Why Lasso Does Not Exhibit the Grouping Effect

1. Independent Treatment of Predictors:
   The L_1 penalty treats each predictor independently, without considering correlations between them. As a result, lasso arbitrarily selects one predictor from a group of highly correlated variables and shrinks the rest to zero.

2. Geometric Explanation:
   - Lasso's constraint region is diamond-shaped, which encourages sparsity by forcing some coefficients to zero.
   - For highly correlated predictors, the optimal solution often lies on a vertex of the diamond, where only one coefficient is nonzero. This breaks the grouping effect.

3. Example from Tibshirani (1996):
   - Consider a simple linear model with two predictors ($p = 2$)

     $$|\hat{\beta}_1 - \hat{\beta}_2| = |cos(\theta)|$$

     where, θ: Angle between $\boldsymbol{y}$ and the difference vector $\boldsymbol{x}_1 - \boldsymbol{x}_2$
   - As $\rho = corr(\boldsymbol{x}_1, \boldsymbol{x}_2) \to 1$, $cos(\theta)$ does not necessarily vanish. This means the coefficients $\hat{\beta}_1$ and $\hat{\beta}_2$ may still differ significantly, even for highly correlated predictors.

Deficiency of the Naïve Elastic Net

Deficiency in Regression Prediction:

- The naïve elastic net follows a two-stage procedure:
  1. Two-stage procedure: For each fixed λ_2, we first find the Ridge regression coefficients, and then we do the LASSO-type shrinkage along the LASSO coefficient solution paths.
- This results in a double amount of shrinkage, which:
  1. Fails to significantly reduce variance.
  2. Introduces unnecessary extra bias.

Conclusion:

- The double shrinkage of the naïve elastic net is less effective compared to pure lasso or ridge shrinkage.
- The Naïve elastic net does not perform satisfactory unless it is very close to either ridge or lasso.

Elastic Net Computation

- Given data $(y, \boldsymbol{X})$, penalty parameters (λ_1, λ_2), and augmented data $(y^*, \boldsymbol{X}^*)$, the naive elastic net solves a LASSO-type problem:

$$\hat{\beta}^* = \arg\min_{\beta^*} \|y^* - \boldsymbol{X}^*\beta^*\|^2 + \frac{\lambda_2}{\sqrt{1+\lambda_2}}\|\beta^*\|_1.$$

- The elastic net (corrected) estimates of β are defined by:

$$\hat{\beta}(\text{elastic net}) = \sqrt{1+\lambda_2}\hat{\beta}^*.$$

- Recall that $\hat{\beta}(\text{naive elastic net}) = \frac{1}{\sqrt{1+\lambda_2}}\hat{\beta}^*$, and thus:

$$\hat{\beta}(\text{elastic net}) = (1+\lambda_2)\hat{\beta}(\text{naive elastic net}).$$

- The Elastic Net coefficient is a rescaled Naive Elastic Net coefficient.

The Elastic Net Estimate

- Previously, $\hat{\beta}(\text{elastic net}) = (1+\lambda_2)\hat{\beta}(\text{naive elastic net})$ is defined to overcome two steps of shrinkage (ridge and LASSO) by the penalty in the elastic net estimates.
- For sampling correlation of $\hat{\Sigma}$,which ($\hat{\Sigma} = \boldsymbol{X}^\top \boldsymbol{X}$), the newly defined

  $$\hat{\Sigma}_{\lambda_2} = \left(\frac{1}{1+\lambda_2}\right)\hat{\Sigma} + \left(\frac{\lambda_2}{1+\lambda_2}\right)\mathrm{I},$$

  yields to reduce the correlation matrix for the predictors.
- Under the OLS, the ridge coefficients are

  $$\hat{\beta}_{\text{ridge}} = \frac{1}{1+\lambda_2}\hat{\Sigma}_{\lambda_2}^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

- **Theorem 2.** Suppose data is given as $(\boldsymbol{y}, \boldsymbol{X})$, then $\hat{\beta}_{\text{enet}}$ is given as

  $$\hat{\beta}_{\text{enet}} = \arg\min_{\beta} \beta^T \left(\frac{\boldsymbol{X}^T\boldsymbol{X} + \lambda_2 \mathrm{I}}{1+\lambda_2}\right)\beta - 2\boldsymbol{y}^T\boldsymbol{X}\beta + \lambda_1|\beta|,$$

  as an explicit optimization. For orthogonal design ($\boldsymbol{X}^T\boldsymbol{X} = \mathrm{I}$), $\hat{\beta}_{\text{enet}}$ reduces the $\hat{\Sigma}$ into I.

Proof of Theorem 2

- Let $\hat{\boldsymbol{\beta}}$ be the elastic net estimates. By definition and Equation (10), we have:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left|\left| \boldsymbol{y}^* - \boldsymbol{X}^* \frac{\boldsymbol{\beta}}{\sqrt{1+\lambda_2}} \right|\right|^2 + \frac{\lambda_1}{\sqrt{1+\lambda_2}} \left|\left| \frac{\boldsymbol{\beta}}{\sqrt{1+\lambda_2}} \right|\right|_1 .$$

- This simplifies to:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \boldsymbol{\beta}^\top \left(\frac{\boldsymbol{X}^{*\top}\boldsymbol{X}^*}{1+\lambda_2} \right) \boldsymbol{\beta} - 2 \frac{\boldsymbol{y}^{*\top}\boldsymbol{X}^*}{\sqrt{1+\lambda_2}} + \frac{\lambda_1}{1+\lambda_2} ||\boldsymbol{\beta}||_1 + \boldsymbol{y}^{*\top}\boldsymbol{y}^* .$$

- Substituting the following identities:

$$\boldsymbol{X}^{*\top}\boldsymbol{X}^* = \frac{\boldsymbol{X}^\top\boldsymbol{X} + \lambda_2 \boldsymbol{I}}{1+\lambda_2},$$

$$\boldsymbol{y}^{*\top}\boldsymbol{X}^* = \frac{\boldsymbol{y}^\top\boldsymbol{X}}{\sqrt{1+\lambda_2}}, \quad \boldsymbol{y}^{*\top}\boldsymbol{y}^* = \boldsymbol{y}^\top\boldsymbol{y},$$

  into the minimization problem, we obtain:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{1+\lambda_2} \left\{ \boldsymbol{\beta}^\top \left(\frac{\boldsymbol{X}^\top\boldsymbol{X} + \lambda_2 \boldsymbol{I}}{1+\lambda_2} \right) \boldsymbol{\beta} - 2\boldsymbol{y}^\top\boldsymbol{X}\boldsymbol{\beta} + \lambda_1 ||\boldsymbol{\beta}||_1 \right\} + \boldsymbol{y}^\top\boldsymbol{y}$$

Final Formulation

- Removing the constant term $\boldsymbol{y}^\top \boldsymbol{y}$, the optimization simplifies to:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \boldsymbol{\beta}^\top \left(\frac{\boldsymbol{X}^\top \boldsymbol{X} + \lambda_2 \boldsymbol{I}}{1+\lambda_2} \right) \boldsymbol{\beta} - 2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} + \lambda_1 ||\boldsymbol{\beta}||_1.$$

- This is the desired form of Theorem 2:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \boldsymbol{\beta}^\top \left(\frac{\boldsymbol{X}^\top \boldsymbol{X} + \lambda_2 \boldsymbol{I}}{1+\lambda_2} \right) \boldsymbol{\beta} - 2\boldsymbol{y}^\top \boldsymbol{X}\boldsymbol{\beta} + \lambda_1 ||\boldsymbol{\beta}||_1.$$

** The term:

$$\frac{\mathrm{X}^\top \mathrm{X} + \lambda_2 \mathrm{I}}{1+\lambda_2} = (1-\gamma)\hat{\Sigma} + \gamma \mathrm{I},$$

where $\gamma = \frac{\lambda_2}{1+\lambda_2}$, shrinks $\hat{\Sigma}$ towards the identity matrix.

- This regularization replaces $\hat{\Sigma}$ with its shrunk version in the lasso formulation.
- Rescaling after elastic net penalization is mathematically equivalent to replacing $\hat{\Sigma}$ with its shrunken version.

LARS-EN Algorithm

- For LARS (Least Angle Regression) algorithm, the elastic net solution paths increase gradually in a predictable manner.
- The idea of the LARS algorithm is to solve for the whole LASSO problems effectively as to compute for the same steps of fitted OLS.
- Within fixed individual λ_2, the LARS-EN (λ_1 and λ_2) uses the single OLS fit to solve for the whole elastic net solutions ($\lambda_1, s,$ or k).
- For the kth step, efficient update or downdate of the Cholesky factorization for the inverted

  $$G_{A_k} = {X^*_{A_k}}^T {X^*}_{A_k} = \frac{1}{1+\lambda_2}\left(X^T_{A_k-1}X_{A_k-1} + \lambda_2 I\right)$$

  is recorded for non-zero coefficients, not to explicitly use X^* to compute for all quantities.
- Algorithm LARS-EN sequentially updates the elastic net fits. As an empirical evidence, the real and experiment simulations show the optimal results for early stops of LARS-EN.

Choice of Tuning Parameters: Elastic Net (1/3)

Parameters in Elastic Net:

- The Elastic Net involves two regularization parameters:

  $$\lambda_1 : \text{Lasso penalty (L1 norm)}, \quad \lambda_2 : \text{Ridge penalty (L2 norm)}.$$

- Elastic Net criterion:

  $$L(\lambda_1, \lambda_2, \beta) = \|y - X\beta\|^2 + \lambda_2\|\beta\|^2 + \lambda_1\|\beta\|_1.$$

- These parameters control sparsity (λ_1) and the grouping effect (λ_2).

Choice of Tuning Parameters: Elastic Net (2/3)

Alternative Parameterizations:

- Instead of (λ_1, λ_2), we can use:
  1. (λ_2, t), where $t = \|\beta\|_1$, the L1 norm of coefficients.
  2. (λ_2, s), where $s = t/\|\beta_{\text{max}}\|_1$, a normalized fraction of the L1 norm.
  3. In algorithm LARS, the lasso is described as a forward stagewise fitting procedure.This view adopts the number of steps k as a tuning parameter for the lasso.
- Normalization ensures $s \in [0, 1]$, simplifying comparisons across models.
- Parameters define piecewise linear solution paths, which are computationally efficient to evaluate.

Tuning Parameters in Elastic Net

1. **Grid Search:** A grid of λ_2 values (e.g., $\{0, 0.01, 0.1, 1, 10, 100\}$) is explored. For each λ_2, the **LARS-EN algorithm** computes the entire solution path for λ_1.
2. **Cross-Validation (CV):** Tenfold CV estimates prediction error. The λ_2 minimizing CV error is selected, and λ_1 is tuned along its path.
3. **Efficiency:**
   - Computational cost for CV is equivalent to 10 OLS fits per λ_2.
   - Feasible for both $n > p$ and $p \gg n$ (cost grows linearly with p).
4. **Early Stopping:** In the $p >> n$ case, early stopping is used to ease the computational burden.

   suppose that $n = 30$ and $p = 5000$; if we do not want more than 200 variables in the final model, we may stop algorithm LARS-EN after 500 steps and consider only the best k within 500.

Simulation

- The simulated data comes from the true model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \sigma\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, 1).$$

- Each simulated dataset is divided into training set, validation set, and test set. Models were fitted on the training set only, and the validation data were used to select the tuning parameters.
- The test error (the mean-squared error) was computed on the test set.

Simulation Example 1, 2 and 3

- **Simulation example 1:** 50 data sets were simulated consisting of 20/20/200 observations and 8 predictors:

$$\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0), \quad \sigma = 3, \quad \text{and} \quad \text{cov}(\boldsymbol{x}_i, \boldsymbol{x}_j) = (0.5)^{|i-j|}$$

- **Simulation example 2:** Same as example 1, except $\beta_j = 0.85$ for all j.
- **Simulation example 3:** 50 data sets were simulated consisting of 100/100/400 observations and 40 predictors:

$$\boldsymbol{\beta} = \left(\underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10}, \underbrace{0, \ldots, 0}_{10}, \underbrace{2, \ldots, 2}_{10}\right), \quad \sigma = 15,$$

$$\text{cor}(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0.5 \quad \text{for all } i, j = 1, \ldots, 40.$$

Simulation Example 4

- **Simulation example 4:** 50 data sets were simulated consisting of 50/50/400 observations and 40 predictors:

$$\boldsymbol{\beta} = \left(\underbrace{3, \ldots, 3}_{15}, \underbrace{0, \ldots, 0}_{25} \right), \quad \sigma = 15.$$

$$\boldsymbol{x}_i = Z_1 + \epsilon_i^x, \quad Z_1 \sim \mathcal{N}(0,1), \quad i = 1, \ldots, 5.$$

$$\boldsymbol{x}_i = Z_2 + \epsilon_i^x, \quad Z_2 \sim \mathcal{N}(0,1), \quad i = 6, \ldots, 10.$$

$$\boldsymbol{x}_i = Z_3 + \epsilon_i^x, \quad Z_3 \sim \mathcal{N}(0,1), \quad i = 11, \ldots, 15.$$

$$\boldsymbol{x}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1), \quad i = 16, \ldots, 40.$$

$$\epsilon_i^x \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,0.01), \quad i = 1, \ldots, 15.$$

Simulated Examples - Median MSE

Method	**Ex.1**	**Ex.2**	**Ex.3**	**Ex.4**
Ridge	4.49 (0.46)	2.84 (0.27)	39.5 (1.80)	64.5 (4.78)
Lasso	3.06 (0.31)	3.87 (0.38)	65.0 (2.82)	46.6 (3.96)
Elastic Net	2.51 (0.29)	3.16 (0.27)	56.6 (1.75)	34.5 (1.64)
Naive Elastic Net	5.70 (0.41)	2.73 (0.23)	41.0 (2.13)	45.9 (3.72)

Table: Median MSE for the simulated examples and 4 methods.

- Elastic Net is more accurate than the LASSO in all four examples, even when the LASSO is significantly more accurate than Ridge regression.
- The Naive Elastic Net performs very poorly with the highest mean-squared error in Example 1. In Example 2 and 3, it behaves very similar to Ridge regression, and in Example 4 it behaves similar to the LASSO.

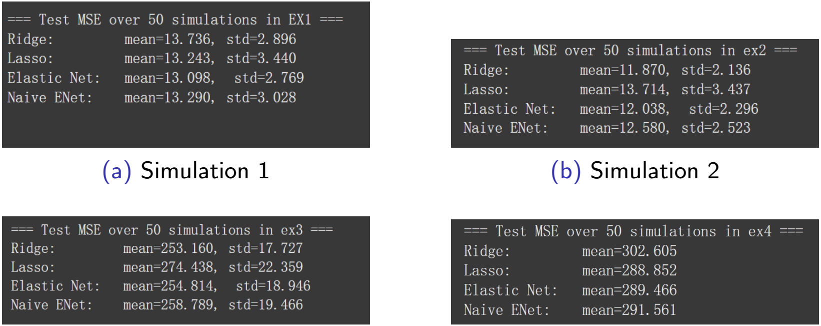
Simulated Examples - Variable Selection

Method	Ex. 1	Ex. 2	Ex. 3	Ex. 4
Lasso	5	6	24	11
Elastic Net	6	7	27	16

Table: Median number of non-zero coefficients

- Elastic Net selects more predictors than the LASSO due to the grouping effect.
- Elastic Net behaves like the ideal model in Example 4, where grouped selection is needed.
- Therefore, the Elastic Net has the additional ability to perform grouped variable selection, which makes it a better variable selection method than the LASSO.

Test MSE over 50 simulations in 4 simulations



```
=== Test MSE over 50 simulations in EX1 ===
Ridge:          mean=13.736，std=2.896
Lasso:          mean=13.243，std=3.440
Elastic Net:    mean=13.098，  std=2.769
Naive ENet:     mean=13.290，std=3.028
```

(a) Simulation 1

```
=== Test MSE over 50 simulations in ex2 ===
Ridge:          mean=11.870，std=2.136
Lasso:          mean=13.714，std=3.437
Elastic Net:    mean=12.038，  std=2.296
Naive ENet:     mean=12.580，std=2.523
```

(b) Simulation 2

```
=== Test MSE over 50 simulations in ex3 ===
Ridge:          mean=253.160，std=17.727
Lasso:          mean=274.438，std=22.359
Elastic Net:    mean=254.814，  std=18.946
Naive ENet:     mean=258.789，std=19.466
```

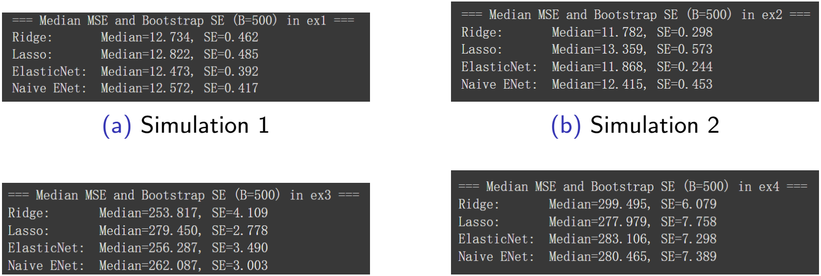
(c) Simulation 3

```
=== Test MSE over 50 simulations in ex4 ===
Ridge:          mean=302.605
Lasso:          mean=288.852
Elastic Net:    mean=289.466
Naive ENet:     mean=291.561
```

(d) Simulation 4

Figure: Test MSE over 50 simulations

Median MSE and Bootstrap SE (B=500) in 4 simulations



```
=== Median MSE and Bootstrap SE (B=500) in ex1 ===
Ridge:        Median=12.734, SE=0.462
Lasso:        Median=12.822, SE=0.485
ElasticNet:   Median=12.473, SE=0.392
Naive ENet:   Median=12.572, SE=0.417
```

(a) Simulation 1

```
=== Median MSE and Bootstrap SE (B=500) in ex2 ===
Ridge:        Median=11.782, SE=0.298
Lasso:        Median=13.359, SE=0.573
ElasticNet:   Median=11.868, SE=0.244
Naive ENet:   Median=12.415, SE=0.453
```

(b) Simulation 2

```
=== Median MSE and Bootstrap SE (B=500) in ex3 ===
Ridge:        Median=253.817, SE=4.109
Lasso:        Median=279.450, SE=2.778
ElasticNet:   Median=256.287, SE=3.490
Naive ENet:   Median=262.087, SE=3.003
```

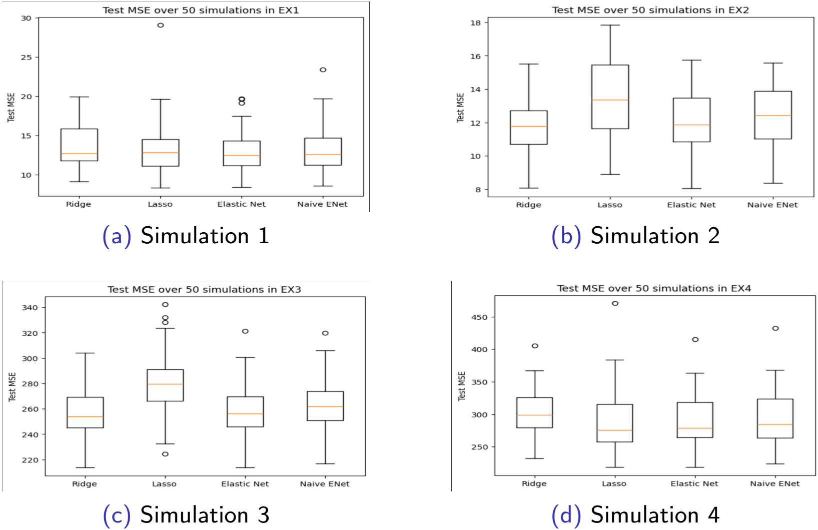
(c) Simulation 3

```
=== Median MSE and Bootstrap SE (B=500) in ex4 ===
Ridge:        Median=299.495, SE=6.079
Lasso:        Median=277.979, SE=7.758
ElasticNet:   Median=283.106, SE=7.298
Naive ENet:   Median=280.465, SE=7.389
```

(d) Simulation 4

Figure: Median MSE and Bootstrap SE (B=500)

the boxplot of MSE



(a) Simulation 1

(b) Simulation 2

(c) Simulation 3

(d) Simulation 4

Figure: Median MSE and Bootstrap SE (B=500)

Conclusion

- The LASSO can select at most n predictors in the $p > n$ case and cannot perform grouped selection. Furthermore, ridge regression usually has a better prediction performance than the LASSO when there are high correlations between predictors in the $n > p$ case.
- The Elastic Net can produce a sparse model with good prediction accuracy, while selecting group(s) of strongly correlated predictors. It can also potentially select all p predictors in all situations.
- A new algorithm called LARS-EN can be used for computing elastic net regularization paths efficiently, similar to the LARS algorithm for LASSO.
- The Elastic Net has two tuning parameters as opposed to one tuning parameter like the LASSO, which can be selected using a training and validation set.
- Simulation results indicate that the Elastic Net dominates the LASSO, especially under collinearity.