A significance test for the lasso

Robert Tibshirani, Stanford University

Gold medal address, SSC 2013

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

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

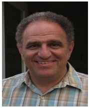
Ryan Tibshirani
Asst Prof, CMU
(PhD Student of Taylor,
Stanford 2011)





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

- Not a review, but instead some recent (unpublished work) on inference in the lasso.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 Future 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 Future 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 (**??**):

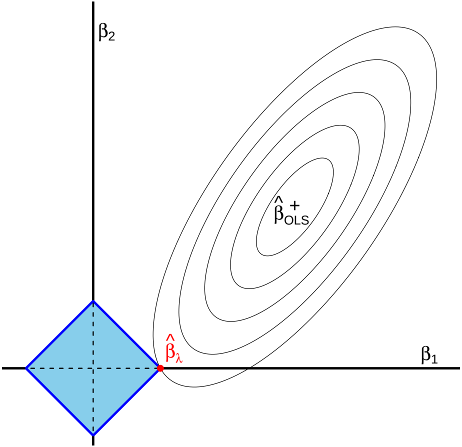
$$\underset{\beta_0 \in \mathbb{R},\ \beta \in \mathbb{R}^p}{\text{minimize}} \frac{1}{2}\|y - \beta_0 1 - X\beta\|^2 + \lambda\|\beta\|_1$$

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

The Lasso

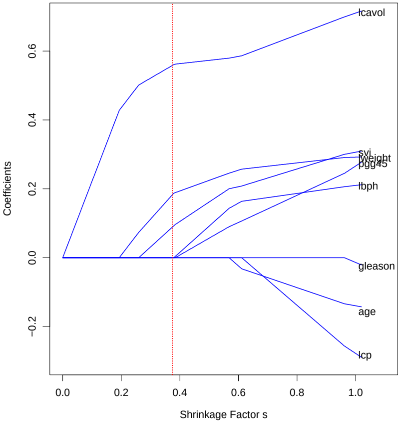
Why does ℓ_1-penalty give sparse $\hat{\beta}_\lambda$?

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



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

Old SSC logo

New SSC logo? (Thanks to Jacob Bien)

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

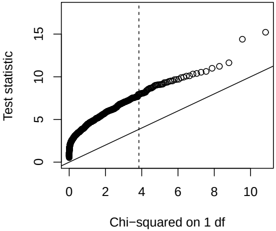
Forward stepwise regression

- 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 $y \sim N(\mu, \sigma^2 I_n)$ [**?**].

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

Stein's formula can be used to evaluate df [**?**].

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:

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

- 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) **?**, 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?

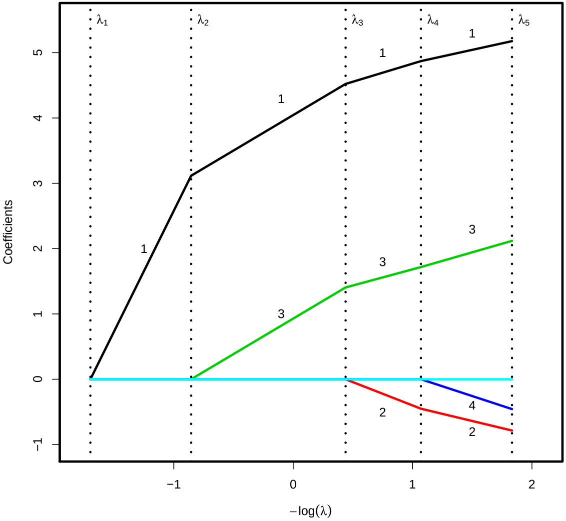
Quick review of least angle regression

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

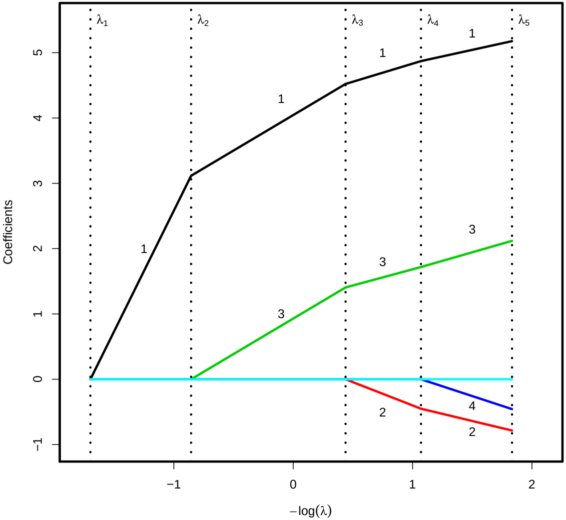


Talk Outline

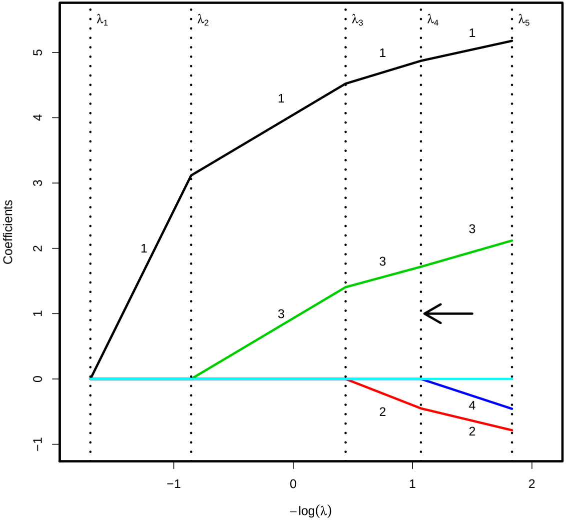
1. Review of lasso, LARS, forward stepwise
2. The covariance test statistic
3. Null distribution of the covariance statistic
4. Theory for orthogonal case
5. Simulations of null distribution
6. General **X** results
7. Example
8. Case of Unknown σ
9. Extensions to elastic net, generalized linear models,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

The covariance test statistic

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

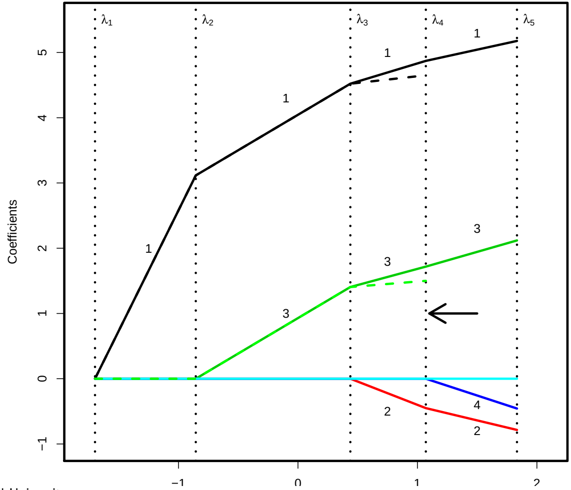


Compute covariance at λ_4: $\langle \mathbf{y}, \mathbf{X}\hat{\beta}(\lambda_4)\rangle$



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

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



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

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

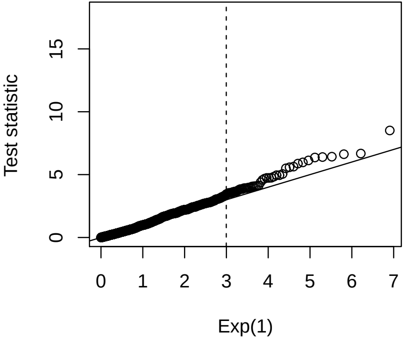
Comments on the covariance test

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

- 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

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*

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

Lemma

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

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

Proof uses a uses theorem from **?**. We were unable to find these remarkably simple results in the literature.

Heuristically, the 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 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 $\mathbf{X}$ results
7. Example
8. Case of Unknown σ
9. Extensions to elastic net, generalized linear models, Cox model
10. Discussion and Future work

Simulations of null distribution

TABLES OF SIMULATION RESULTS ARE BORING !!!!

SHOW SOME MOVIES INSTEAD

Proof sketch

We use a theorem from **?** on the asymptotic distributions of extreme order statistics.

1. Let E_1, E_2 be independent standard exponentials. There are constants a_n and b_n such that
$$W_{1n} \equiv b_n(V_1 - a_n) \longrightarrow W_1 = \log(E_1)$$
2. For those same constants put $W_{2n} = b_n(X_2 - a_n)$. Then
$$(W_{1n}, W_{2n}) \longrightarrow (W_1, W_2) = (-\log(E_1), -\log(E_1 + E_2))$$
3. 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 Exp(1).

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,

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

Jon Taylor: "Something magical happens in the math"

Sketch of proof: $k = 1$

- Assume that $y \sim N(X\beta_0, \sigma^2 I)$, and, for simplicity $\text{diag}(X^TX) = 1$. Let $U_j = X_j^T y$, $R = X^TX$,
- We are interested in $T_1 = \lambda_1(\lambda_1 - \lambda_2)/\sigma^2$. Can show that $\lambda_1 = \|X^Ty\|_\infty = \max_{j,s_j} s_j X_j^T y$ and

$$\lambda_2 = \max_{j \neq j_1,\, s \in \{-1,1\}} \frac{sU_j - sR_{j,j_1} U_{j_1}}{1 - ss_1 R_{j,j_1}}. \tag{7}$$

- Define

$$g(j,s) = sU_j \quad \text{for } j = 1, \ldots p,\; s \in \{-1,1\}. \tag{8}$$

$$h^{(j_1,s_1)}(j,s) = \frac{g(j,s) - ss_1 R_{j,j_1} g(j_1,s_1)}{1 - ss_1 R_{j,j_1}} \quad \text{for } j \neq j_1,\; s \in \{-1,1\}. \tag{9}$$

$$M(j_1,s_1) = \max_{j \neq j_1,\, s} h^{(j_1,s_1)}(j,s), \tag{10}$$

Sketch of proof— continued

- Key fact:

$$\Big\{g(j_1,s_1) > g(j,s) \text{ for all } j,s\Big\} = \Big\{g(j_1,s_1) > M(j_1,s_1)\Big\},$$

  and $M(j_1,s_1)$ is independent of $g(j_1,s_1)$. Motivated from expected Euler characteristic for a Gaussian random field [Adler, Taylor, Worsley]

- Use this to write

$$\mathbb{P}(T_1 > t) = \sum_{j_1,s_1} \mathbb{P}\Big(g(j_1,s_1)\big(g(j_1,s_1) - M(j_1,s_1)\big)/\sigma^2 > t,\ \ g(j_1,s_1) > M(j_1,s_1)\Big)$$

  Condition on M, assume $M \to \infty$, and use Mill's ratio applied to tail of Gaussian to get the result.

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 $\log p$.

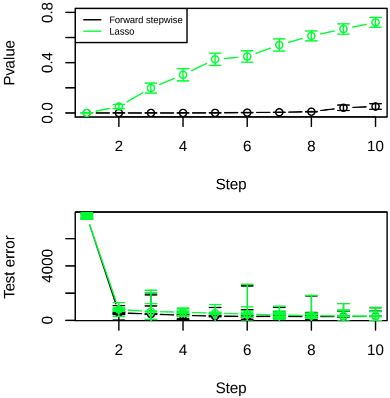
Talk Outline

HIV mutation data

$N = 1057$ samples

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

y = a measure of drug resistance



Talk Outline

Case of Unknown σ

Let

$$W_k = \Big(\langle y, X\hat{\beta}(\lambda_{k+1}) \rangle - \langle y, X_A \hat{\beta}_A(\lambda_{k+1}) \rangle \Big). \tag{11}$$

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

$$F_k = \frac{W_k}{\hat{\sigma}^2} \sim F_{2,n-p} \tag{12}$$

[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.

Talk Outline

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 Future work

Future work

- Generic (non-sequential) lasso testing problem: given a lasso fit at a knot $\lambda = \lambda_k$, what is the p-value for dropping any predictor from model? We think we know how to do this, but the details are yet to be worked out
- model selection and FDR using the p-values proposed here
- More general framework! For essentially any regularized loss+penalty problem, can derive a p-value for each event along the path. [Group lasso, Clustering, PCA, graphical models ...]
- 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.

Stepping back: food for thought

- Does this work suggest something fundamental about lasso/LARS, and the knots $\lambda_1, \lambda_2, \ldots$?
- perhaps LARS/lasso is more "correct" than forward stepwise?

  In forward stepwise, a predictor needs to win just one "correlation contest" to enter the model, and then its coefficient is unconstrained; $- >$ overfitting

  In LARS, a predictor needs to win a continuous series of correlation contests, at every step, to increase its coefficient.

The covariance test suggests that LARS is taking exactly the right-sized step.

References