‘Regularization and variable selection via the elastic net’
Hui Zou, Trevor Hastie

Presented by: James Pope and Michal Kozlowski
Background

• Ordinary Least Squares:

\[ \hat{\beta}^{OLS} = \arg \min \| y - X\beta \|^2 \]

• Two important aspects to evaluate quality of model:
  • Accuracy of prediction on future data – good generalisation
  • Interpretation of the model – simpler model often preferred
Background

Bias – Variance trade-off
Ridge Regression and LASSO

• Penalization methods to resolve this:
  • Ridge Regression – minimisation of residual sum of squares, bound on the $L_2$ norm

$$\hat{\beta}_{ridge} = \arg \min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$$

• Lasso – penalised least squares, imposing $L_1$ penalty – shrinkage and variable selection at once

$$\hat{\beta}_{lasso} = \arg \min_{\beta} \|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$
Ridge Regression and LASSO
Ridge Regression and LASSO

- LASSO is limited however:
  - In the p>n case, the lasso selects at most n variables before it saturates.
  - If there is a group of highly correlated independent variables, LASSO chooses one and discards the others.
Naïve Elastic Net

• At high dimensional data, there can be variables which are highly correlated with each other – multicollinearity.

• Methods up to now failed to perform grouped selection.

• Can be thought of as a hybrid of ridge and LASSO:

\[
\hat{\beta}_{elastic} = \arg\min_{\beta} \| y - X\beta \|_2^2 + \lambda_2 \|\beta\|_2^2 + \lambda_1 \|\beta\|_1
\]
Naïve Elastic Net

If $\alpha = 1$, we have ridge

If $\alpha = 0$, we have LASSO

If $0 < \alpha < 1$ we have elastic net

In this figure $\alpha = 0.5$

\[
J(\beta) = \alpha \|\beta\|^2 + (1-\alpha) \|\beta\|_1
\]

\[
\alpha = \frac{\lambda_2}{\lambda_1 + \lambda_2}
\]
Naïve Elastic Net is insufficient

$$\hat{\beta}_{\text{elastic}} = \arg \min \| y - X\beta \|_2^2 + \lambda_2 \| \beta \|_2^2 + \lambda_1 \| \beta \|_1$$

- The Naïve EN estimator is a two-stage procedure:
  - For each fixed $\lambda_2$, we first find the ridge coefficient $\lambda_2 \| \beta \|_2^2$ and then perform LASSO shrinkage $\lambda_1 \| \beta \|_1$

- We effectively perform double shrinkage!

- Extra bias, no variance reduction...
Elastic Net with scaling correction

- Correct it, by introducing the quadratic penalty. Given augmented data \((\mathbf{y}^*, \mathbf{X}^*)\), naïve net solves the lasso problem:

\[
\hat{\beta}^* = \arg \min_{\beta^*} |\mathbf{y}^* - \mathbf{X}^* \beta^*|^2 + \frac{\lambda_1}{\sqrt{1 + \lambda_2}} |\beta^*|_1
\]

- Corrected estimates defined by:

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

- Naïve net can be shown to be \(\hat{\beta} (\text{naïve elastic net}) = \{1/\sqrt{1 + \lambda_2}\} \hat{\beta}^*\); thus:

\[
\hat{\beta} (\text{elastic net}) = (1 + \lambda_2) \hat{\beta} (\text{naïve elastic net})
\]

- That way, we retain the grouping effect, and perform shrinkage once!
Elastic Net

- Elastic Net estimate

L1 Norm: Sparsity inducing
L2 Norm: Weight sharing
L1 + L2 Norm: Compromise... Two parameters...
Univariate Soft Thresholding

• Consider that when $\lambda_2 = 0$, elastic net effectively becomes LASSO

• Special case of elastic net, when $\lambda_2 \to \infty$

• Elastic net applies soft thresholding on univariate coefficients.

• UST ignores the dependence between predictors, and treats them as independent variables.

• Used for significance analysis of microarrays (Tusher, 2001)

• Elastic net bridges the gap between the LASSO and UST.
LARS-EN – Elastic Net Computation

• First proposed as LARS by Efron et al., 2004, expanded in this paper
• Resembles forward stepwise regression.
• The solution path is piecewise linear.
• Given a fixed $\lambda_2$, we can solve the entire ENet solution path!
  • At step $k$, update Cholesky factorisation of $X_{A_{k-1}}^TX_{A_{k-1}} + \lambda_2I$
  • Record the non-zero coefficients at each LARS-EN step

• Not necessary to run it all the way to the end for $p>>n$!
(early stopping)
Summary on Elastic Net

• Ridge regression fails to choose any variables. Either all or none.

• LASSO solves this, but disregards clusters of highly correlated data. It favours a single variable.

• Elastic net is a good compromise between the two.
Elastic Net Evaluation

• Study: Prostate Cancer Data
• Simulation A: To evaluate prediction performance (Lasso)
• Simulation B: To evaluate feature selection and grouping performance
• Study: Leukaemia Data (p>>n)
Study: Prostate Cancer Data

• Eight clinical measures as predictors
  1. Cancer volume
  2. Prostate weight
  3. Age
  4. Benign prostatic hyperplasia
  5. Seminal vesicle invasion
  6. Capsular penetration
  7. Gleason score
  8. Gleason score 4/5

• The response is the prostate-specific antigen (l- PSA)

• Data divided into training (67 observations) and test (30 observations)
  • Compared methods OLS, Ridge, Lasso, Naïve Elastic Net, Elastic Net
  • Used tenfold CV on training to determine model tuning parameters.
Study: Methods Prediction Compared

- For $\lambda = 1000$, this essentially becomes UST (Why 1000?)
  - (previously $\beta_{\hat{\text{hat}}}(\infty)$, i.e. where $\lambda_2 \rightarrow \infty$)
- Elastic Net dominates Lasso, Lasso hurt by high correlation (e.g. 7, and 8)

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameter(s)</th>
<th>Test mean-squared error</th>
<th>Variables selected</th>
</tr>
</thead>
<tbody>
<tr>
<td>OLS</td>
<td></td>
<td>0.586 (0.184)</td>
<td>All</td>
</tr>
<tr>
<td>Ridge regression</td>
<td>$\lambda = 1$</td>
<td>0.566 (0.188)</td>
<td>All</td>
</tr>
<tr>
<td>Lasso</td>
<td>$s = 0.39$</td>
<td>0.499 (0.161)</td>
<td>(1,2,4,5,8)</td>
</tr>
<tr>
<td>Naïve elastic net</td>
<td>$\lambda = 1$, $s = 1$</td>
<td>0.566 (0.188)</td>
<td>All</td>
</tr>
<tr>
<td>Elastic net</td>
<td>$\lambda = 1000$, $s = 0.26$</td>
<td>0.381 (0.105)</td>
<td>(1,2,5,6,8)</td>
</tr>
</tbody>
</table>
Simulation A

• Purpose to show that elastic net dominates lasso.
  • Prediction accuracy
  • Variable selection

• Simulate using model:

\[ y = X \beta + \sigma \varepsilon, \; \varepsilon \sim N(0,1) \]

• Four Examples (first three from Lasso), training / validation / test

1. 50 data sets, 20/20/200, 8 predictors $\beta = \{3,1.5,0,0,2,0,0,0\}$, $\sigma = 3$
   Pairwise correlation between $x_i$ and $x_j$ corr($i,j$)=$0.5^{|i-j|}$

2. Same as 1, except $\beta = \{0.85,0.85,...,0.85\}$

3. 50 data sets, 100/100/400, 40 predictors, $\beta = \{10x0,10x2,10x0,10x2\}$, $\sigma = 15$
   Pairwise correlation between $x_i$ and $x_j$ corr($i,j$)=$0.5$
Simulation A: Example 4

• Example specifically designed to show grouping selection
  4. 50 data sets, 50/50/400 observations
     40 predictors $\beta = \{15 \times 3, 25 \times 0\}$, $\sigma = 15$
     $x_i = Z_1 + \text{noise}, \ i = 1, \ldots, 5$
     $x_i = Z_2 + \text{noise}, \ i = 6, \ldots, 10$
     $x_i = Z_3 + \text{noise}, \ i = 11, \ldots, 15$
     $x_i = \text{noise}, \ i = 16, \ldots, 40$

• In this model there are three important groups
  • Ideally would select 15 features and reject 25 noise features
Simulation A: Results

- Elastic net dominates Lasso (under collinearity).
- But in some cases worse than Ridge or Naïve!

Table 2. Median mean-squared errors for the simulated examples and four methods based on 50 replications†

<table>
<thead>
<tr>
<th>Method</th>
<th>Example 1</th>
<th>Example 2</th>
<th>Example 3</th>
<th>Example 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lasso</td>
<td>3.06 (0.31)</td>
<td>3.87 (0.38)</td>
<td>65.0 (2.82)</td>
<td>46.6 (3.96)</td>
</tr>
<tr>
<td>Elastic net</td>
<td>2.51 (0.29)</td>
<td>3.16 (0.27)</td>
<td>56.6 (1.75)</td>
<td>34.5 (1.64)</td>
</tr>
<tr>
<td>Ridge regression</td>
<td>4.49 (0.46)</td>
<td>2.84 (0.27)</td>
<td>39.5 (1.80)</td>
<td>64.5 (4.78)</td>
</tr>
<tr>
<td>Naïve elastic net</td>
<td>5.70 (0.41)</td>
<td>2.73 (0.23)</td>
<td>41.0 (2.13)</td>
<td>45.9 (3.72)</td>
</tr>
</tbody>
</table>

†The numbers in parentheses are the corresponding standard errors (of the medians) estimated by using the bootstrap with \( B = 500 \) resamplings on the 50 mean-squared errors.
Simulation A (Fig.4): Error Distribution

- Paper does not comment, elastic net lots of outliers
Simulation A: Results Sparse Solutions

• Elastic net selects more features than Lasso due to grouping/correlation
  • Example 4 is close to optimal

<table>
<thead>
<tr>
<th>Method</th>
<th>Example 1</th>
<th>Example 2</th>
<th>Example 3</th>
<th>Example 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lasso</td>
<td>5</td>
<td>6</td>
<td>24</td>
<td>11</td>
</tr>
<tr>
<td>Elastic net</td>
<td>6</td>
<td>7</td>
<td>27</td>
<td>16</td>
</tr>
</tbody>
</table>

Non-zero $\beta$: 3 3 20 15
Simulation B: Grouping Selection (Lasso)

• Given two independent variable $Z_1$, and $Z_2$, the response variable $y$

\[
Y \sim N(Z_1 + 0.1 Z_2, 1)
\]

• Create $X$ from two groups (note that $x_2$ and $x_5$ are negated)

\[
\begin{align*}
  x_1 &= Z_1 + \text{noise} \\
  x_2 &= -Z_1 + \text{noise} \\
  x_3 &= Z_1 + \text{noise} \\
  x_4 &= Z_2 + \text{noise} \\
  x_5 &= -Z_2 + \text{noise} \\
  x_6 &= Z_2 + \text{noise}
\end{align*}
\]

• Within group correlation roughly 1
• Between group correlation roughly 0
• Important variates are the $Z_1$-group, how well do EN and Lasso perform?
Simulation B: EN vs Lasso Solution Paths

- Recall good grouping will set coefficients to similar values.
- Lasso very unstable.
- Elastic Net selects same (absolute) coefficient for the $Z_1$-group

![Graph showing Lasso and Elastic Net solution paths](image-url)
Great, so far elastic net dominates Lasso (prediction and feature selection). What about feature selection when $p >> n$ (e.g. $p = 10000$ and $n = 100$)?

When $p >> n$, good classification method should have:

A. Gene selection built into the procedure

B. Not be limited to the fact at $p >> n$ (Lasso is limited to selecting $n$)

C. Genes sharing same biological pathway (correlated?), should group into model

Current methods arbitrarily select one of the related gene predictors.
Microarray Classification and Gene Selection
Other Methods

• Lasso:
  • Good at A (built in selection)
  • Fails B (handle p>>n) and C (grouped selection)

• Support Vector Machine (Guyon et al., 2002) and Penalized Logistic Regression (Zhu and Hastie, 2004) fail at C.
  • Use either univariate ranking or recursive feature elimination
• Leukaemia data 7192 genes and 72 samples.  
  • Training set 38 samples, of which 27 are Type 1 and 11 are Type 2  
  • Test set 34  
• Goal: Construct rule to predict $y = 0-1$ response (Type1=0, Type2=1)  
• Pre-screening to select 1000 genes from 7192 before elastic net  
  • Used t-statistic to make “computation more manageable”  
  • Authors claim does not affect results (NEED TO UNDERSTAND MORE)  
• Used ten-fold CV to select tuning parameters.
Leukaemia Classification

- Early stopping strategy $k = 200$ (iterations). Less computational cost (?)

- Full path (i.e. when using $s$ as fraction of $L_1$-norm to know when to stop).

$$\lambda = 0.01$$
Leukaemia Classification Summary

- Elastic net selects more than n predictors (training set determines n).
- Lasso would only be able to select a maximum of 38 genes.

**Table 4. Summary of the leukaemia classification results**

<table>
<thead>
<tr>
<th>Method</th>
<th>Tenfold CV error</th>
<th>Test error</th>
<th>Number of genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Golub</td>
<td>3/38</td>
<td>4/34</td>
<td>50</td>
</tr>
<tr>
<td>Support vector machine–recursive feature elimination</td>
<td>2/38</td>
<td>1/34</td>
<td>31</td>
</tr>
<tr>
<td>Penalized logistic regression–recursive feature elimination</td>
<td>2/38</td>
<td>1/34</td>
<td>26</td>
</tr>
<tr>
<td>Nearest shrunken centroids</td>
<td>2/38</td>
<td>2/34</td>
<td>21</td>
</tr>
<tr>
<td>Elastic net</td>
<td>3/38</td>
<td>0/34</td>
<td>45</td>
</tr>
</tbody>
</table>
Leukaemia/Gene: Solution Paths

• Not clear that the 45 selected genes are properly grouped, though some coefficients seem to converge.
• No evidence that better than Golub (or others) “...best classification and internal gene selection”.

Number of Genes Selected

k=82
45 Selected
Discussion

• Elastic appears to be better than Lasso for prediction and feature selection.

• Other issues are unclear.
  • Elastic net is better than or comparable to Ridge Regression for predictions.
  • Elastic net is better at grouping than other grouping methods (sans Lasso).
References

• Hui Zou and Trevor Hastie, ‘Regularization and variable selection via the elastic net’, 2005
• Derek Kane, ‘Data Science - Part XII - Ridge Regression, LASSO, and Elastic Nets’, 2015
• Vivian S. Zhang, ‘Ridge regression, lasso and elastic net’, 2014
• http://www.ds100.org/sp17/assets/notebooks/linear_regression/Regularization.html
• https://stats.stackexchange.com/questions/123205/relation-between-the-tuning-parameter-lambda-parameter-estimates-beta-i-a
• https://codingstartups.com/practical-machine-learning-ridge-regression-vs-lasso/
• https://datascience.stackexchange.com/questions/361/when-is-a-model-underfitted