Inference after model selection in high dimensional linear regression
Lecture 2

Richard Lockhart, Simon Fraser University

University of Cambridge: Mini-course, Lent term, 2017

January 23, 2017

Outline

- Finish (quickly) illustrative example.
- Review LASSO / LARS path.
- Discuss testing hypotheses generated by data analysis – often after model selection.
- Large sample approach – version of extreme value theory.

Basic Lasso Tactic

- Balance Error SS against size of parameter vector β.
- Minimize

$$J(\beta) \equiv \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\beta\|^2 + \text{Penalty}(\beta).$$

- Class includes Ridge regression, SCAD, and others. LASSO:

$$\begin{aligned} J_\lambda(\beta) &= \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\beta\|^2 + \lambda \sum_i |\beta_i| \\ &= \frac{1}{2}\mathbf{Y}^T\mathbf{Y} + \frac{1}{2}\beta^T\mathbf{X}^T\mathbf{X}\beta - \mathbf{U}^T\beta + \lambda \sum_i |\beta_i| \end{aligned}$$

- Minimum depends only on $\mathbf{Y}$ only via $\mathbf{U} = \mathbf{X}^T\mathbf{Y}$.

Scaling, intercepts

- Don't shrink the intercept: $\mathbf{Y}$ and columns of $\mathbf{X}$ centred.
- You can't (shouldn't) add apples to oranges.
- The penalty does unless we standardize somehow.
- Scale $\mathbf{X}$ so that $\mathbf{X}^T\mathbf{X}$ is a correlation matrix.
- Notice β effectively grows with n, like $\sqrt{n}$.

Choosing λ

- Lots of ways to do that; not our focus.

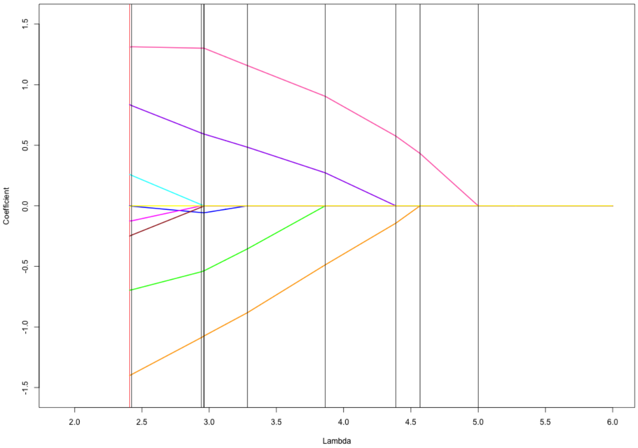
Choosing λ

- Lots of ways to do that; not our focus.
- Start λ out very large.
- For all large λ all components of $\hat{\beta}(\lambda) = 0$.
- Shrink λ gradually till one variable enters model.
- At critical value (knot) of λ, say λ_1, variable J_1 enters model.
- For λ slightly smaller than λ_1 only $\hat{\beta}_{J_1}$ is non-zero.

Choosing λ

- Lots of ways to do that; not our focus.
- Start λ out very large.
- For all large λ all components of $\hat{\beta}(\lambda) = 0$.
- Shrink λ gradually till one variable enters model.
- At critical value (knot) of λ, say λ_1, variable J_1 enters model.
- For λ slightly smaller than λ_1 only $\hat{\beta}_{J_1}$ is non-zero.
- Do we need this variable in our model?

LASSO path plot



- We test the hypothesis $\beta = 0$.
- Related to the *random* hypothesis $\beta_{J_1} = 0$?
- Bigger difference when we get to next variable.
- As we shrink λ new variables enter at knots

$$\lambda_1 > \lambda_2 > \cdots .$$

- ith variable entering is J_i with sign $S_i \in \{\pm 1\}$.
- As λ goes from λ_i to λ_{i+1}, $\hat{\beta}_{J_i}(\lambda)$ grows (linearly).
- Measure improvement of fit by change in covariance between predictor $(\mathbf{X}\hat{\beta}(\lambda))$ and $\mathbf{Y}$ between λ_i and λ_{i+1} scaled by estimate of the error variance σ^2.

Why we need to worry

- Regress log riboflavin production on variables 1278, 4003, 1516, 2564, 1588; first 5 variables in.
- Overall F test: $P = 2.2 \times 10^{-16}$.
- Individual t-test P-values: 4×10^{-5}, 5×10^{-6}, 4×10^{-3}, 1×10^{-4} and 0.34.
- But, of course, this is cherry picking.
- Our test statistic is

$$T_1 = \frac{\lambda_1(\lambda_1 - \lambda_2)}{\hat{\sigma}^2} = 24 \text{ or } 2.55.$$

- Our P-value is either 3.7×10^{-11} or 0.078.

Why we need to worry

- Regress log riboflavin production on variables 1278, 4003, 1516, 2564, 1588; first 5 variables in.
- Overall F test: $P = 2.2 \times 10^{-16}$.
- Individual t-test P-values: 4×10^{-5}, 5×10^{-6}, 4×10^{-3}, 1×10^{-4} and 0.34.
- But, of course, this is cherry picking.
- Our test statistic is

$$T_1 = \frac{\lambda_1(\lambda_1 - \lambda_2)}{\hat{\sigma}^2} = 24 \text{ or } 2.55.$$

- Our P-value is either 3.7×10^{-11} or 0.078.
- Estimation of σ is crucial and hard, I think.

More specifically: KKT conditions

Fix some $\lambda > 0$. The estimate $\hat{\beta}_\lambda$ is the vector β^* if:

$$\beta_j^* \neq 0 \Rightarrow \left.\frac{\partial J(\beta)}{\partial \beta_i}\right|_{\beta=\beta^*} = 0 \textbf{ and}$$

$$\beta_j^* = 0 \Rightarrow \left.\frac{\partial J(\beta-)}{\partial \beta_i}\right|_{\beta=\beta^*} \leq 0 \textbf{ and}$$

$$\beta_j^* = 0 \Rightarrow \left.\frac{\partial J(\beta+)}{\partial \beta_i}\right|_{\beta=\beta^*} \geq 0.$$

Here $\beta\pm$ indicate a right $(+)$ / left $(-)$ partial derivatives.

The two derivatives differ, when $\beta_j^* = 0$ by 2λ.

What are these conditions

- At β^* these derivatives take one of three forms depending on the value of β_j^*.
- For $\beta_j^* > 0$ the derivative is

$$\left(X^T X\beta^*\right)_j - U_j + \lambda = X_j^T X\beta^* - U_j + \lambda$$

- For $\beta_j^* < 0$ the derivative is

$$X_j^T X\beta^* - U_j - \lambda$$

- At $\beta_j^* = 0$ above are the right and left derivatives.

More

- Compactly. Let S_i be the sign of β_j^* and $A = \{i : \beta_j^* \neq 0\}$
- Then
$$X\beta^* = X_A\beta_A^*$$
and
$$X_A^T X_A \beta_A^* = X_A^T Y - S_A\lambda.$$
- Simplest case: $\beta^* = 0$ means that for all j we have
$$-U_j - \lambda \leq 0 \text{ and } -U_j + \lambda \geq 0$$
or
$$|U_j| \leq \lambda.$$

The first and second knots

- Except in pathological situations there is a unique $j = J_1$ such that
$$|U_{J_1}| = \max_j\{|U_j|\}.$$
- For that to fail we would have to have a pair $i \neq j$ with
$$|X_i^T Y| - |X_j^T Y| = \left|(X_i \pm X_j)^T Y\right| = 0$$
which won't happen for absolutely continuous errors unless there is a choice of signs making
$$X_i \pm X_j = 0$$
- We assume this silly design out of existence; *general position*.

The first and second knots: more

- Set $\lambda_1 = \max_i\{|U_i|\}$.
- Use J_1 for the maximizing index and S_1 for the sign of U_{J_1}.
- For $\lambda > \lambda_1$ we have $\hat{\beta}_\lambda = 0$.
- For $\lambda = \lambda_1 - \epsilon$, with $\epsilon > 0$ small enough:

$$\begin{aligned}
\hat{\beta}_{\lambda,j} =& 0 \text{ for } j \neq J_1 \\
\hat{\beta}_{\lambda,J_1} =& U_{J_1} - S_1\lambda \\
&= U_{J_1} - S_1(S_1 U_{J_1} - \epsilon) \\
&= S_1\epsilon.
\end{aligned}$$

Proof

- Check to see that this β^* satisfies the conditions.
- We are saying $A = \{J_1\}$ and solving the equation

$$X_A^T X_A \beta_A - U_{J_1} + S_1 \lambda = 0$$

  remembering that $\mathbf{X}^T\mathbf{X}$ is the identity.
- For $j \neq J_1$ the left and right derivatives are

$$X_j^T X_A \beta_A - U_j \pm \lambda$$

More Proof

- Write ρ_{jk} for the jk^{th} entry in $X^T X$.
- Note
$$\text{Cov}(U_j, U_K) = \text{Corr}(U_j, U_k) = \rho_{jk}.$$
- Left and right derivatives are on opposite sides of 0 if
$$\rho_{jJ_1}(U_{J_1} - \lambda S_1) - U_j - \lambda < 0 < \rho_{jJ_1}(U_{J_1} - \lambda S_1) - U_j + \lambda$$
which becomes
$$-\lambda(1 + \rho_{jJ_1} S_1) \le U_j - \rho_{jJ_1} U_{J_1} \le \lambda(1 - \rho_{jJ_1} S_1)$$
or
$$\max\left\{ \frac{U_j - \rho_{jJ_1} U_{J_1}}{1 - \rho_{jJ_1} S_1}, \frac{-(U_j - \rho_{jJ_1} U_{J_1})}{1 + \rho_{jJ_1} S_1} \right\} < \lambda$$

Conclusion of Proof

- So if

$$\lambda_2 \equiv \max_{j \neq J_1, s \in \{-1,1\}} \left\{ \frac{s(U_j - \rho_{jJ_1} U_{J_1})}{1 - s\rho_{jJ_1} S_1} \right\} < \lambda < \lambda_1$$

  then

$$\hat{\beta}_{\lambda j} = \begin{cases} 0 & j \neq J_1 \\ U_{J_1} - \lambda S_1 & j = J_1. \end{cases}$$

- Lockhart et al. [2014] compared the fit at λ_1 and λ_2 to get a test of the global null $\beta = 0$.
- At $\lambda = \lambda_1$ the fitted predictor is 0 and the covariance with Y is 0
- At $\lambda = \lambda_2$ the fitted predictor is $X\hat{\beta}_{\lambda_2}$ the "covariance" is

$$Y^T X \hat{\beta}_{\lambda_2}$$

Simplification

$$
\begin{aligned}
Y^T X \hat{\beta}_{\lambda_2} =& U_{J_1} \hat{\beta}_{\lambda_2 J_1} \\
=& U_{J_1} \left(U_{J_1} - \lambda_2 S_1 \right) \\
=& U_{J_1}^2 - \lambda_2 |U_{J_1}| \\
=& \lambda_1^2 - \lambda_1 \lambda_2 \\
&= \lambda_1(\lambda_1 - \lambda_2)
\end{aligned}
$$

This has to be scaled for the scale of Y so our test statistic is

$$T = \frac{\lambda_1(\lambda_1 - \lambda_2)}{\sigma^2}$$

I will discuss estimation of σ later.

Toy example: global null hypothesis true

- Approximate theory usually depends on limits.
- For p fixed that limit is normally $n \to \infty$.
- But our focus is on big p.
- Orthogonal design first. $\mathbf{X}^T\mathbf{X} = \mathbf{I}$.
- Fix $\sigma = 1$ known.
- Now $U_1, \ldots, U_p$ iid N(0,1).
- Our statistic for $i = 1$ boils down to

  $$|U_{[1]}|(|U_{[1]}| - |U_{[2]}|);$$

  subscript denotes descending order of absolute values.
- So this is an extreme value problem.

What does extreme value theory tell us?

- For a_p and b_p both more or less $\sqrt{2\log p}$ we have

$$a_p(|U_{[1]}|-b_p), a_p(|U_{[2]}|-b_p), \ldots, a_p(|U_{[K]}|-b_p)$$

  has joint extreme value limit distribution; Weissman [1978].
- Weak limit $W_1, \ldots, W_k$ has joint density

$$\exp\left(-w_1-\cdots-w_k-e^{-w_k}\right) 1(w_k<\cdots<w_1)$$

- In fact we may take

$$a_p=\sqrt{2\log p}$$

  and

$$b_p=a_p-\frac{\log\log p+\log\pi}{2a_p}.$$

Consequences

- Implication:

$$a_p(|U_{[1]}| - |U_{[2]}|) \Longrightarrow \text{Exponential}(1).$$

- And $|U_{[1]}|/a_p \to 1$ so

$$|U_{[1]}|(|U_{[1]}| - |U_{[2]}|) \Longrightarrow \text{Exponential}(1).$$

- Indeed under the global null with Gaussian errors

$$U_{[1]}|(|U_{[1]}| - |U_{[2]}|), \ldots, U_{[k]}(U_{[k]} - U_{[k+1]})$$

  converges in law to

$$E_1, E_2/2, \ldots, E_k/k$$

  where the E_i are iid standard exponential.

How to handle general $\mathbf{X}^T\mathbf{X}$ 21

- Notice $U_{[1]}$ is NOT independent of $U_{[2]}$.
- But given $J_1 = j_1$, $U_{[2]}$ computed from the U_j with $j \neq j_1$.
- So conditional law of $U_{[1]}$ given $J_1 = j_1, S_1 = 1$ AND $U_{[2]}$ is Gaussian truncated to range

$$(|U_{[2]}|, \infty).$$

- This part remains true for general designs!
- So what is conditional law of

$$U_{j_1}(U_{j_1} - \lambda_2)$$

given other U_j and $J_1 = j_1$ and $S_1 = 1$?

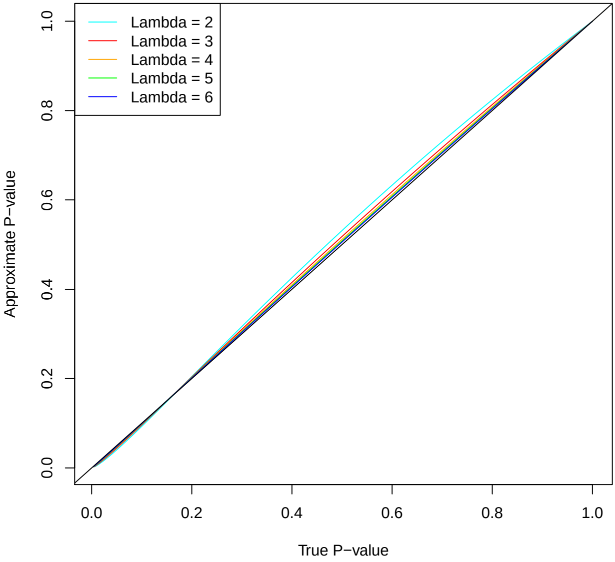
The tail of the normal distribution is exponential

- Assume $Z \sim N(0,1)$ and $\mathrm{E}(Z) = 0$ and let $\lambda \to \infty$.
- Then

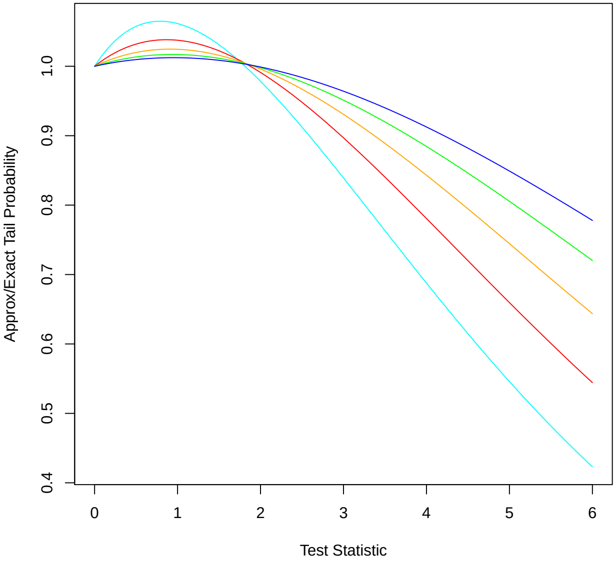
$$\lim_{\lambda \to \infty} P(Z(Z-\lambda) > x | Z > \lambda) = e^{-x} \text{ for } x > 0.$$

- Much better approx than usual extreme value theory.

Exact versus Approximate – Optimistic version



Exact versus Approximate – Pessimistic version



References

Richard Lockhart, Jonathan Taylor, Ryan J. Tibshirani, and Robert Tibshirani. A significance test for the lasso. *Ann. Statist.*, 42(2): 413–468, 04 2014. doi: 10.1214/13-AOS1175. URL http://dx.doi.org/10.1214/13-AOS1175.

Ishay Weissman. Estimation of parameters and larger quantiles based on the k largest observations. *Journal of the American Statistical Association*, 73(364):812–815, 1978. ISSN 01621459. URL http://www.jstor.org/stable/2286285.