“A Significance Test for the Lasso”

Lockhart R, Taylor J, Tibshirani R, and Tibshirani R

Ashley Petersen

June 6, 2013

Motivation

- Problem: Many clinical covariates – which are important to a certain medical outcome?
- Want to choose the important variables and say how important these variables are
- Bad solution: Forward stepwise regression $\rightarrow$ very anti-conservative p-values
- Better solution: Lasso with p-values from newly proposed covariance test statistic

Forward stepwise regression

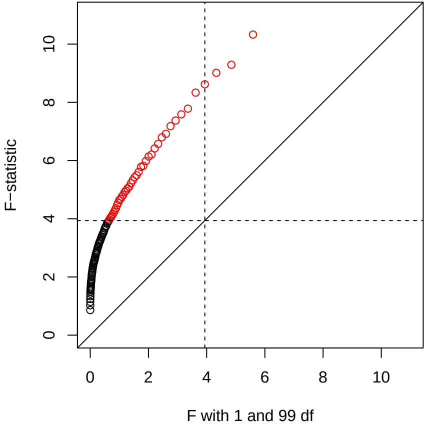
- Enter covariates into the model one at a time
- At each step choose the covariate with the largest F-statistic (smallest p-value)

$$F_k = \frac{RSS_{null} - RSS}{RSS/(n-k)}$$

- Compare to F distribution with 1 and $n-k$ df to obtain p-value

Evidence against taking those p-values seriously...

- Simulation of distribution of F-statistic for first covariate to enter model under global null ($\beta = 0$)
- $n = 100$, $p = 10$
- Type I error of 42%



Why does this matter?

- Just look at the literature – abundance of incorrect p-values
- Much desire to do adaptively fit a model and produce valid p-values

Explaining variations in prescribing costs across England

Tony Morton-Jones, Mike Pringle

TABLE II—*Regression coefficients, significances, and percentage contributions of factors used in net ingredient cost per patient multiple regression model*

Regression detail	List inflation	Standardised mortality ratio	% Pensioners	% Prepayment certificates	Constant
Regression coefficient	−0·307	0·175	0·877	0·0254	33·81
t	−8·09	9·07	6·84	4·62	5·86
Significance	<0·0001	<0·0001	<0·0001	<0·0001	<0·0001
% Variation explained	44·7	65·0	75·8	80·7	0

Framework

Consider regression setup with outcome vector $y \in \mathbb{R}^n$ with covariate matrix $X \in \mathbb{R}^{n \times p}$ and

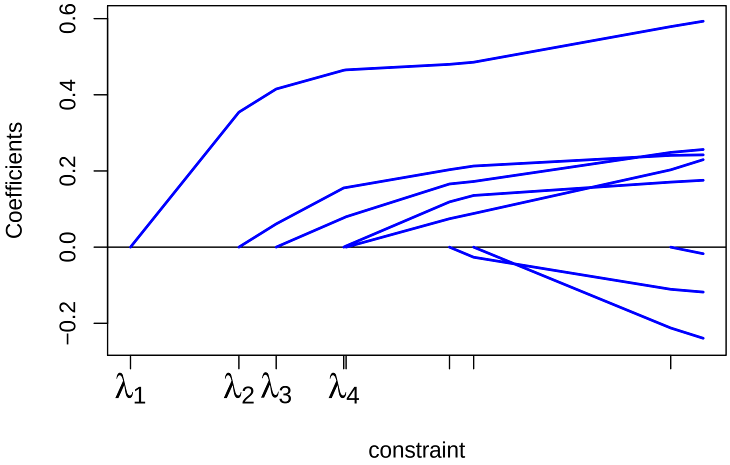
$$y = \boldsymbol{\beta} X + \boldsymbol{\epsilon} \text{ with } \boldsymbol{\epsilon} \sim N(0, \sigma^2 I).$$

The lasso estimator is obtained by finding $\boldsymbol{\beta}$ that minimizes

$$\frac{1}{2}\|y - X\boldsymbol{\beta}\|^2 + \lambda \sum_{i=1}^{p} |\beta_i|,$$

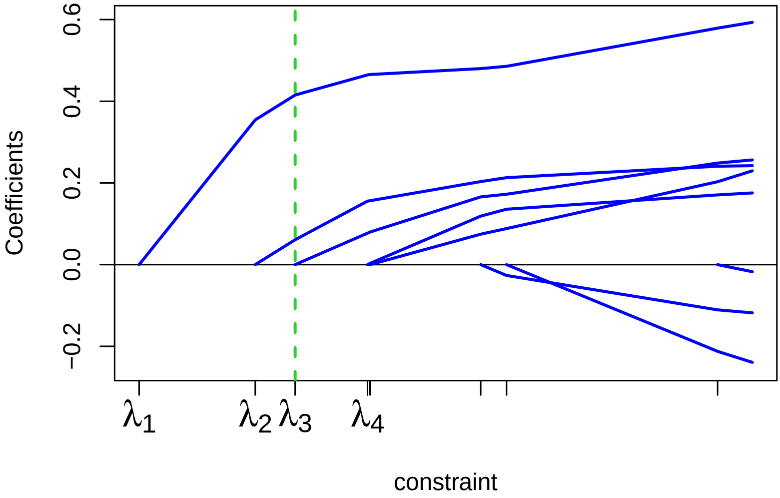
where λ is the lasso penalty.

Lasso solution path ($\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4 > \dots$)



$$\hat{\beta}_{lasso} = \arg\min_{\beta} \frac{1}{2}\|y - X\beta\|^2 + \lambda \sum_{i=1}^{p} |\beta_i|$$

Obtain p-value for covariate entering the model



Covariance test statistic

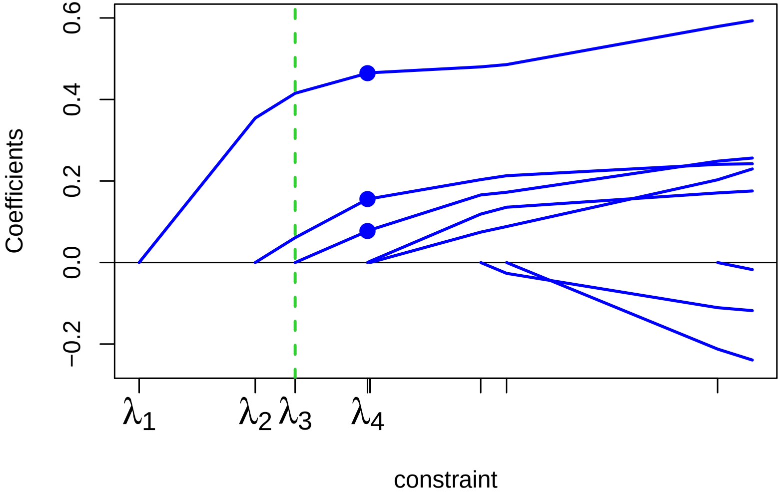
The covariance test statistic for testing the predictor that enters at the kth step is

$$
\begin{aligned}
T_k &= \frac{y^T \hat{y} - y^T \hat{y}_{null}}{\sigma^2} \\
&= \frac{y^T X \hat{\beta}(\lambda_{k+1}) - y^T X_{null} \hat{\beta}_{null}(\lambda_{k+1})}{\sigma^2}.
\end{aligned}
$$

We assume σ^2 is known for now...

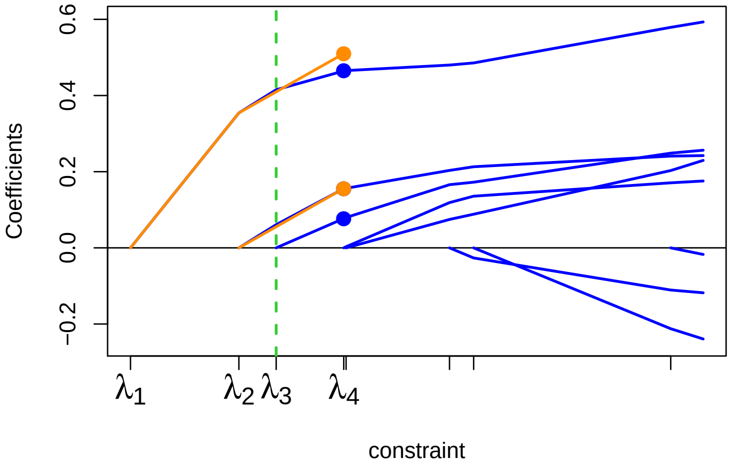
What is $\hat{y}$?

- testing that variable that enters at λ_3 has $\beta = 0$
- $\hat{y} = X\hat{\beta}(\lambda_4)$



What about $\hat{y}_{null}$?

- testing that variable that enters at λ_3 has $\beta = 0$
- $\hat{y} = X\hat{\beta}(\lambda_4)$
- $\hat{y}_{null} = X_{null}\hat{\beta}_{null}(\lambda_4)$



Asymptotic distribution

Under the null where there are k_0 truly active covariates (and they have entered the model), then

$$\begin{aligned} T_{k_0+1} &\to_d \text{Exp}(1) \\ T_{k_0+2} &\to_d \text{Exp}(1/2) \\ T_{k_0+3} &\to_d \text{Exp}(1/3) \\ &\vdots \end{aligned}$$

for orthogonal predictor matrix X.

What to do when σ^2 is unknown?

When $p < n$:

- Estimate in the usual way:

$$\hat{\sigma}^2 = \frac{RSS}{n-p} = \frac{\|y - X\hat{\beta}_{LS}\|^2}{n-p}$$

- Asymptotic distribution is now $F_{2,n-p}$

When $p \geq n$:

- Estimate from least squares fit from model selected by cross-validation
- No rigorous theory here (fingers crossed!)

Simulation setup

Generate data where

$$y = \boldsymbol{\beta} X + \boldsymbol{\epsilon} \text{ with } \boldsymbol{\epsilon} \sim N(0, I).$$

Goal: See how well Exp(1) approximates empirical distribution

- We'll use the mean, variance, and tail probability to summarize the empirical distribution

Simulation setup

Simulation 1:

- Correlated, multivariate normal predictors where $\boldsymbol{\beta} = 0$
- Consider distribution of T_1

Simulation 2:

- Correlated, multivariate normal predictors where $\boldsymbol{\beta}$ has k non-zero elements
- Consider distribution of T_{k+1}

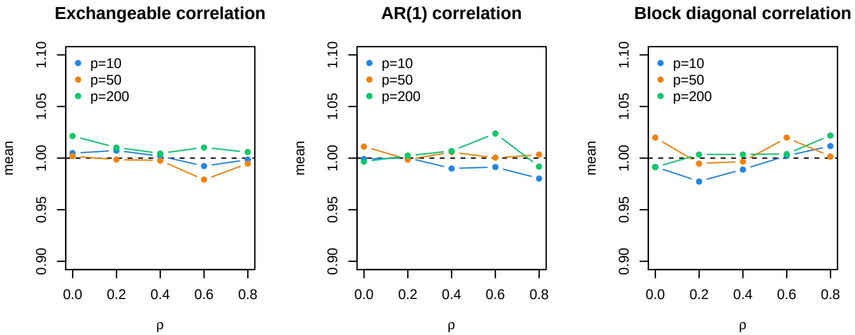
Simulation 3:

- Non-normal predictors where $\boldsymbol{\beta} = 0$
- Consider distribution of T_1

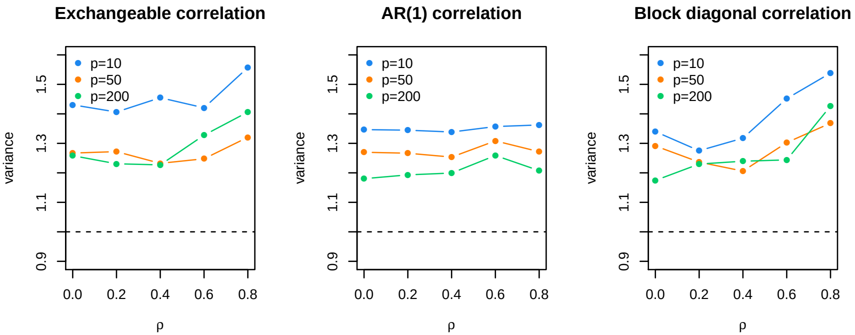
Simulation 1

- $n = 100$ and $p \in (10, 50, 200)$
- Correlated, multivariate normal predictors
- Varying correlation structure of predictors with $\rho \in (0, 0.2, 0.4, 0.6, 0.8)$
  - Exchangeable
  - AR(1)
  - Block diagonal
- $\boldsymbol{\beta} = 0$
- Consider distribution of T_1

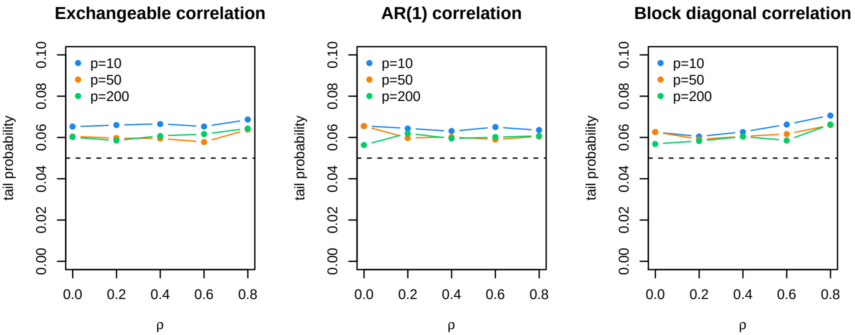
Simulation 1 results – mean



Simulation 1 results – variance



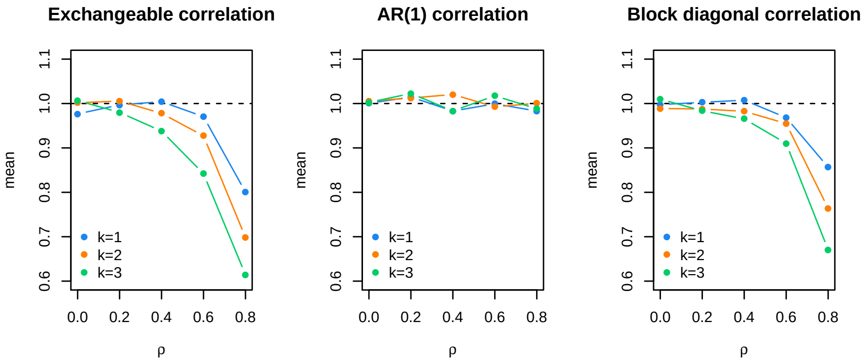
Simulation 1 results – tail probability



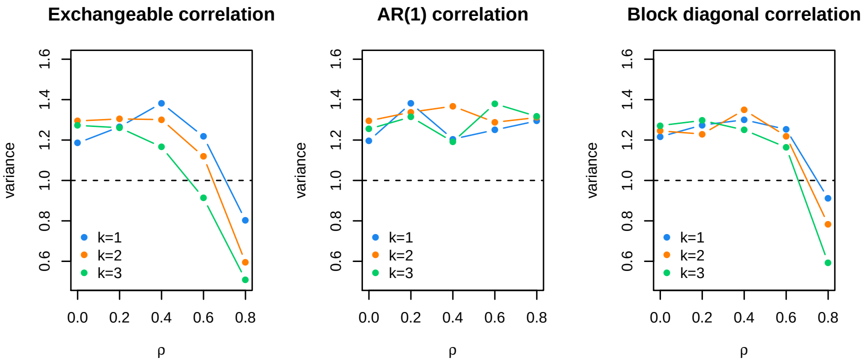
Simulation 2

- $n = 100$ and $p = 50$
- Correlated, multivariate normal predictors
- Varying correlation structure of predictors with $\rho \in (0, 0.2, 0.4, 0.6, 0.8)$
  - Exchangeable
  - AR(1)
  - Block diagonal
- $\boldsymbol{\beta}$ has k non-zero elements
- Consider distribution of T_{k+1} for $k \in (1, 2, 3)$

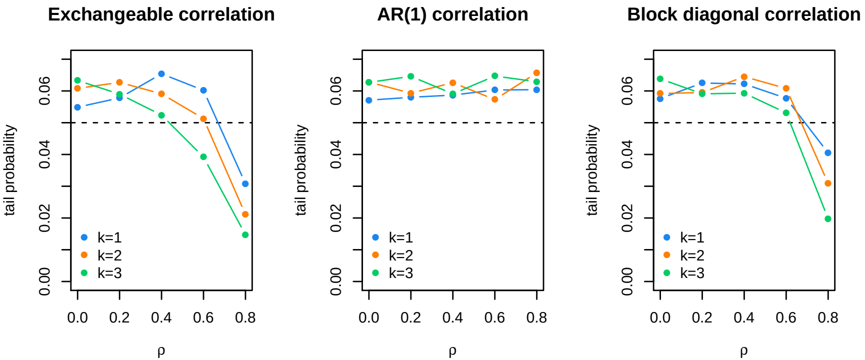
Simulation 2 results – mean



Simulation 2 results – variance



Simulation 2 results – tail probability



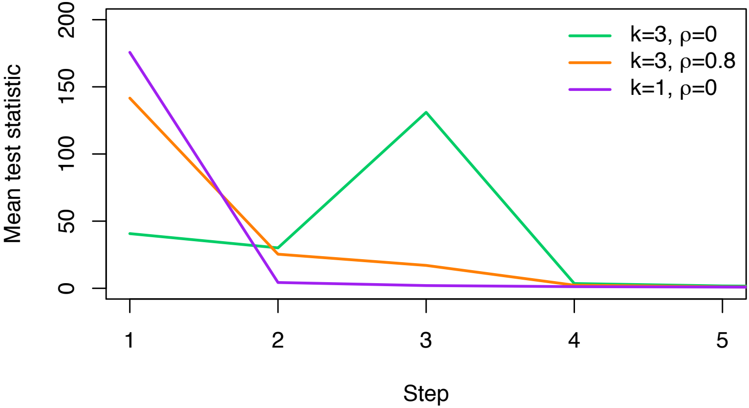
Simulation 2 results – explanation

- With high correlation, effective number of active covariates is reduced
- Test statistic does not have a distribution like that for first inactive predictor

$$
\begin{aligned}
T_{k_0+1} &\to_d \mathrm{Exp}(1) \\
T_{k_0+2} &\to_d \mathrm{Exp}(1/2) \\
T_{k_0+3} &\to_d \mathrm{Exp}(1/3) \\
&\vdots
\end{aligned}
$$

Simulation 2 results – explanation

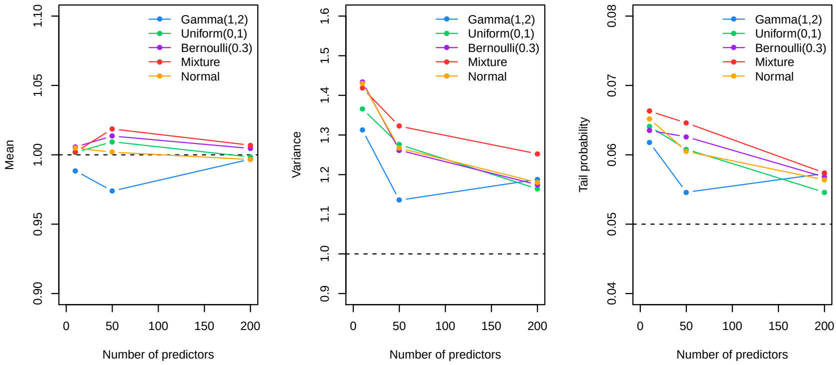
- $n = 100$ and $p = 50$ with k active covariates and correlation of ρ between predictors



Simulation 3

- $n = 100$ and $p \in (10, 50, 200)$
- Non-normal predictors
  - Gamma(1,2)
  - Uniform(0,1)
  - Bernoulli(0.3)
  - Mixture
- $\boldsymbol{\beta} = 0$
- Consider distribution of T_1

Simulation 3 results



Prostate Cancer Data

- Outcome of log PSA, 8 clinical covariates
- 67 observations

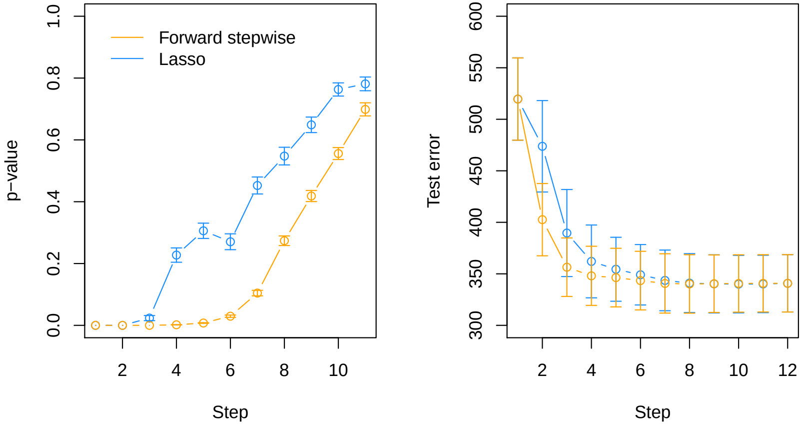
Step Number	Predictor Entered	Forward Stepwise	Predictor Entered	Lasso
1	lcavol	< 0.001	lcavol	< 0.001
2	lweight	< 0.001	lweight	0.051
3	svi	0.040	svi	0.173
4	lbph	0.045	lbph	0.929
5	pgg45	0.226	pgg45	0.352
6	lcp	0.085	age	0.650
7	age	0.142	lcp	0.050
8	gleason	0.883	gleason	0.978

Wine Quality Data

- Outcome of wine quality, 11 covariates
- 1599 observations

Step Number	Predictor Entered	Forward Stepwise	Predictor Entered	Lasso
1	alcohol	< 0.001	alcohol	< 0.001
2	volatile.acidity	< 0.001	volatile.acidity	< 0.001
3	sulphates	< 0.001	sulphates	0.001
4	total.sulfur.dioxide	0.008	total.sulfur.dioxide	0.286
5	chlorides	0.008	fixed.acidity	0.711
6	pH	0.036	chlorides	0.016
7	free.sulfur.dioxide	0.172	pH	0.568
8	fixed.acidity	0.443	free.sulfur.dioxide	0.566
9	density	0.502	density	0.824
10	residual.sugar	0.552	residual.sugar	0.848
11	citric.acid	0.952	citric.acid	0.996

Wine Quality Data



Critique

Implementation:

- More simulations needed to obtain accurate estimates
- Better to display simulation results as graphs than tables

Methods:

- Motivation: "practitioner will undoubtedly seek some sort of inferential guarantees for his or her computed lasso model"
- But...actually want inference for all coefficients from a model for a specific lasso penalty

What's the big idea?

- Use covariance test statistic to obtain p-value for covariate as it enters the lasso model
- Compare to asymptotic distribution – Exp(1) – to obtain p-values
- Reasonable performance in finite samples
- Using same data set to adaptively fit model and do inference