
Robert Tibshirani, Stanford University

Georgia Statistics Day, 2015
An exciting time for statisticians!

- BIG DATA is everywhere
- Data Science is emerging as a new field
- Enrollment in Statistics/Machine Learning/Data Science classes are way up!
- Graduates are in **big demand**: Google, FaceBook, Twitter, LinkedIn
For Statisticians: 15 minutes of fame

• 2009: “I keep saying the sexy job in the next ten years will be statisticians.” Hal Varian, Chief Economist Google

Sexiest man alive?
Don’t believe everything on the Web
Outline

Focus is on **predictive modeling**

1. **Prediction**: Example application of the lasso for cancer diagnosis
2. **Inference** (How to obtain p-values and confidence intervals) after fitting Forward Stepwise Regression or the Lasso: **New tools for post-selective inference**
A Cancer detection problem

- I am currently working in a cancer diagnosis project with co-workers at Stanford; Livia Eberlin (PI) —PostDoc (Chemistry); Richard Zare (Chemistry) and George Poulsides (Surgery)
- They have collected samples of tissue from a number of patients undergoing surgery for stomach cancer.
Livia Eberlin

Richard Zare

George Poulisides
Left in patient

Extracted part

Normal margin

Cancer

Stromal

Epithelial

Robert Tibshirani, Stanford University

Lasso
The challenge

- Build a classifier than can distinguish three kinds of tissue: normal epithelial, normal stromal and cancer.
- Such a classifier could be used to assist surgeons in determining, in real time, whether they had successfully removed all of the tumor. Current pathologist error rate for the real-time call can be as high as 20%.
Technology to the rescue!
DESI (Desorption electrospray ionization)
An electrically charged “mist” is directed at the sample; surface ions are freed and enter the mass spec.
The data for one patient

Spectrum sampled at 11,000 m/z values
Details

- 20 patients, each contributing a sample of epithelial, stromal and cancer tissue.
- Labels determined after 2 weeks of testing in pathology lab.
- At each pixel in the image, the intensity of metabolites is measured by DESI. Peaks in the spectrum representing different metabolites.
- The spectrum has been finely sampled, with the intensity measured at about 11,000 \( m/z \) sites across the spectrum, for each of about 8000 pixels.
The overall data

<table>
<thead>
<tr>
<th>Patient</th>
<th>Label of pixel:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1=Epithelial</td>
</tr>
<tr>
<td>2</td>
<td>2= Cancer</td>
</tr>
<tr>
<td>3</td>
<td>3=Stromal</td>
</tr>
</tbody>
</table>

--- 11,000 m/z sites ---

Robert Tibshirani, Stanford University
Lasso
Selected negative ion mode DESI-MS ion images of sample GC727.
What we need to tackle this problem

- A statistical classifier (algorithm) that sorts through the large number of features, and finds the most informative ones: a \textit{sparse} set of features.
- \textbf{the Lasso!}
Review of the Lasso

The **Lasso** is an estimator defined by the following optimization problem:

$$\min_{\beta_0, \beta} \frac{1}{2} \sum_i (y_i - \beta_0 - \sum_j x_{ij}\beta_j)^2 \quad \text{subject to} \quad \sum |\beta_j| \leq s$$

- Penalty $\Rightarrow$ sparsity (feature selection)
- Convex problem (good for computation and theory)
- *Ridge regression* uses penalty $\sum j \beta_j^2 \leq s$ and does not yield sparsity
Prostate cancer example
$N = 88, \ p = 8$. Predicting log-PSA, in men after prostate cancer surgery
Why does the lasso give a sparse solution?
Back to our problem

- $K = 3$ classes (epithelial, stromal, cancer): multinomial model

\[
\log \frac{Pr(Y_i = k|x)}{\sum_k Pr(Y_i = k|x)} = \beta_{0k} + \sum_j x_{ij}\beta_{jk}, \quad k = 1, 2, \ldots K
\]

Here $x_{ij}$ is height of spectrum for sample $i$ at $jth$ $m/z$ position

- We replace the least squares objective function by the multinomial log-likelihood
- Add lasso penalty $\sum |\beta_j| \leq s$; optimize, using cross-validation to estimate best value for budget $s$.
- yields a pixel classifier, and also reveals which $m/z$ sites are informative.
Fast computation is essential

- Our lab has written a open-source R language package called \texttt{glmnet} for fitting lasso models. Available on CRAN. Largely written in \texttt{FORTRAN}!!!!!!
- It is very fast- can solve the current problem in a few minutes on a PC. Some built-in parallelization too.
- Not “off-the-shelf”: Many clever computational tricks were used to achieve the impressive speed.
- Lots of features- Gaussian, Logistic, Poisson, Survival models; elastic net; grouping; parameter constraints; Available in R and Matlab.

Jerry Friedman
Trevor Hastie
## Results

**Cross-validation** - min at 129 peaks; overall error rate = 4.2%

<table>
<thead>
<tr>
<th></th>
<th>Epi</th>
<th>Canc</th>
<th>Strom</th>
<th>Prop correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epi</td>
<td>3277.00</td>
<td>80.00</td>
<td>145.00</td>
<td>0.94</td>
</tr>
<tr>
<td>Canc</td>
<td>73.00</td>
<td>5106.00</td>
<td>13.00</td>
<td>0.98</td>
</tr>
<tr>
<td>Strom</td>
<td>79.00</td>
<td>86.00</td>
<td>2409.00</td>
<td>0.94</td>
</tr>
</tbody>
</table>

**Test set**: overall error rate = 5.7%

<table>
<thead>
<tr>
<th></th>
<th>Epi</th>
<th>Canc</th>
<th>Strom</th>
<th>Prop correct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Epi</td>
<td>1606.00</td>
<td>149.00</td>
<td>19.00</td>
<td>0.91</td>
</tr>
<tr>
<td>Canc</td>
<td>23.00</td>
<td>1622.00</td>
<td>5.00</td>
<td>0.98</td>
</tr>
<tr>
<td>Strom</td>
<td>67.00</td>
<td>5.00</td>
<td>1222.00</td>
<td>0.94</td>
</tr>
</tbody>
</table>
Cross-validated estimates of class probabilities
<table>
<thead>
<tr>
<th>Peak #</th>
<th>m/z value</th>
<th>Epi</th>
<th>Canc</th>
<th>Strom</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>101.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>107.5</td>
<td>0.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>110.5</td>
<td>0.44</td>
<td>0.56</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>123.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>132.5</td>
<td>0.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>134.5</td>
<td>0.10</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>135.5</td>
<td>0.13</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>137.5</td>
<td>-0.01</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>145.5</td>
<td>-0.71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>146.5</td>
<td>-0.41</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>151.5</td>
<td>-0.15</td>
<td>0.35</td>
<td>-0.25</td>
</tr>
<tr>
<td>12</td>
<td>157.5</td>
<td>-0.07</td>
<td>-0.17</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>170.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>171.5</td>
<td>0.05</td>
<td>0.14</td>
<td>0.54</td>
</tr>
<tr>
<td>15</td>
<td>174.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>175.5</td>
<td></td>
<td>0.14</td>
<td>0.54</td>
</tr>
<tr>
<td>17</td>
<td>179.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>188.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>212.5</td>
<td>-0.01</td>
<td>-0.13</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>214.5</td>
<td>-0.01</td>
<td>-0.13</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>215.5</td>
<td>-1.17</td>
<td>-1.07</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>222.5</td>
<td>0.29</td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>224.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>225.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>231.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>244.5</td>
<td>-0.31</td>
<td>-0.41</td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>247.5</td>
<td>0.21</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>258.5</td>
<td>0.21</td>
<td>-0.14</td>
<td>-0.24</td>
</tr>
<tr>
<td>29</td>
<td>270.5</td>
<td>-0.14</td>
<td>-0.24</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>278.5</td>
<td>0.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>279.5</td>
<td>0.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>280.5</td>
<td>0.39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>285.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>289.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>293.5</td>
<td>0.25</td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>297.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>299.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>301.5</td>
<td>-0.45</td>
<td>0.21</td>
<td>0.11</td>
</tr>
<tr>
<td>39</td>
<td>312.5</td>
<td>0.10</td>
<td>-0.44</td>
<td>-0.24</td>
</tr>
<tr>
<td>40</td>
<td>322.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>41</td>
<td>324.5</td>
<td>0.06</td>
<td></td>
<td></td>
</tr>
<tr>
<td>42</td>
<td>325.5</td>
<td>0.03</td>
<td>-0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>43</td>
<td>333.5</td>
<td>0.82</td>
<td>0.55</td>
<td></td>
</tr>
<tr>
<td>44</td>
<td>340.5</td>
<td>-0.02</td>
<td>-0.07</td>
<td></td>
</tr>
<tr>
<td>45</td>
<td>341.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>46</td>
<td>347.5</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>47</td>
<td>349.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>48</td>
<td>353.5</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>49</td>
<td>355.5</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
patient= 105  at m/z= 788.6

true

predicted

epithelial
cancer
stromal
Other approaches

- **Support vector machines**: classification error was a little higher than lasso; doesn’t give a sparse solution easily
- **Deep learning** (with help from a student of Geoff Hinton): reported that it didn’t work any better than lasso; thought that non-linearities were likely unimportant for this problem, and sparsity was more important
A challenge

- “Abstentions”: sometimes a classifier should not make a prediction; instead it should say “I don’t know” For example, when the query feature vector is far away from the training set features.
- This problem happened in some tests of our system
- Can’t rely on the fact that the largest posterior probability will be close to $1/K$ ($K=$ number of classes):
Inference for Forward Stepwise regression and the Lasso

Richard Lockhart
Simon Fraser University
Vancouver
PhD, Student of David Blackwell,
Berkeley, 1979

Jonathan Taylor
Stanford University
PhD Student of Keith Worsley, 2001

Ryan Tibshirani
CMU. PhD student of Taylor
2011

Rob Tibshirani
Stanford

Matching Results
from picadilo.com

Robert Tibshirani, Stanford University

Lasso
Richard Lockhart
Simon Fraser University
Vancouver
PhD, Student of David Blackwell, Berkeley, 1979

Ryan Tibshirani
CMU

Jonathan Taylor
Stanford University
PhD Student of Keith Worsley, 2001

Rob Tibshirani
Stanford

81%

Top matches from picadilo.com

71%
Richard Lockhart  
Simon Fraser University  
Vancouver  
PhD, Student of David Blackwell, Berkeley, 1979

Ryan Tibshirani  
CMU

81%

Jonathan Taylor  
Stanford University  
PhD Student of Keith Worsley, 2001

Rob Tibshirani  
Stanford

71%  69%

Robert Tibshirani, Stanford University  
Lasso
Confidence— the strength of evidence— matters!
• Regression problem: We observe \( n \) feature-response pairs \((x_i, y_i)\), where \( x_i \) is a \( p \)-vector and \( y_i \) is real-valued.

• Let \( x_i = (x_{i1}, x_{i2}, \ldots x_{ip}) \)

• Consider a linear regression model:

\[
y_i = \beta_0 + \sum_j x_{ij} \beta_j + \epsilon_i
\]

where \( \epsilon_i \) is an error term with mean zero. \( \beta_j \) is the weight given feature \( j \)

• Least squares fitting is defined by

\[
\minimize_{\beta_0, \beta} \frac{1}{2} \sum_i (y_i - \beta_0 - \sum_j x_{ij} \beta_j)^2
\]
Forward Stepwise regression

- start with fit equal to the mean $\bar{y}$
- at each step, add to model predictor that most decreases training error
- continue until $\min(n, p)$ predictors are in the model
- use a criterion like cross-validation to select best member in the path of models.
The Lasso is an estimator defined by the following optimization problem:

\[
\begin{align*}
\text{minimize} \quad & \frac{1}{2} \sum_i (y_i - \beta_0 - \sum_j x_{ij} \beta_j)^2 \\
\text{subject to} \quad & \sum_j |\beta_j| \leq s
\end{align*}
\]

- Penalty \(\Rightarrow\) sparsity (feature selection)
- Convex problem (good for computation and theory)
The Lasso—continued

Equivalent Lagrangian form:

$$\min_{\beta_0, \beta} \frac{1}{2} \sum_i (y_i - \beta_0 - \sum_j x_{ij} \beta_j)^2 + \lambda \sum |\beta_j|$$

- We choose \( s \) or \( \lambda \) by cross-validation.
Prostate cancer example

\( N = 88, p = 8 \). Predicting log-PSA, in men after prostate cancer surgery
Illustration of the post-selection inference problem

- Suppose that we apply forward stepwise regression to the prostate data, and our software prints out the usual p-values from the F-test at each stage.
- The test statistic is the reduction in sum of squares divided by an estimate of the error variance.

Results:

<table>
<thead>
<tr>
<th></th>
<th>P-value from F-test</th>
</tr>
</thead>
<tbody>
<tr>
<td>lcavol</td>
<td>0.000</td>
</tr>
<tr>
<td>lweight</td>
<td>0.000</td>
</tr>
<tr>
<td>svi</td>
<td>0.047</td>
</tr>
<tr>
<td>lbph</td>
<td>0.047</td>
</tr>
<tr>
<td>pgg45</td>
<td>0.234</td>
</tr>
<tr>
<td>lcp</td>
<td>0.083</td>
</tr>
<tr>
<td>age</td>
<td>0.137</td>
</tr>
<tr>
<td>gleason</td>
<td>0.883</td>
</tr>
</tbody>
</table>

Do we believe these p-values?
Illustration of the post-selection inference problem: Forward stepwise-regression- chi-square statistic

$N = 100, p = 10$, true model null

Test is too liberal: for nominal size 5%, actual type I error is 39%. Can get proper $p$-values by sample splitting: but messy, loss of power
Post-selection inference

John Tukey held that too much emphasis in statistics was placed on statistical hypothesis testing (confirmatory data analysis); more emphasis needed to be placed on using data to suggest hypotheses to test.

"... confusing the two types of analyses and employing them on the same set of data can lead to systematic bias owing to the issues inherent in testing hypotheses suggested by the data."

We will show how to get valid post-selection p-values and confidence intervals for LAR, lasso, forward stepwise regression and many other procedures.
Our solution

- gives exact p-values that account for selection;
- applicable very generally— to forward stepwise regression, fixed-λ lasso and beyond— e.g. principal components analysis,
Example: Forward Stepwise regression

<table>
<thead>
<tr>
<th></th>
<th>FS, naive</th>
<th>FS, adjusted</th>
</tr>
</thead>
<tbody>
<tr>
<td>lcavol</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>lweight</td>
<td>0.000</td>
<td>0.012</td>
</tr>
<tr>
<td>svi</td>
<td>0.047</td>
<td>0.849</td>
</tr>
<tr>
<td>lbph</td>
<td>0.047</td>
<td>0.337</td>
</tr>
<tr>
<td>pgg45</td>
<td>0.234</td>
<td>0.847</td>
</tr>
<tr>
<td>lcp</td>
<td>0.083</td>
<td>0.546</td>
</tr>
<tr>
<td>age</td>
<td>0.137</td>
<td>0.118</td>
</tr>
<tr>
<td>gleason</td>
<td>0.883</td>
<td>0.311</td>
</tr>
</tbody>
</table>

Table: Prostate data example: $n = 88, p = 8$. Naive and selection-adjusted forward stepwise sequential tests
Lasso with fixed-$\lambda$

HIV data: selection intervals for lasso with fixed tuning parameter $\lambda$. 

Robert Tibshirani, Stanford University
The polyhedral lemma

• Response vector \( y \sim N(\mu, \Sigma) \). Suppose we make a selection that can be written as

\[
Ay \leq b
\]

with \( A, b \) not depending on \( y \).

• Then for any vector \( \eta \)

\[
F_{\eta^\top \mu, \sigma^2 \eta^\top \eta}^\left[V^{-}, V^{+}\right](\eta^\top y) \{A_L, \lambda y \leq b_L, \lambda\} \sim \text{Unif}(0, 1)
\]

(truncated Gaussian distribution), where \( V^{-}, V^{+} \) are (computable) constants that are functions of \( \eta, A, b \).

• Result is **exact** (finite-sample) and works for any \( X! \)
\[
\mathcal{V}^-(y) + \mathcal{V}^+(y) = P_{\eta^\perp} y + \eta^T y = \{A y \leq b\}.
\]
The polyhedral lemma: some intuition

- Suppose that we have taken one step of forward stepwise regression, entering predictor $x_3$. We wish to make inferences about the coefficient of $x_3$.
- We can view the stepwise procedure as a competition among inner products: $x_3$ has a larger inner product with $y$ than the second place finisher, say $x_2$.
- More generally, the set of outcome vectors $y$ that would lead to the same results of the competition can be written in the form $Ay \leq b$. 

Robert Tibshirani, Stanford University

Lasso
• Now without selection, the inner product between $x_3$ and $y$ can have any value between $-\infty$ and $\infty$. **But after selection**, with orthogonal variables, we know it must be at least as large as the inner product between $x_2$ and $y$.

• **Conditioning on the results of the competition** leads to constraints on the size of the inner products, and these are computable from $A$ and $b$, namely $V_m$ and $V_p$.

• We carry out inference using the Gaussian distribution **truncated** to lie in $(V_m, V_p)$.
Polyhedral lemma again

\[
F_{\eta^T \mu, \sigma^2 \eta^T \eta} (\eta^T y) \mid \{A_L, \lambda y \leq b_L, \lambda\} \sim \text{Unif}(0, 1)
\]

- **Very powerful**: can be used to construct p-values and confidence intervals for parameters after any selection event of the form \(AY \leq b\).
- Can generalize to quadratic selections event
- With sampling, can obtain more powerful tests in exponential family: *Optimal inference after model selection* (Fithian, Sun, Taylor)
R package now available on CRAN

**selectiveInference**

Joshua Loftus, Jon Taylor, Ryan Tibshirani, Rob Tibshirani

- Forward Stepwise regression, including AIC stopping rule, categorical variables
  \[ \text{fsInf}(x,y) \]
- Fixed lambda Lasso
  \[ \text{fixedLassoInf}(x,y) \]
- Least angle regression
  \[ \text{larInf}(x,y) \]

These compute p-values and selection intervals
Ongoing work on selective inference

- LAR+forward stepwise (Taylor, Lockhart, Tibs times 2)
- Forward stepwise with grouped variables (Loftus and Taylor)
- **Sequential selection procedures** (G’Sell, Wager, Chouldechova, Tibs) JRSSB
- PCA (Choi, Taylor, Tibs)
- Marginal screening (Lee and Taylor)
- Many means problem (Reid, Taylor, Tibs)
- Asymptotics (Tian and Taylor)
- Bootstrap versions (Ryan Tibshirani+friends)
PCA example

Scree Plot of (true) rank = 2

Singular values (SNR=1.5)

Singular values (SNR=0.23)

(p-values for right: 0.030, 0.064, 0.222, 0.286, 0.197, 0.831, 0.510, 0.185, 0.126.)
Some resources

available online for free

Courses: google → tibshirani → teaching
- MOOC (free Stanford Online course) on Statistical Learning - Jan 2016
- 2-day in-person course with Trevor Hastie- given twice a year (Started at CDC!)
Conclusions

• The Lasso is a method well-suited for today’s big data problems. The sparsity yields interpretable models and improves predictive accuracy.

• I have presented some new tools for post-selection inference for the lasso and related methods.

• Working with real scientists and real problems is fulfilling and sometimes suggests new statistical challenges.