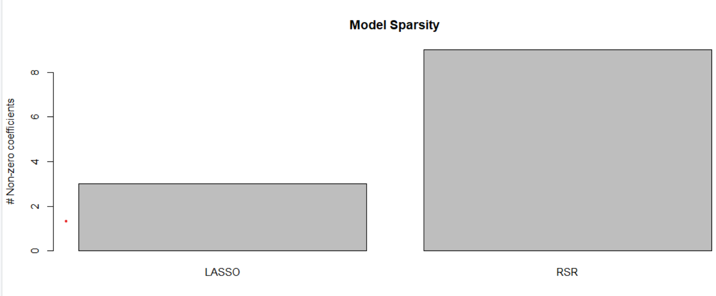


Model Prediction Performance: LASSO vs RSR

Interpretation of CV Error Curves

- The plots show cross-validated mean squared error (CV MSE) as a function of the regularization parameter $\log(\lambda)$.
- **LASSO (left):**
  - The CV error has a U-shaped pattern.
  - Small λ leads to overfitting, while large λ causes excessive shrinkage and underfitting.
  - Prediction accuracy deteriorates when all coefficients are penalized equally.
- **RSR (right):**
  - CV error decreases as λ increases.
  - Stronger regularization removes high-rank interaction and polynomial terms first.
  - Important main effects remain in the model longer, improving generalization.
- Overall, RSR achieves comparable or better prediction accuracy by enforcing structured sparsity and avoiding unnecessary model complexity.

Prediction Accuracy: LASSO vs RSR

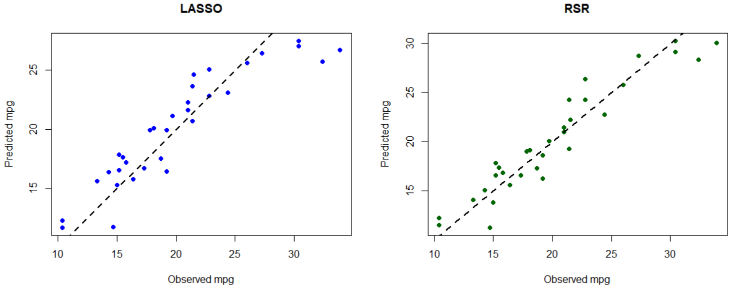


Comparison of Model Sparsity

Model Sparsity and Variable Selection

- Model sparsity is measured by the number of non-zero coefficients in the final fitted model.
- **LASSO:**
  - Selects a very small number of predictors.
  - Produces a highly sparse and interpretable model.
  - Primarily retains main effects.
- **RSR:**
  - Selects a larger set of predictors.
  - Includes interaction and polynomial terms.
  - Captures more complex relationships in the data.

Prediction Accuracy: LASSO vs RSR



Prediction Accuracy

Conclusion

- LASSO provides a simple and interpretable approach to variable selection by enforcing unstructured sparsity.
- Ranked Sparsity Regularization (RSR) extends LASSO by incorporating feature hierarchy, penalizing higher-order terms more strongly.
- In the `mtcars` application, both methods achieved good predictive performance.
- RSR selected richer interaction structures and showed modest improvements in prediction accuracy for some observations.

Thank You