

A significance test for the lasso

Robert Tibshirani, Stanford University

Presidential invited address, ENAR 2014

Joint work with *Richard Lockhart* (SFU), *Jonathan Taylor* (Stanford), and *Ryan Tibshirani* (Carnegie-Mellon Univ.)

Paper to appear (with discussion) *Annals of Statistics* 2014

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)



Overview

- Not a review, but instead some recent (unpublished work) on inference in the lasso.
- 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 Example
- 8 Case of Unknown σ
- 9 Extensions to elastic net, generalized linear models, Cox model
- 10 Discussion and more recent developments work

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 Example
- 8 Case of Unknown σ
- 9 Extensions to elastic net, generalized linear models, Cox model
- 10 Discussion and more recent developments work

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):

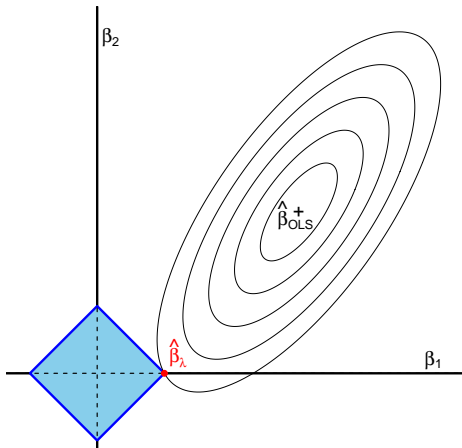
$$\underset{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p}{\text{minimize}} \quad \frac{1}{2} \|y - \beta_0 \mathbf{1} - X\beta\|^2 + \lambda \|\beta\|_1$$

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

The Lasso

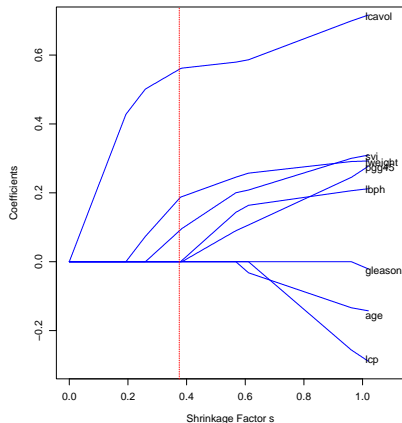
Why does ℓ_1 -penalty give sparse $\hat{\beta}_\lambda$?

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



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, Bühlmann, Candès, 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:

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

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. $N(0, 1)$.

- What we show today: how to provide a p-value for each variable as it is added to lasso model via the LAR algorithm

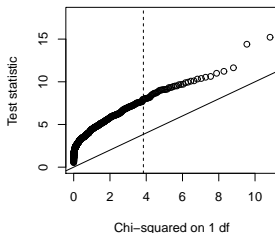
Forward stepwise regression

- 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 RSS_{null} the residual sum of squares before the k th predictor was added, we can form the usual statistic

$$R_k = (RSS_{\text{null}} - RSS)/\sigma^2 \quad (2)$$

(with σ assumed known for now), and compare it to a χ_1^2 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, $\hat{y} = h(y)$:

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

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?

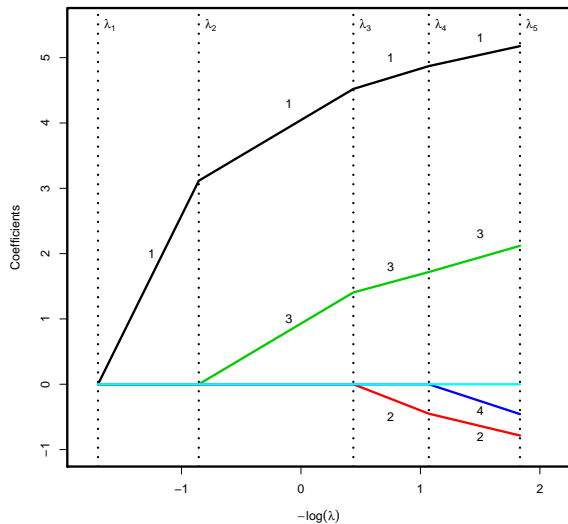
Quick review of least angle regression

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

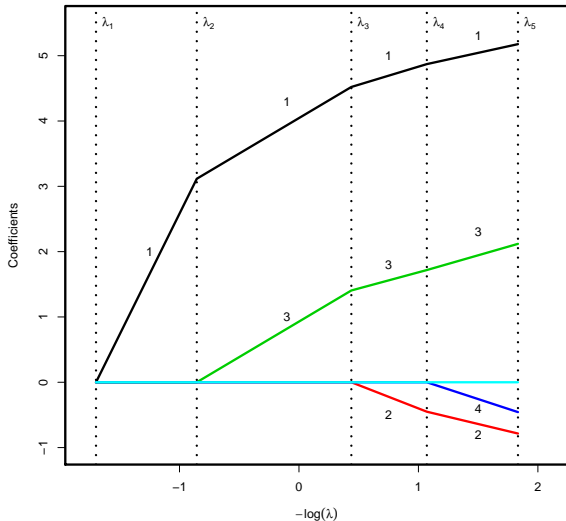


Talk Outline

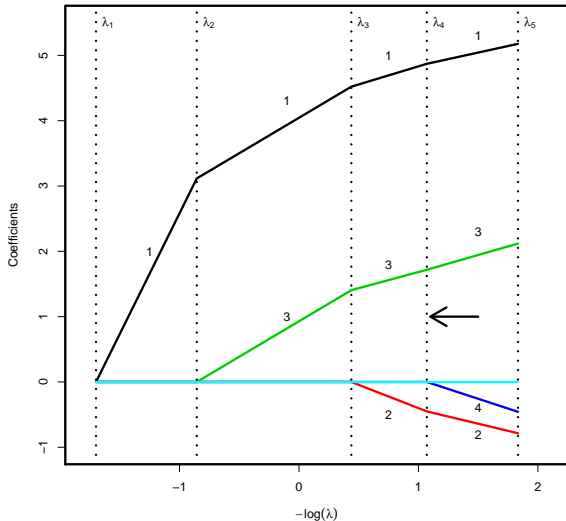
- 1 Review of lasso, LARS, forward stepwise
- 2 The covariance test statistic**
- 3 Null distribution of the covariance statistic
- 4 Theory for orthogonal case
- 5 Simulations of null distribution
- 6 General \mathbf{X} results
- 7 Example
- 8 Case of Unknown σ
- 9 Extensions to elastic net, generalized linear models, Cox model
- 10 Discussion and more recent developments work

The covariance test statistic

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

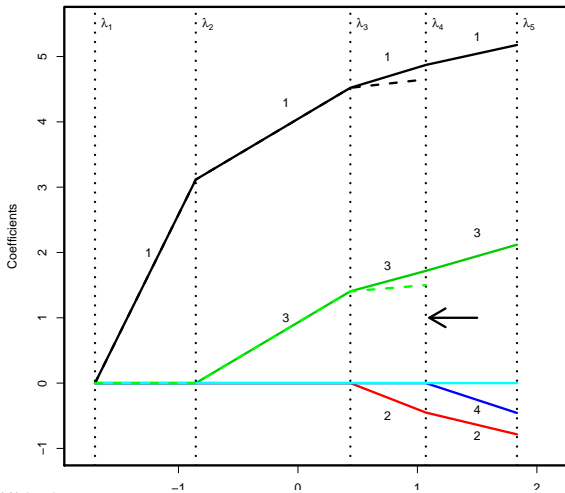


Compute covariance at λ_4 : $\langle \mathbf{y}, \mathbf{X}\hat{\beta}(\lambda_4) \rangle$



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

$$T = \left(\langle \mathbf{y}, \mathbf{X} \hat{\beta}(\lambda_4) \rangle - \langle \mathbf{y}, \mathbf{X}_A \hat{\beta}_A(\lambda_4) \rangle \right) / \sigma^2$$



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

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

- 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 Example
- 8 Case of Unknown σ
- 9 Extensions to elastic net, generalized linear models, Cox model
- 10 Discussion and more recent developments work

Main result

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

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

as $p, n \rightarrow \infty$.

More details to come

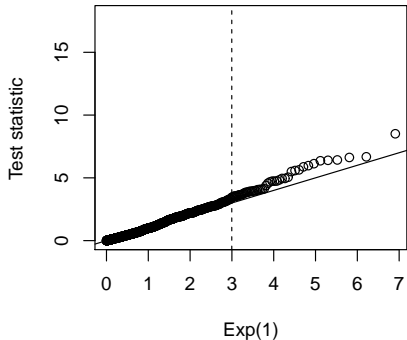
Comments on the covariance test

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

- Generalization of standard χ^2 or F test, designed for fixed linear regression, to adaptive regression setting.
- $\text{Exp}(1)$ is the same as $\chi_2^2/2$; its mean is 1, like χ_1^2 : 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 , $\text{Exp}(1)$ 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 j th largest univariate coefficient in absolute value. Hence

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

Rough summary of theoretical results

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

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 Example
- 8 Case of Unknown σ
- 9 Extensions to elastic net, generalized linear models, Cox model
- 10 Discussion and more recent developments work

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 $V_1 > V_2 \dots > V_n$ 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 $V_1 > V_2 \dots > V_p$ 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*

$$V_1(V_1 - V_2) \rightarrow \text{Exp}(1). \quad (6)$$

Lemma

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

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

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

Heuristically, the $\text{Exp}(1)$ limiting distribution for T_1 can be seen as follows:

- The spacings $|\hat{\beta}_{(1)}| - |\hat{\beta}_{(2)}|$ have an exponential distribution asymptotically, while $|\hat{\beta}_{(1)}|$ has an extreme value distribution with relatively small variance.
- It turns out that $|\hat{\beta}_{(1)}|$ is just the right (stochastic) scale factor to give the spacings an $\text{Exp}(1)$ distribution.

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 X results
- 7 Example
- 8 Case of Unknown σ
- 9 Extensions to elastic net, generalized linear models, Cox model
- 10 Discussion and more recent developments work

Simulations of null distribution

TABLES OF SIMULATION RESULTS ARE BORING !!!!

SHOW SOME MOVIES INSTEAD

Proof sketch

We use a theorem from de Haan & Ferreira (2006) on the asymptotic distributions of extreme order statistics.

- Let E_1, E_2 be independent standard exponentials. There are constants a_n and b_n such that

$$W_{1n} \equiv b_n(V_1 - a_n) \longrightarrow W_1 = \log(E_1)$$

- For those same constants put $W_{2n} = b_n(X_2 - a_n)$. Then

$$(W_{1n}, W_{2n}) \longrightarrow (W_1, W_2) = (-\log(E_1), -\log(E_1 + E_2))$$

- The quantity of interest T is a function of W_{1n}, W_{2n} . A change of variables shows that $T \longrightarrow \log(E_2 + E_1) - \log(E_1) = \log(1 + E_2/E_1)$, which is $\text{Exp}(1)$.

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 X results**
- 7 Example
- 8 Case of Unknown σ
- 9 Extensions to elastic net, generalized linear models, Cox model
- 10 Discussion and more recent developments work

General \mathbf{X} results

Under appropriate condition on \mathbf{X} , as $p, N \rightarrow \infty$,

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

$$T_{k+1} \stackrel{n, p \rightarrow \infty}{\leq} \text{Exp}(1)$$

Jon Taylor: “Something magical happens in the math”

Conditions on X

- The main condition is that for each (j, s_j) the random variable $M_{j,s}(g)$ grows sufficiently fast.
- 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 $\log p$.

Talk Outline

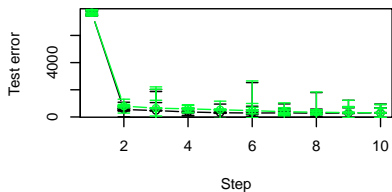
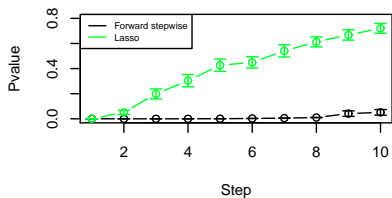
- 1 Review of lasso, LARS, forward stepwise
- 2 The covariance test statistic
- 3 Null distribution of the covariance statistic
- 4 Theory for orthogonal case
- 5 Simulations of null distribution
- 6 General X results
- 7 Example**
- 8 Case of Unknown σ
- 9 Extensions to elastic net, generalized linear models, Cox model
- 10 Discussion and more recent developments work

HIV mutation data

$N = 1057$ samples

$p = 217$ mutation sites ($x_{ij}=0$ or 1)

y = a measure of drug resistance



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 Example
- 8 Case of Unknown σ**
- 9 Extensions to elastic net, generalized linear models, Cox model
- 10 Discussion and more recent developments work

Case of Unknown σ

Let

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

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

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

[W_j/σ^2 is asymptotically $\text{Exp}(1)$ which is the same as $\chi_2^2/2$, $(n - p) \cdot \hat{\sigma}^2/\sigma^2$ is asymptotically χ_{n-p}^2 and the two are independent.]

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

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 Example
- 8 Case of Unknown σ
- 9 Extensions to elastic net, generalized linear models, Cox model
- 10 Discussion and more recent developments work

Extensions

- **Elastic Net:** T_j is simply scaled by $(1 + \lambda_2)$, where λ_2 multiplies the ℓ_2 penalty.
- **Generalized likelihood models:**

$$T_j = [S_0 I_0^{-1/2} \mathbf{X} \hat{\beta}(\lambda_{j+1}) - S_0^T I_0^{-1/2} \mathbf{X}_{\mathcal{A}} \hat{\beta}_{\mathcal{A}}(\lambda_{j+1})] / 2$$

where S_0, I_0 are null score and information matrices, respectively. Works e.g. for generalized linear models and Cox 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 Example
- 8 Case of Unknown σ
- 9 Extensions to elastic net, generalized linear models, Cox model
- 10 Discussion and more recent developments work**

Discussion and more recent developments work

- *Software:* R library

```
covTest(larsobj, x, y),
```

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

- *More recent developments:*
 - Exact post-selection inference with the lasso (2013) (Lee et al) in arXiv
 - Post-selection adaptive inference for Least Angle Regression and the Lasso (2014) (Taylor et al) in arXiv (provides exact p-values for finite n, p)!

$$\frac{1 - \Phi(\lambda_1/\sigma)}{1 - \Phi(\lambda_2/\sigma)} \sim U(0, 1)$$

- False Discovery Rate Control for Sequential Selection Procedures, with Application to the Lasso (2013). (G'Sell et al) in arXiv
- In the works: extensions to PCA, group lasso, forward stepwise regression

References

- 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.