The Lasso: An Application to Cancer Detection and Some New Tools for Selective Inference

Robert Tibshirani, Stanford University

Georgia Statistics Day, 2015

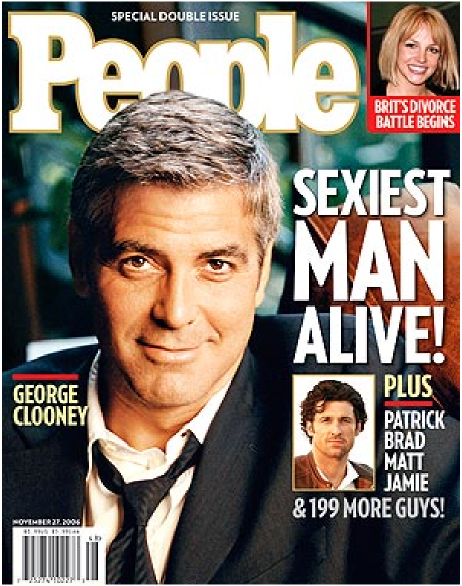
An exciting time for statisticians!

- 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

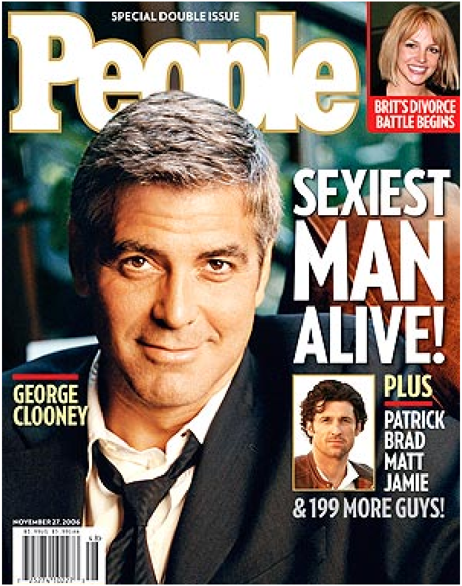
For Statisticians: 15 minutes of fame

- 2009: “ I keep saying the **sexy** job in the next ten years will be **statisticians**.” Hal Varian, Chief Economist Google
- 2012 “**Data Scientist**: The Sexiest Job of the 21st Century” Harvard Business Review

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**

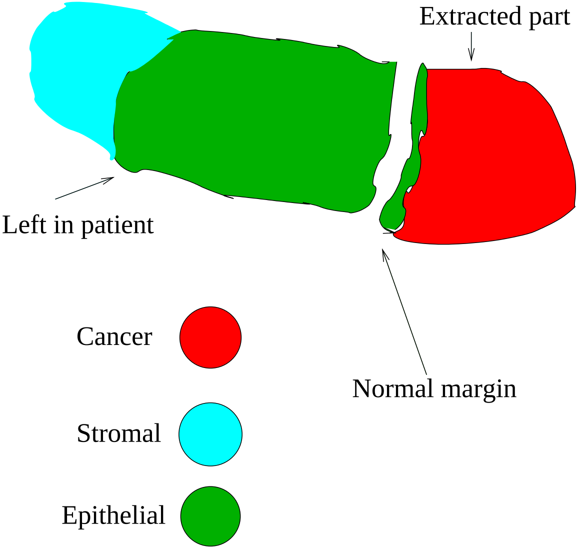
A Cancer detection problem

- 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)
- Eberlin et al (2014). "Molecular assessment of surgical-resection margins of gastric cancer by mass-spectrometric imaging". Proc. Nat. Acad. Sci.
- They have collected samples of tissue from a number of patients undergoing surgery for stomach cancer.

Livia
Eberlin

Richard
Zare

George
Poulsides



Extracted part

Left in patient

Normal margin

Cancer

Stromal

Epithelial

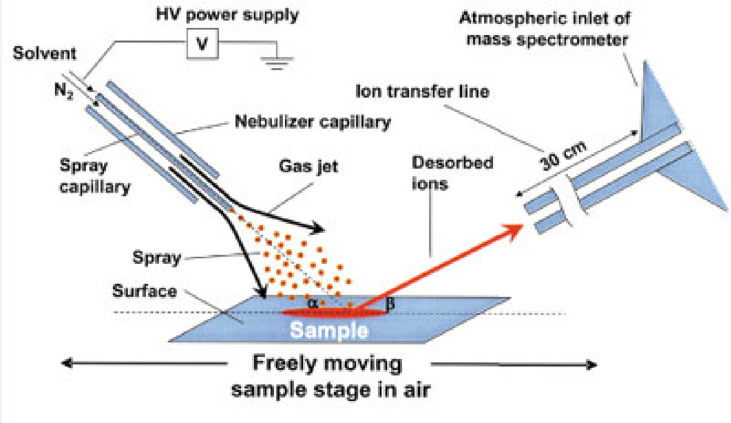
The challenge

- 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%.

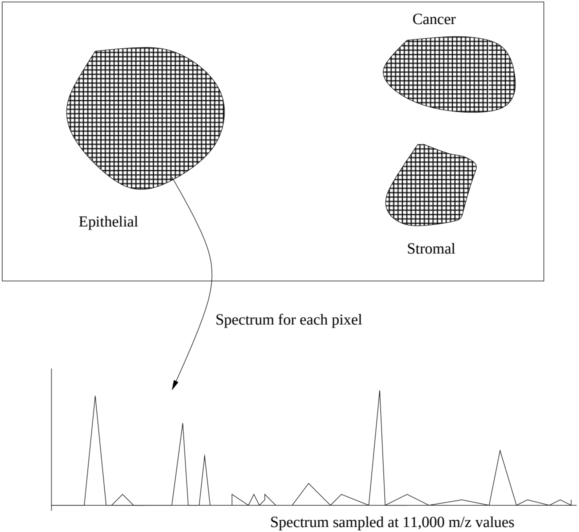
Technology to the rescue!

DESI (Desorption electrospray ionization)

An electrically charged "mist" is directed at the sample; surface ions are freed and enter the mass spec.



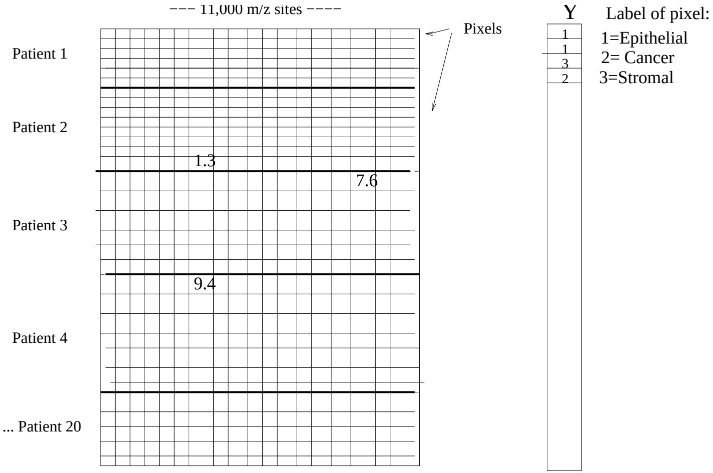
The data for one patient



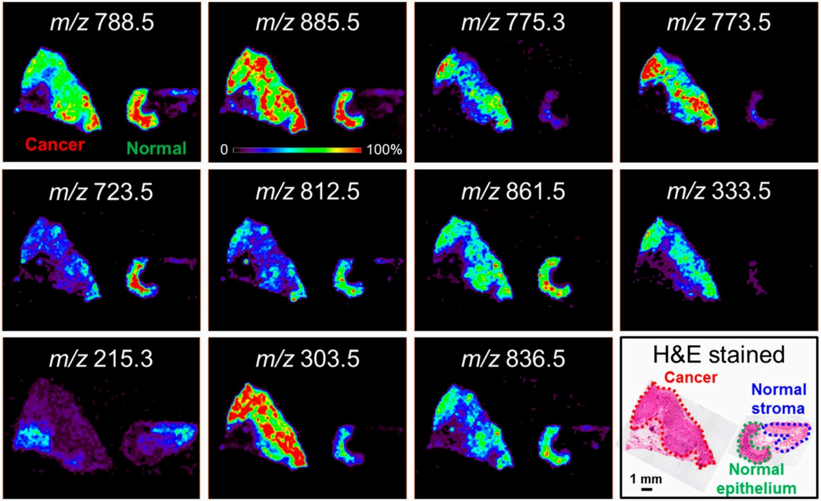
Details

- 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



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 sparse set of features.
- **the Lasso!**

Review of the Lasso

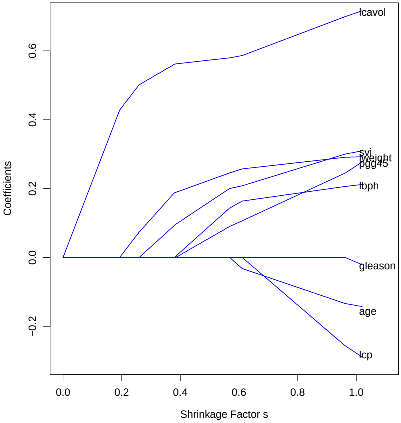
The **Lasso** is an estimator defined by the following optimization problem:

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

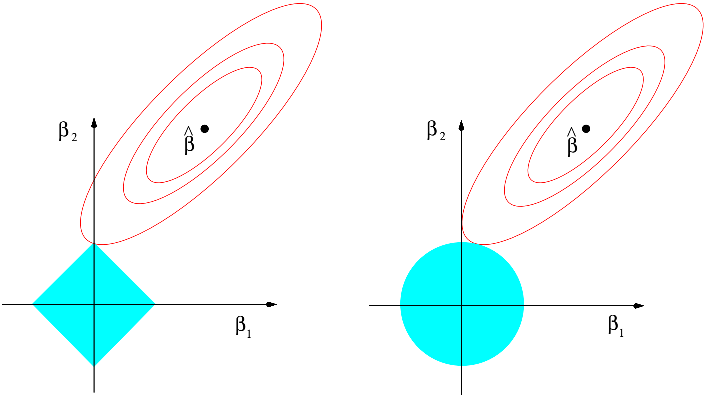
- Penalty $\implies$ sparsity (feature selection)
- Convex problem (good for computation and theory)
- *Ridge regression* uses penalty $\sum_j \beta_j^2 \leq s$ and does not yield sparsity

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

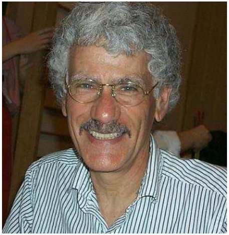
$$\log \frac{Pr(Y_i = k|x)}{\sum_k Pr(Y_i = k|x)} = \beta_{0k} + \sum_j x_{ij}\beta_{jk},\ k = 1, 2, \ldots K$$

  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 `glmnet` for fitting lasso models. Available on CRAN. Largely written in FORTRAN!!!!!
- It is very fast- can solve the current problem in a few minutes on a PC. Some builtin 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

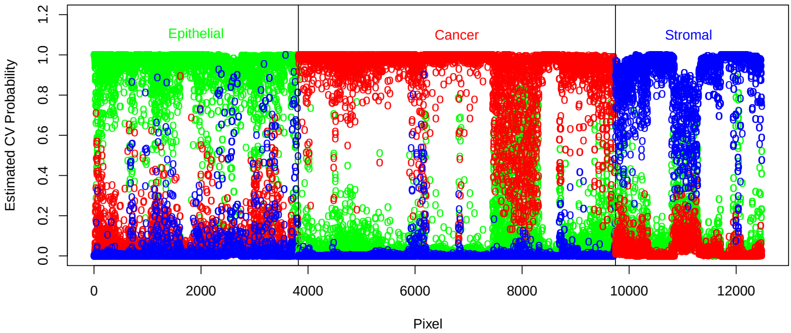
Cross-validation- min at 129 peaks; overall error rate= 4.2%

	Predicted			
true	Epi	Canc	Strom	Prop correct
Epi	3277.00	80.00	145.00	0.94
Canc	73.00	5106.00	13.00	0.98
Strom	79.00	86.00	2409.00	0.94

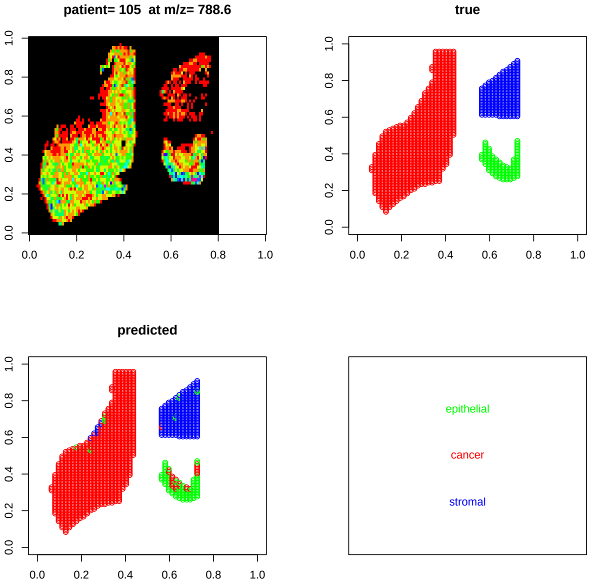
Test set: overall error rate =5.7%

	Predicted			
true	Epi	Canc	Strom	Prop correct
Epi	1606.00	149.00	19.00	0.91
Canc	23.00	1622.00	5.00	0.98
Strom	67.00	5.00	1222.00	0.94

Cross-validated estimates of class probabilities



Peak #	m/z value	Epi	Canc	Strom
1	101.5			
2	107.5	0.09		
3	110.5		0.44	0.56
4	123.5			
5	132.5	0.06		
6	134.5		0.10	0.13
7	135.5		0.13	0.19
8	137.5		-0.01	0.05
9	145.5	-0.71		
10	146.5	-0.41		
11	151.5	-0.15	0.35	-0.25
12	157.5		-0.07	-0.17
13	170.5			
14	171.5			
15	174.5	0.05		
16	175.5		0.14	0.54
17	179.5			
18	188.5			
19	212.5			
20	214.5		-014	-0.13
21	215.5		-1.17	-1.07
22	222.5	0.29		
23	224.5			
24	225.5			
25	231.5			
26	244.5	-0.01		
27	247.5		-0.31	-0.41
28	258.5	0.21		
29	270.5		-0.14	-0.24
30	278.5	0.32		
31	279.5	0.39		
32	280.5			
33	285.5			
34	289.5			
35	293.5	0.25		
36	297.5			
37	299.5			
38	301.5	-0.45	0.21	0.11
39	312.5	0.10	-0.44	-0.24
40	322.5			
41	324.5	0.06		
42	325.5	0.03	-0.00	-0.00
43	333.5		0.82	0.55
44	340.5		-0.02	-0.07
45	341.5			
46	347.5	0.01		
47	349.5			
48	353.5	0.05		
49	355.5	0.02		

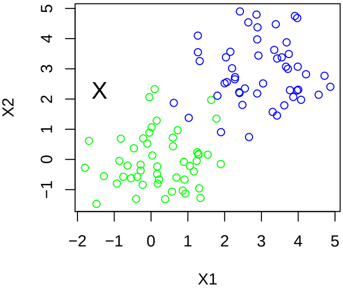

patient= 105 at m/z= 788.6
true
predicted
0.0
0.2
0.4
0.6
0.8
1.0
epithelial
cancer
stromal

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

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

Jonathan Taylor
Stanford University
PhD Student of Keith Worsley, 2001

81%

Ryan Tibshirani
CMU

71%

Rob Tibshirani
Stanford

Top matches from picadilo.com

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

Jonathan Taylor
Stanford University
PhD Student of Keith Worsley, 2001

81%

Ryan Tibshirani
CMU

Rob Tibshirani
Stanford

71%

69%

Conclusion

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 ϵ_i is an error term with mean zero. β_j is the weight given feature j

- Least squares fitting is defined by

$$\underset{\beta_0,\beta}{\text{minimize}} \frac{1}{2} \sum_i (y_i - \beta_0 - \sum_j x_{ij}\beta_j)^2$$

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

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

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

- Penalty $\Longrightarrow$ sparsity (feature selection)
- Convex problem (good for computation and theory)

The Lasso– continued

Equivalent Lagrangian form:

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

- We choose s or λ by cross-validation.

Prostate cancer example

$N = 88, p = 8$. Predicting log-PSA, in men after prostate cancer surgery

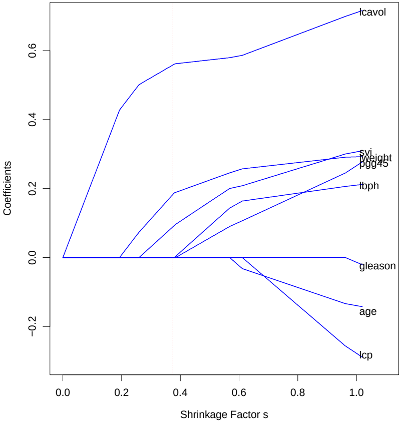


Illustration of the post-selection inference problem

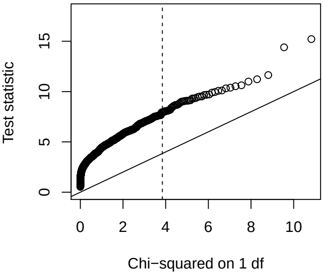
- Suppose that we apply forward stepwise regression to the prostate data, and our software prints out the usual p-values from the F-test at each stage.
- The test statistic is the reduction in sum of squares divided by an estimate of the error variance.

Results:

	P-value from F-test
lcavol	0.000
lweight	0.000
svi	0.047
lbph	0.047
pgg45	0.234
lcp	0.083
age	0.137
gleason	0.883

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,

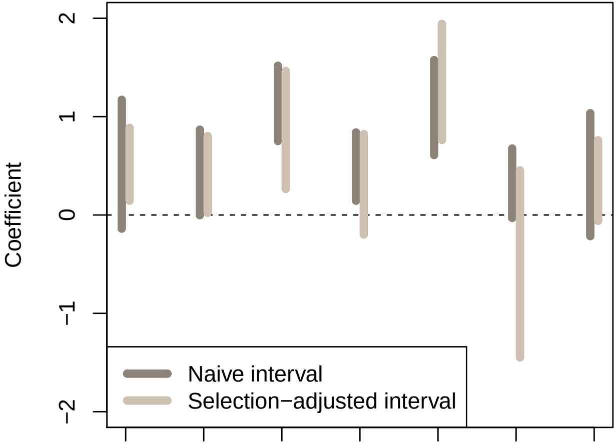
Example: Forward Stepwise regression

	FS, naive	FS, adjusted
lcavol	0.000	0.000
lweight	0.000	0.012
svi	0.047	0.849
lbph	0.047	0.337
pgg45	0.234	0.847
lcp	0.083	0.546
age	0.137	0.118
gleason	0.883	0.311

Table : *Prostate data example:* $n = 88, p = 8$. *Naive and selection-adjusted forward stepwise sequential tests*

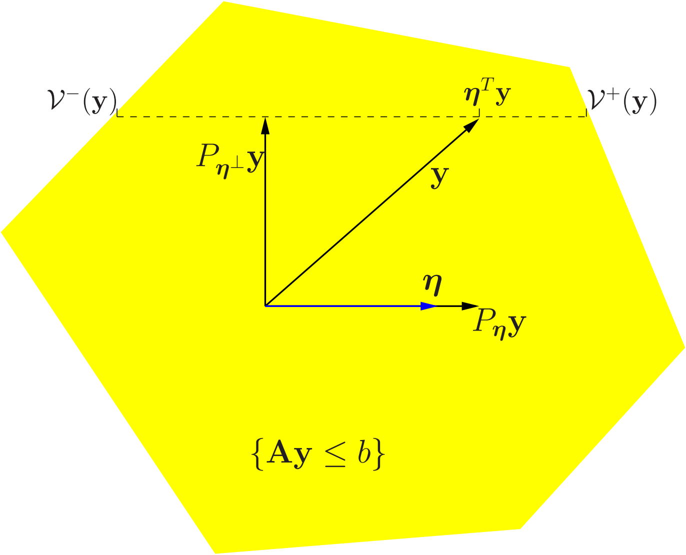
Lasso with fixed-λ

HIV data: selection intervals for lasso with fixed tuning parameter λ.



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 η
$$F^{[\mathcal{V}^-, \mathcal{V}^+]}_{\eta^\top \mu, \sigma^2 \eta^\top \eta}(\eta^\top y) | \{A_{L,\lambda} y \leq b_{L,\lambda}\} \sim \mathrm{Unif}(0,1)$$
(truncated Gaussian distribution), where V^-, V^+ are (computable) constants that are functions of η, A, b.
- Result is **exact** (finite-sample) and works for any X!


$\mathcal{V}^{-}(\mathbf{y})$
$\boldsymbol{\eta}^T\mathbf{y}$
$\mathcal{V}^{+}(\mathbf{y})$
$P_{\boldsymbol{\eta}^\perp}\mathbf{y}$
$\mathbf{y}$
$\boldsymbol{\eta}$
$P_{\boldsymbol{\eta}}\mathbf{y}$
$\{\mathbf{Ay} \leq b\}$

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 $Ay \leq b$.

The polyhedral lemma: intuition continued

- Now without selection, the inner product between x_3 and y can have any value between $-\infty$ and ∞. **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

$$F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\eta^\top\mu,\sigma^2\eta^\top\eta}(\eta^\top y)|\{A_{L,\lambda}y \le b_{L,\lambda}\} \sim \mathrm{Unif}(0,1)$$

- **Very powerful**: can be used to construct p-values and confidence intervals for parameters after any selection event of the form $AY \le b$.
- 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
  `fsInf(x,y)`
- Fixed lambda Lasso
  `fixedLassoInf(x,y)`
- Least angle regression
  `larInf(x,y)`

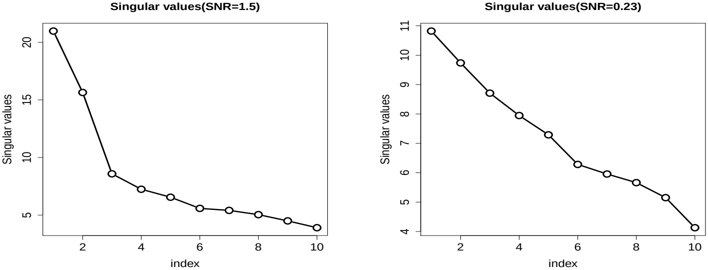
These compute p-values and selection intervals

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

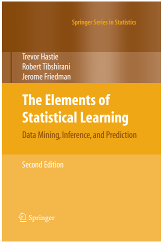
Scree Plot of (true) rank = 2



(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



Springer Series in Statistics

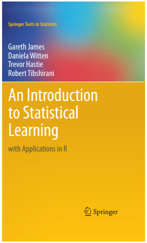
Trevor Hastie
Robert Tibshirani
Jerome Friedman

The Elements of Statistical Learning

Data Mining, Inference, and Prediction

Second Edition

Springer



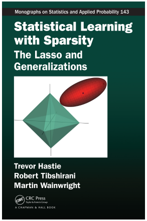
Springer Texts in Statistics

Gareth James
Daniela Witten
Trevor Hastie
Robert Tibshirani

An Introduction to Statistical Learning

with Applications in R

Springer



Monographs on Statistics and Applied Probability 143

Statistical Learning with Sparsity

The Lasso and Generalizations

Trevor Hastie
Robert Tibshirani
Martin Wainwright

CRC Press
A CHAPMAN & HALL BOOK

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.