A significance test for the lasso

Robert Tibshirani, Stanford University

First part: Joint work with *Richard Lockhart* (SFU), *Jonathan Taylor* (Stanford), and *Ryan Tibshirani* (Carnegie-Mellon Univ.)
Second part: Joint work with *Max Grazier G'Sell, Stefan Wager and Alexandra Chouldechova*, Stanford University.

Reaping the benefits of LARS: *A special thanks to Brad Efron, Trevor Hastie and Iain Johnstone*

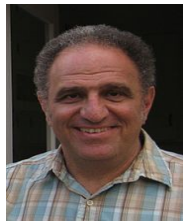
Richard Lockhart
Simon Fraser University
Vancouver
(PhD . Student of David Blackwell, Berkeley, 1979)

Jonathan Taylor

Ryan Tibshirani
Asst Prof, CMU
(PhD Student of Taylor, Stanford 2011)

An Intense and Memorable Collaboration!

With substantial and unique contributions from all four authors:



Quarterback and cheerleader

Expert in "elementary" theory

Expert in "advanced" theory

The closer: pulled together the elementary and advanced views into a coherent whole

Overview

- Although this is "yet another talk on the lasso", it may have something to offer mainstream statistical practice.

Talk Outline

1. Review of lasso, LARS, forward stepwise
2. The covariance test statistic
3. Null distribution of the covariance statistic
4. Theory for orthogonal case
5. Simulations of null distribution
6. General $\mathbf{X}$ results
7. Sequential testing and FDR
8. Graphical models

Talk Outline

1. Review of lasso, LARS, forward stepwise
2. The covariance test statistic
3. Null distribution of the covariance statistic
4. Theory for orthogonal case
5. Simulations of null distribution
6. General $\mathbf{X}$ results
7. Sequential testing and FDR
8. Graphical models

The Lasso

Observe n predictor-response pairs $(x_i,\ y_i)$, where $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$. Forming $X \in \mathbb{R}^{n\times p}$, with standardized columns, the **Lasso** is an estimator defined by the following optimization problem (Tibshirani 1996, Chen et al. 1998):

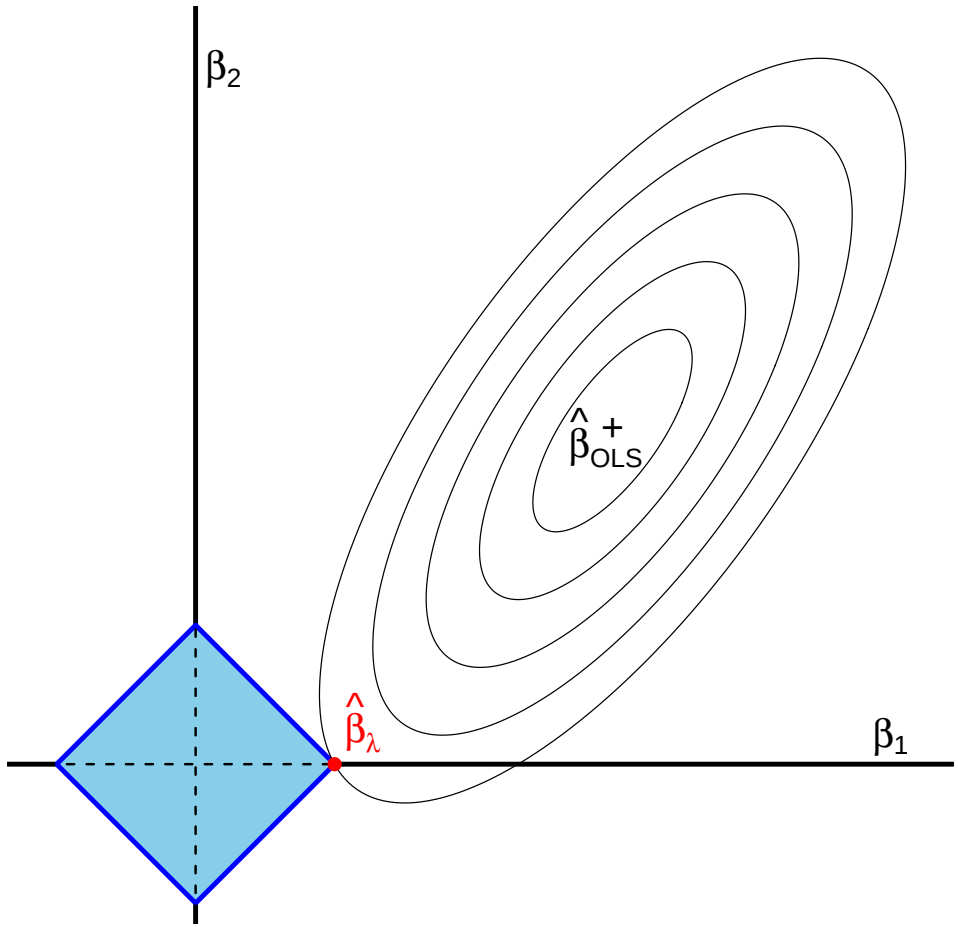
$$\underset{\beta_0\in\mathbb{R},\ \beta\in\mathbb{R}^p}{\text{minimize}}\ \frac{1}{2}\|y-\beta_0 1-X\beta\|^2+\lambda\|\beta\|_1$$

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

The Lasso

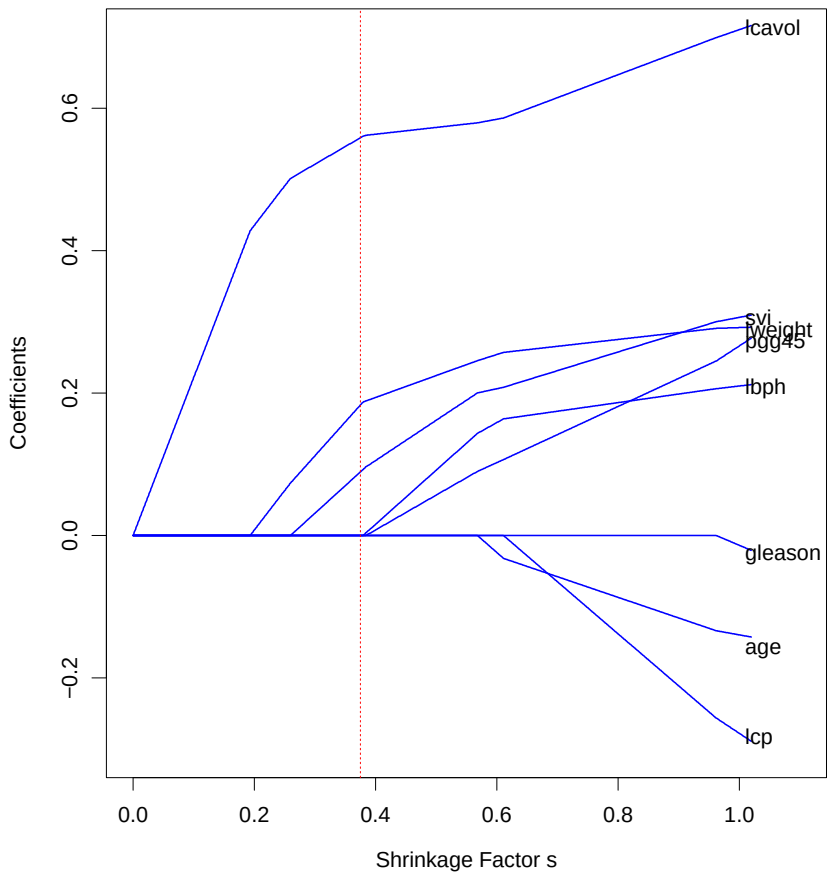
Why does ℓ_1-penalty give sparse $\hat{\beta}_\lambda$?

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \frac{1}{2}\|y - X\beta\|^2 \qquad \text{subject to} \quad \|\beta\|_1 \leq s$$



Prostate cancer example

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



Emerging themes

- Lasso (ℓ_1) penalties have powerful *statistical* and *computational* advantages
- ℓ_1 penalties provide a natural to encourage/enforce sparsity and simplicity in the solution.
- *"Bet on sparsity principle"* (In the *Elements of Statistical learning*). Assume that the underlying truth is sparse and use an ℓ_1 penalty to try to recover it. If you're right, you will do well. If you're wrong— the underlying truth is not sparse—, then no method can do well. [Bickel, Buhlmann, Candes, Donoho, Johnstone,Yu ...]
- ℓ_1 penalties are convex and the assumed sparsity can lead to significant *computational* advantages

Setup and basic question

- Given an outcome vector $\mathbf{y} \in \mathbf{R}^n$ and a predictor matrix $\mathbf{X} \in \mathbf{R}^{n \times p}$, we consider the usual linear regression setup:

$$\mathbf{y} = \mathbf{X}\beta^* + \sigma\epsilon, \tag{1}$$

  where $\beta^* \in \mathbf{R}^p$ are unknown coefficients to be estimated, and the components of the noise vector $\epsilon \in \mathbf{R}^n$ are i.i.d. $N(0, 1)$.
- Given fitted lasso model at some λ can we produce a p-value for each predictor in the model? Difficult! (but we have some ideas for this- future work)
- What we show today: how to provide a p-value for each variable as it is added to lasso model

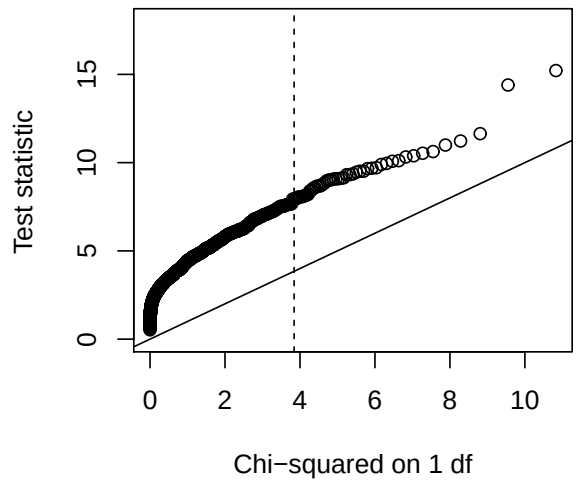
Forward stepwise regression

- This procedure enters predictors one a time, choosing the predictor that most decreases the residual sum of squares at each stage.
- Defining RSS to be the residual sum of squares for the model containing k predictors, and $\mathrm{RSS}_{\mathrm{null}}$ the residual sum of squares before the kth predictor was added, we can form the usual statistic

$$R_k = (\mathrm{RSS}_{\mathrm{null}} - \mathrm{RSS})/\sigma^2 \tag{2}$$

(with σ assumed known for now), and compare it to a χ^2_1 distribution.

Simulated example- Forward stepwise- F 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

Degrees of Freedom

Degrees of Freedom used by a procedure, $\hat{y} = h(y)$:

$$df_h = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{cov}(\hat{y}_i, y_i)$$

where $y \sim N(\mu, \sigma^2 I_n)$ [Efron (1986)].

Measures total self-influence of y_i's on their predictions.

Stein's formula can be used to evaluate df [Stein (1981)].

For fixed (non-adaptive) linear model fit on k predictors, $df = k$.

But for forward stepwise regression, df after adding k predictors is $> k$.

Degrees of Freedom – Lasso

- Remarkable result for the Lasso:

$$df_{\text{lasso}} = E[\#\text{nonzero coefficients}]$$

- For least angle regression, df is exactly k after k steps (under conditions). *So LARS spends one degree of freedom per step*!
- Result has been generalized in multiple ways in (Ryan Tibs & Taylor) Tibshirani & Taylor (2012), e.g. for general X, p, n.

Question that motivated today's work

Is there a statistic for testing in lasso/LARS having one degree of freedom?

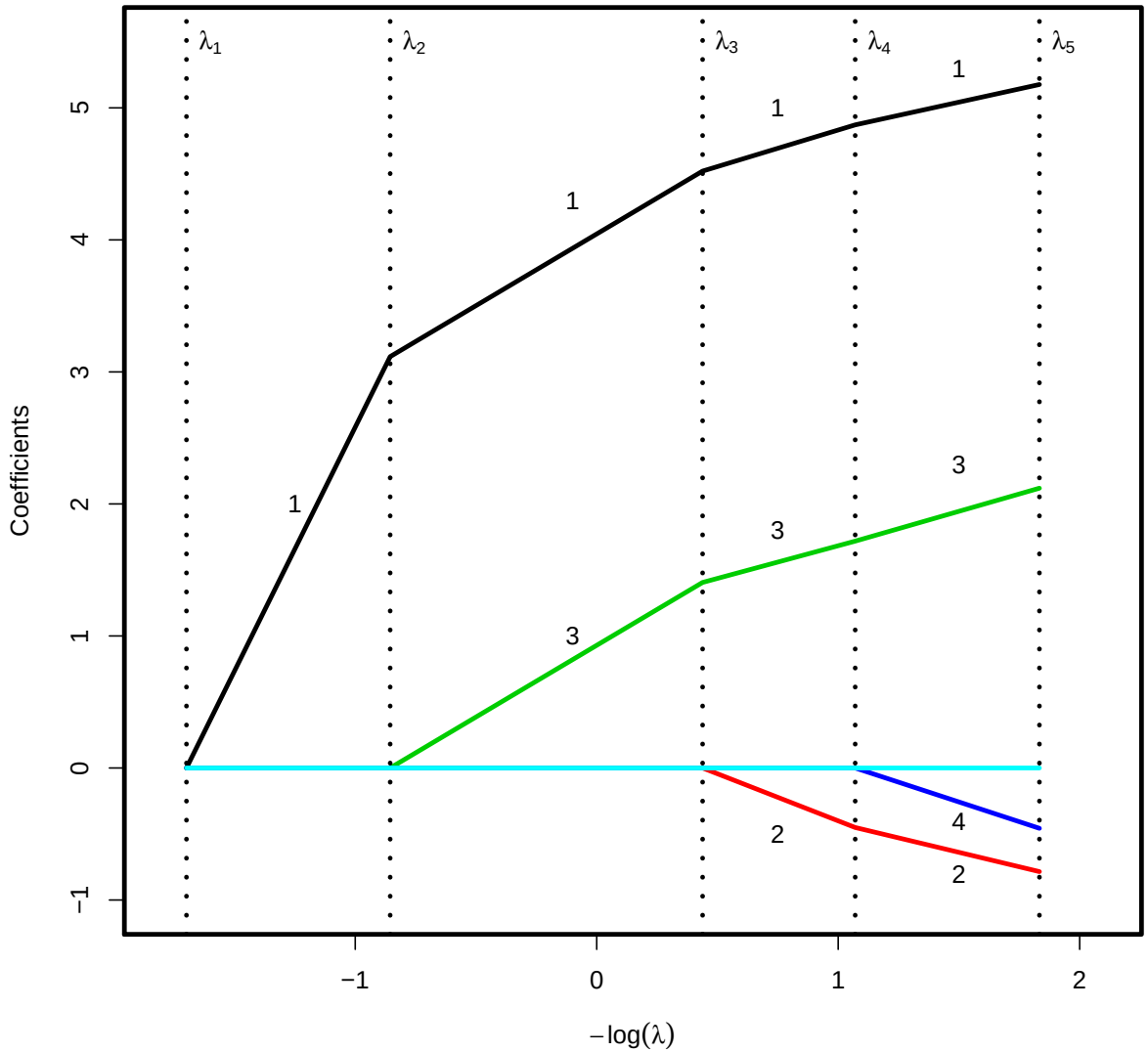
Quick review of least angle regression

Least angle regression is a method for constructing the path of lasso solutions.

A more democratic version of forward stepwise regression.

- find the predictor *most correlated* with the outcome,
- move the parameter vector in the least squares direction until some other predictor has as much correlation with the current residual.
- this new predictor is added to the active set, and the procedure is repeated.
- If a non-zero coefficient hits zero, that predictor is dropped from the active set, and the process is restarted. [This is "lasso" mode, which is what we consider here.]

Least angle regression

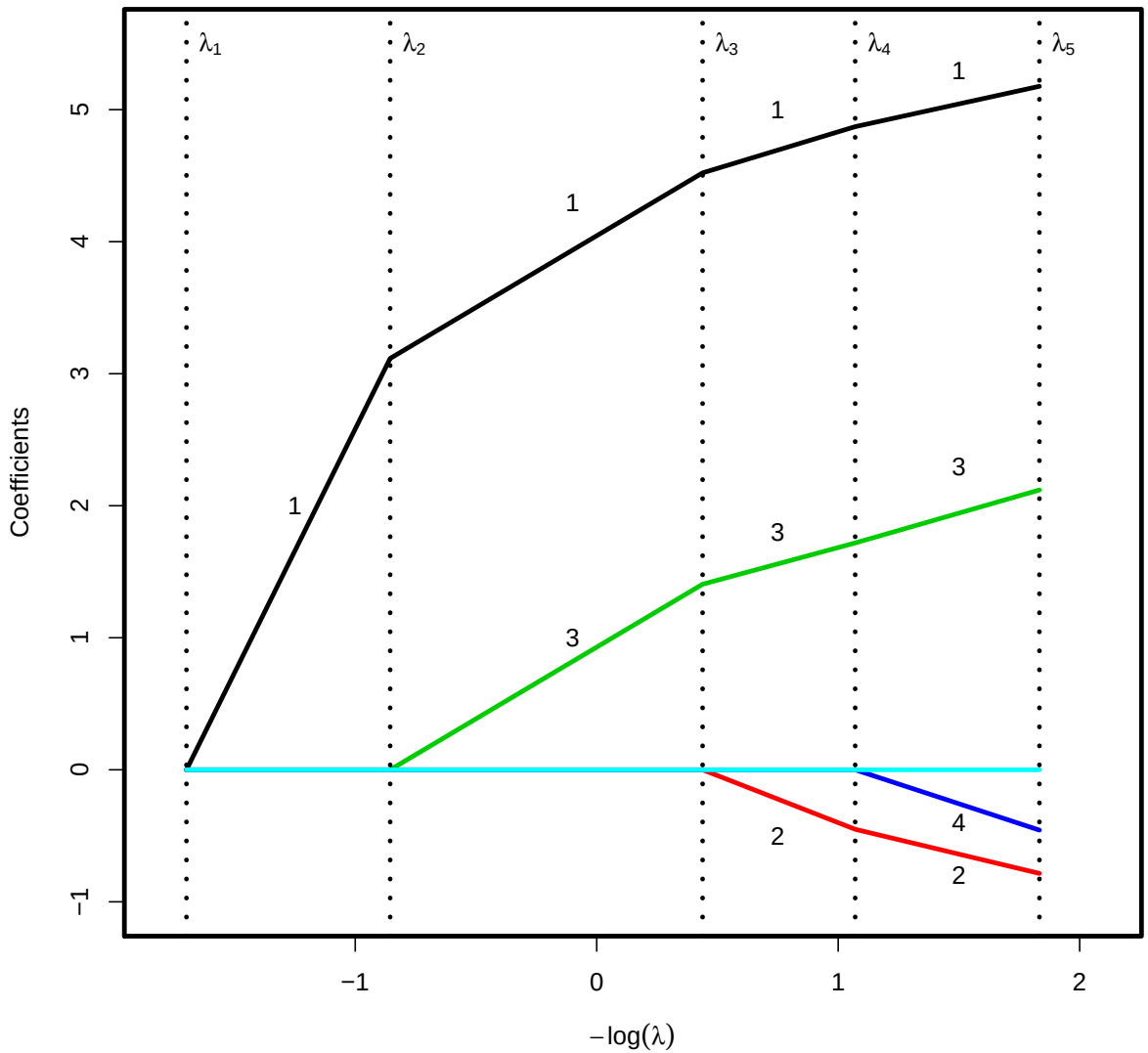


Talk Outline

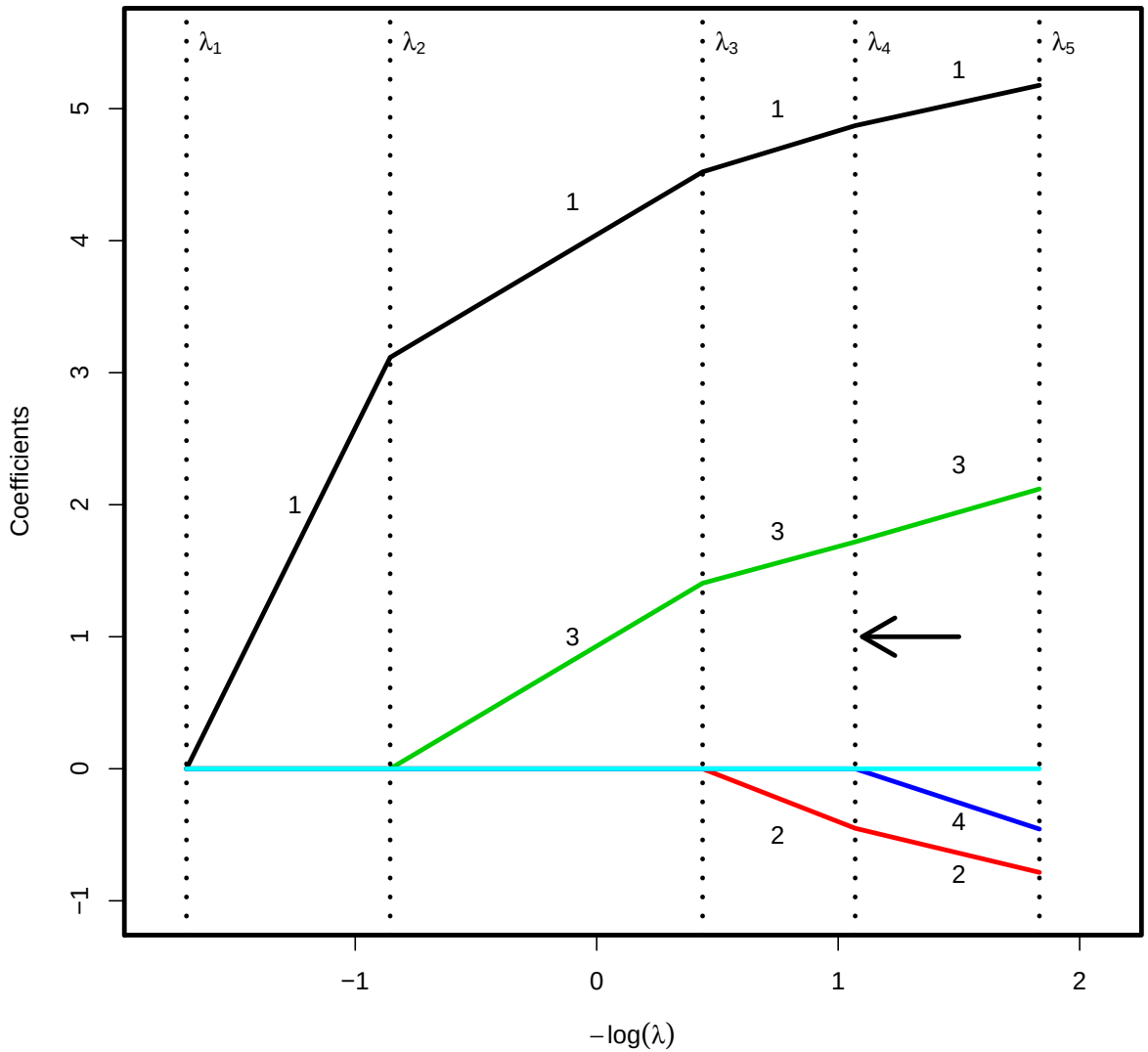
1. Review of lasso, LARS, forward stepwise
2. The covariance test statistic
3. Null distribution of the covariance statistic
4. Theory for orthogonal case
5. Simulations of null distribution
6. General $\mathbf{X}$ results
7. Sequential testing and FDR
8. Graphical models

The covariance test statistic

Suppose that we want a p-value for predictor 2, entering at the 3rd step.

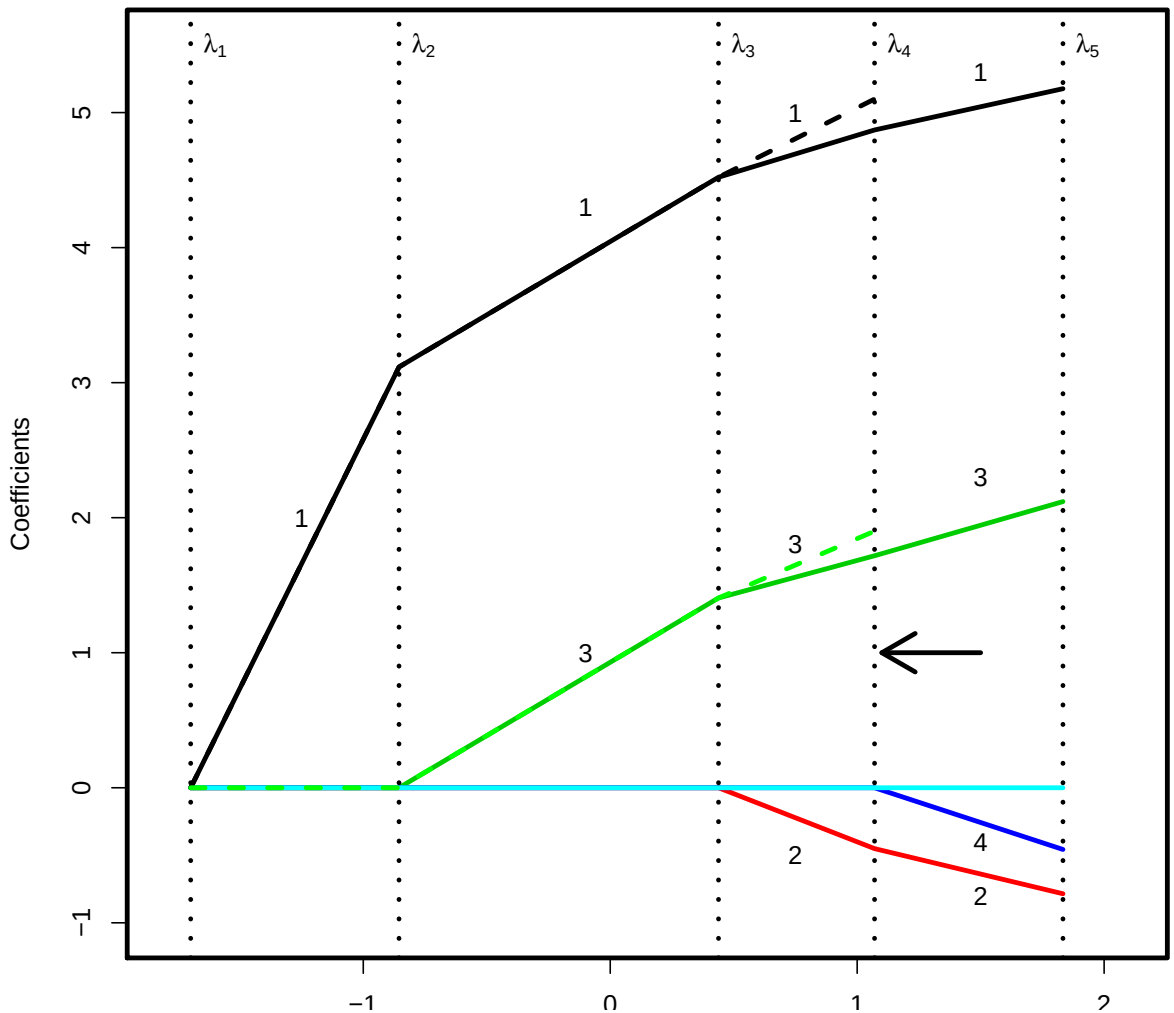


Compute covariance at λ_4: $\langle \mathbf{y}, \mathbf{X}\hat{\beta}(\lambda_4)\rangle$



Drop x_2, yielding active yet A; refit at λ_4, and compute resulting covariance at λ_4, giving

$$T = \Big(\langle \mathbf{y}, \mathbf{X}\hat{\beta}(\lambda_4)\rangle - \langle \mathbf{y}, \mathbf{X}_{\mathcal{A}}\hat{\beta}_{\mathcal{A}}(\lambda_4)\rangle\Big)/\sigma^2$$



The covariance test statistic: formal definition

- Suppose that we wish to test significance of predictor that enters LARS at λ_j.
- Let A be the active set before this predictor added
- Let the estimates at the end of this step be $\hat{\beta}(\lambda_{j+1})$
- We refit the lasso, keeping $\lambda = \lambda_{j+1}$ but using just the variables in $\mathcal{A}$. This yields estimates $\hat{\beta}_{\mathcal{A}}(\lambda_{j+1})$. The proposed *covariance test statistic* is defined by

$$T_j = \frac{1}{\sigma^2} \cdot \Big(\langle \mathbf{y}, \mathbf{X}\hat{\beta}(\lambda_{j+1}) \rangle - \langle \mathbf{y}, \mathbf{X}_{\mathcal{A}}\hat{\beta}_{\mathcal{A}}(\lambda_{j+1}) \rangle \Big). \tag{3}$$

- measures how much of the covariance between the outcome and the fitted model can be attributed to the predictor which has just entered the model.

Talk Outline

1. Review of lasso, LARS, forward stepwise
2. The covariance test statistic
3. Null distribution of the covariance statistic
4. Theory for orthogonal case
5. Simulations of null distribution
6. General $\mathbf{X}$ results
7. Sequential testing and FDR
8. Graphical models

Main result

Under the null hypothesis that all signal variables are in the model:

$$T_j = \frac{1}{\sigma^2} \cdot \Big(\langle \mathbf{y}, \mathbf{X}\hat{\beta}(\lambda_{j+1}) \rangle - \langle \mathbf{y}, \mathbf{X}_{\mathcal{A}}\hat{\beta}_{\mathcal{A}}(\lambda_{j+1}) \rangle \Big) \to \mathrm{Exp}(1)$$

as $p, n \to \infty$.

More details to come

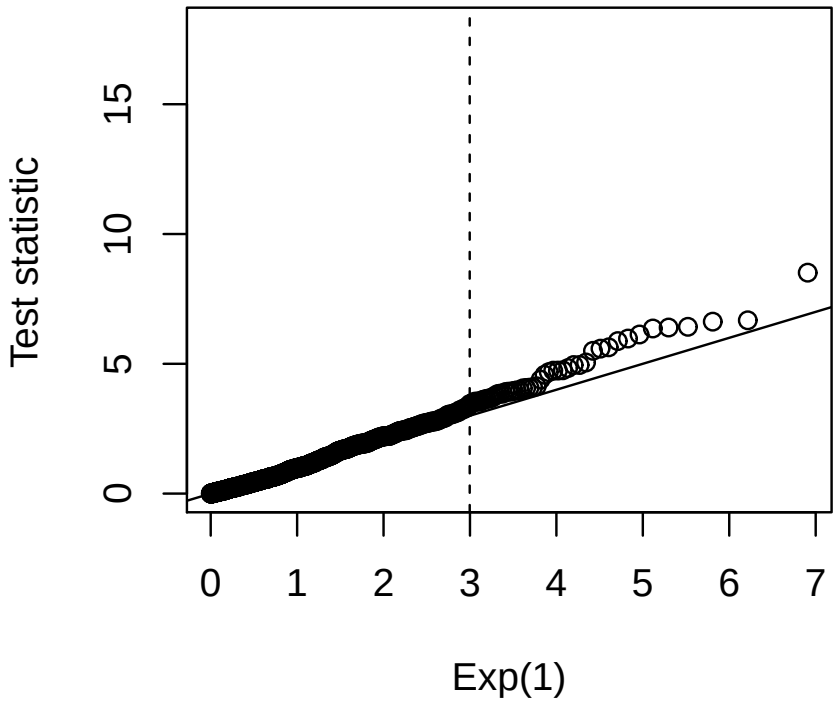
Comments on the covariance test

$$T_j = \frac{1}{\sigma^2} \cdot \Big(\langle \mathbf{y}, \mathbf{X}\hat{\beta}(\lambda_{j+1}) \rangle - \langle \mathbf{y}, \mathbf{X}_{\mathcal{A}}\hat{\beta}_{\mathcal{A}}(\lambda_{j+1}) \rangle \Big). \tag{4}$$

- Generalization of standard χ^2 or F test, designed for fixed linear regression, to adaptive regression setting.
- Exp(1) is the same as $\chi^2_2/2$; its mean is 1, like χ^2_1: overfitting due to adaptive selection is offset by shrinkage of coefficients
- Form of the statistic is very specific- uses covariance rather than residual sum of squares (they are equivalent for projections)
- Covariance must be evaluated at specific knot λ_{j+1}
- Applies when $p > n$ or $p \leq n$: although asymptotic in p, Exp(1) seem to be a very good approximation even for small p

Simulated example- Lasso- Covariance statistic

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



Example: Prostate cancer data

	Stepwise	Lasso
lcavol	0.000	0.000
lweight	0.000	0.052
svi	0.041	0.174
lbph	0.045	0.929
pgg45	0.226	0.353
age	0.191	0.650
lcp	0.065	0.051
gleason	0.883	0.978

Simplifications

- For any design, the first covariance test T_1 can be shown to equal $\lambda_1(\lambda_1 - \lambda_2)$.
- For orthonormal design, $T_j = \lambda_j(\lambda_j - \lambda_{j+1})$ for all j; for general designs, $T_j = c_j\lambda_j(\lambda_j - \lambda_{j+1})$
- For orthonormal design, $\lambda_j = |\hat{\beta}_{(j)}|$, the jth largest univariate coefficient in absolute value. Hence

$$T_j = (|\hat{\beta}_{(j)}|(|\hat{\beta}_{(j)}| - |\hat{\beta}_{(j+1)}|). \tag{5}$$

Rough summary of theoretical results

Under somewhat general conditions, after all signal variables are in the model, distribution of T for kth null predictor $\to \mathrm{Exp}(1/k)$

Talk Outline

1. Review of lasso, LARS, forward stepwise
2. The covariance test statistic
3. Null distribution of the covariance statistic
4. Theory for orthogonal case
5. Simulations of null distribution
6. General $\mathbf{X}$ results
7. Sequential testing and FDR
8. Graphical models

Theory for orthogonal case

Global null case: first predictor to enter

Recall that in this setting,

$$T_j = \lambda_j(\lambda_j - \lambda_{j+1})$$

and $\lambda_j = |\hat{\beta}_{(j)}|$, $\hat{\beta}_j \sim N(0,1)$

So the question is:

Suppose $V_1 > V_2 \ldots > V_n$ are the order statistics from a χ_1 distribution (absolute value of a standard Gaussian).

What is the asymptotic distribution of $V_1(V_1 - V_2)$?

[Ask Richard Lockhart!]

Theory for orthogonal case

Global null case: first predictor to enter

Lemma

Lemma 1: Top two order statistics*: Suppose* $V_1 > V_2 \ldots > V_p$ *are the order statistics from a* χ_1 *distribution (absolute value of a standard Gaussian) with cumulative distribution function* $F(x) = (2\Phi(x) - 1)1(x > 0)$*, where* $\Phi(x)$ *is standard normal cumulative distribution function. Then*

$$V_1(V_1 - V_2) \to \text{Exp}(1). \tag{6}$$

Lemma

Lemma 2: All predictors*. Under the same conditions as Lemma 1,*

$$(V_1(V_1 - V_2), \ldots, V_k(V_k - V_{k+1})) \to (\text{Exp}(1), \text{Exp}(1/2), \ldots \text{Exp}(1/k))$$

Proof uses a uses theorem from de Haan & Ferreira (2006). We were unable to find these remarkably simple results in the literature.

Talk Outline

1. Review of lasso, LARS, forward stepwise
2. The covariance test statistic
3. Null distribution of the covariance statistic
4. Theory for orthogonal case
5. Simulations of null distribution
6. General $\mathbf{X}$ results
7. Sequential testing and FDR
8. Graphical models

Simulations of null distribution

TABLES OF SIMULATION RESULTS ARE BORING !!!!

SHOW SOME MOVIES INSTEAD

Talk Outline

General **X** results

Under appropriate condition on **X**, as $p, N \to \infty$,

1. *Global null case*: $T_1 = \lambda_1(\lambda_1 - \lambda_2) \to \mathrm{Exp}(1)$.
2. *Non-null case:* After the k strong signal variables have entered, under the null hypothesis that the rest are weak,

$$T_{k+1} \overset{n,p\to\infty}{\leq} \mathrm{Exp}(1)$$

This is true under orthogonal design, approximately true under general design.

Jon Taylor: "Something magical happens in the math"

Conditions on X

- A sufficient condition: for any j, we require the existence of a subset S not containing j such that the variables U_i, $i \in S$ are not too correlated, in the sense that the conditional variance of any one on all the others is bounded below. This subset S has to be of size at least $\log p$.

Case of Unknown σ

Let

$$W_k = \Big(\langle y, X\hat{\beta}(\lambda_{k+1})\rangle - \langle y, X_A\hat{\beta}_A(\lambda_{k+1})\rangle\Big). \tag{7}$$

and assuming $n > p$, let $\hat{\sigma}^2 = \sum_{i=1}^n (y_i - \hat{\mu}_{\text{full}})^2/(n-p)$. Then asymptotically

$$F_k = \frac{W_k}{\hat{\sigma}^2} \sim F_{2,n-p} \tag{8}$$

[W_j/σ^2 is asymptotically Exp(1) which is the same as $\chi^2_2/2$, $(n-p)\cdot\hat{\sigma}^2/\sigma^2$ is asymptotically χ^2_{n-p} and the two are independent.]

When $p > n$, σ^2 miust be estimated with more care.

Extensions

- Elastic Net
- Generalized likelihood models: GLMs, Cox model. Natural extensions, but detailed theory not yet developed.

Generalizations

Taylor, Loftus, Ryan Tibshirani (2013)

- $$\underset{\beta_0 \in \mathbb{R},\ \beta \in \mathbb{R}^p}{\text{minimize}}\ \frac{1}{2}\|y - \beta_0 1 - X\beta\|^2 + \lambda P(\beta)$$

  $P(\beta)$ is any semi-norm.
- They derive a global test for $\beta = 0$ that is *exact* for finite n, p. Asymptotically equivalent (and numerically close) to covariance test in the lasso setting.
- Upcoming: exact selection intervals for coefficients from lasso solution at a knot.

Software

R library

`covTest(larsobj,x,y)`,

where larsobj is fit from LARS or glmpath [logistic or Cox model (Park and Hastie)]. Produces p-values for predictors as they are entered. More coming!

Two important problems

- Estimation of σ^2 when $p > n$ (we're working on it)
- Sequential testing and FDR

Talk Outline

1. Review of lasso, LARS, forward stepwise
2. The covariance test statistic
3. Null distribution of the covariance statistic
4. Theory for orthogonal case
5. Simulations of null distribution
6. General $\mathbf{X}$ results
7. Sequential testing and FDR
8. Graphical models

Sequential testing and FDR

- How can we use the covariance test p-values to from a selection rule with guaranteed FDR control?
- Max G'Sell, Stefan Wager, Alex Chouldechova, Rob Tibshirani (2013) (Really the students' work!)
- Consider a hypothesis testing scenario where we have a p-value p_j for each of a set of hypotheses H_1, H_2, ..., H_m, and these hypotheses must be rejected in a sequential manner.
- Because of this, we cannot apply the standard Benjamini-Hochberg procedure

The idea

- Transform the p-values p_1, ..., p_m into statistics $q_1 < ... < q_m$, such that the q_i behaved like a sorted list of p-values. Then, we apply the BH procedure on the q_i,
- Under the global null where p_1, ..., $p_m \sim iid\ U([0,1])$, we can achieve such a transformation using the Rényi representation theorem
- Rényi showed that if Y_1, ..., Y_m are independent standard exponential random variables, then

$$\left(\frac{Y_1}{m}, \frac{Y_1}{m} + \frac{Y_2}{m-1}, ..., \sum_{i=1}^{m} \frac{Y_i}{m-i+1} \right) \sim E_{1,m}, E_{2,m}, ..., E_{m,m},$$

  where the $E_{i,m}$ are exponential order statistics,
- Idea: is Uniform$->$ Exponential $->$ CumSum (Exponential again) $->$ Uniform

ForwardStop procedure

$$
\begin{aligned}
Y_i &= -\log(1-p_i), \\
Z_i &= \sum_{j=1}^{i} Y_j/(m-j+1), \text{and} \\
q_i &= 1-e^{-Z_i}.
\end{aligned}
$$

Apply BH to the q_i; with one more simplification gives

$$
\hat{k}_F = \max\left\{ k \in \{1, \ldots, m\} : \frac{1}{k}\sum_{i=1}^{k} Y_i \leq \alpha \right\},
$$

We show that if the null p-values are i.i.d $U[0,1]$, then then this procedure controls the FDR at level α

Example

Apply to covariance test p-values with Exp(1) Null;

$n = 50, p = 10, \beta_1 = 2, \beta_3 = 4, \beta_2 = \beta_4 = \beta_5 \ldots \beta_{10} = 0$

LARS step	1	2	3	4	5	6	7	8	9
Predictor	3	1	4	7	9	10	2	6	5
p-value	0.00	0.07	0.89	0.84	0.98	0.96	0.93	0.95	0.99
Ave of $-\log(1-p)$	0.00	0.04	0.76	1.03	1.60	1.87	1.99	2.11	2.39

Note that independence assumptions needed for ForwardStop only met for orthogonal design

Paper has much more, including cumSum rules from the end: this exploits approximate $Exp(1/k)$ behaviour

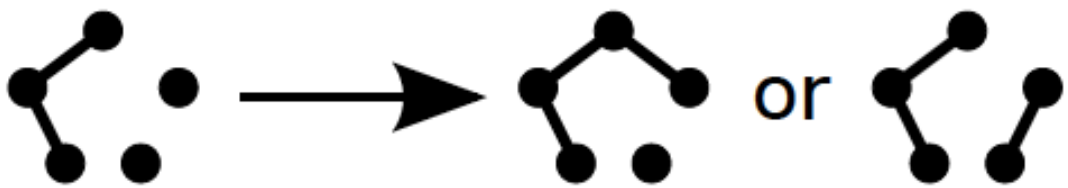
Talk Outline

1. Review of lasso, LARS, forward stepwise
2. The covariance test statistic
3. Null distribution of the covariance statistic
4. Theory for orthogonal case
5. Simulations of null distribution
6. General $\mathbf{X}$ results
7. Sequential testing and FDR
8. Graphical models

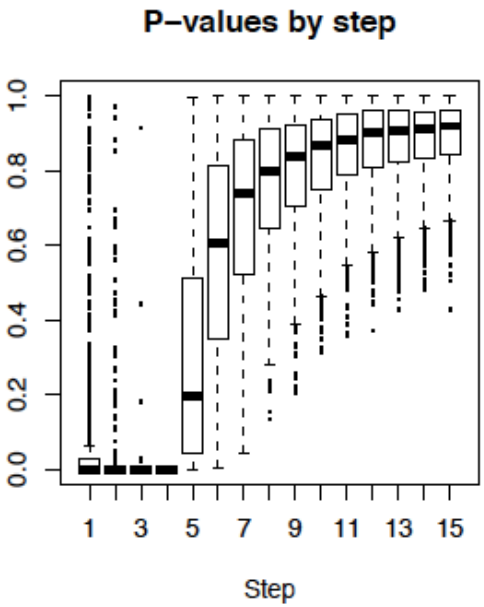
Application to graphical models

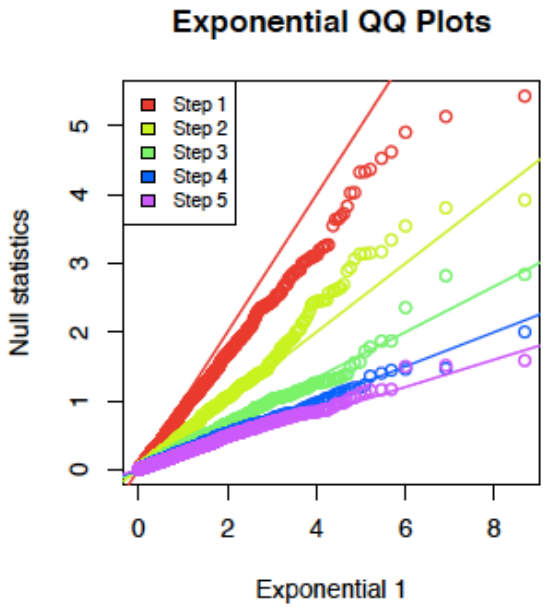
Max G'Sell, Taylor, Tibshirani (2013)

- Graphical model estimation through ℓ_1-penalized log-likelihood (graphical lasso)
- As λ decreases, connected components are fused together. We get a LARS-like path with knots λ_j.
- We derive the corresponding covariance test $n\lambda_j(\lambda_j - \lambda_{j+1}) \sim \mathrm{Exp}(1/\mathrm{j})$ for testing the significance of an edge



Example


P−values by step
0.0
0.2
0.4
0.6
0.8
1.0
1
3
5
7
9
11
13
15
Step


Exponential QQ Plots
Step 1
Step 2
Step 3
Step 4
Step 5
Null statistics
0
1
2
3
4
5
0
2
4
6
8
Exponential 1

THANK YOU!

Chen, S. S., Donoho, D. L. & Saunders, M. A. (1998), 'Atomic decomposition by basis pursuit', *SIAM Journal on Scientific Computing* pp. 33–61.

de Haan, L. & Ferreira, A. (2006), *Extreme Value Theory: An Introduction*, Springer.

Efron, B. (1986), 'How biased is the apparent error rate of a prediction rule?', *Journal of the American Statistical Association* **81**, 461–70.

Stein, C. (1981), 'Estimation of the mean of a multivariate normal distribution', *Annals of Statistics* **9**, 1131–1151.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society Series B* **58**(1), 267–288.

Tibshirani, R. J. & Taylor, J. (2012), 'Degrees of freedom in lasso problems', *Annals of Statistics* **40**(2), 1198–1232.