INFERENCE IN HIGH-DIMENSIONAL LINEAR MODELS COURSE NOTES

RICHARD LOCKHART

Contents

1. Introduction 3
1.1. Motivating Analysis of Riboflavin Data 3
1.2. Some inference and modelling issues 11
2. Model selection by the LASSO 12
2.1. Scaling, intercepts 14
2.2. Asymptotic Tests following Lockhart et al. [2014] 16
2.3. Tests for the selected variable 18
2.4. Toy example: orthogonal design, global null hypothesis true 23
2.5. General design, global null hypothesis true 26
2.6. Extensions and criticism 31
3. Conditional Inference 36
3.1. Global null test 36
3.2. Corresponding confidence intervals 38
3.3. Forward Stepwise Algorithm, General Step 39
3.4. Extra ideas 43
4. De-biasing and de-sparsifying 45
4.1. Zhang and Zhang [2014] 45
4.2. van de Geer et al. [2014] and Javanmard and Montanari [2014b] 50
4.3. Theoretical Results 51
5. PoSI: Berk et al. [2013] 52
6. Limits to inference, Leeb and Pötscher [2005] 57
7. Some simulations 60
References 61

Version of 8 March 2017.

Course schedule

Note: almost certain to change as time goes by

Jan 23	Introduction: framing of issues in high dimensional inference; an example data set; some primitive inference methods; discussion of scientific contexts.
Jan 30	LASSO for model selection before inference; Unconditional limit theory for LASSO path; Lockhart et al. [2014]
Feb 6	Lockhart et al. [2014], continued.
Feb 13	Conditional inference given selection; Tibshirani et al. [2016].
Feb 20	Finish conditional inference; start debiasing/desparsifying; Zhang and Zhang [2014].
Feb 27	Debiasing/desparsifying; van de Geer et al. [2014], Javanmard and Montanari [2014b]
Mar 6	Debiasing results, PoSI, Berk et al. [2013]
Mar 13	Limits of inference, Leeb and Ptscher [2006]. Synthesis: comparison, strengths, weaknesses, my view of open issues.

1. Introduction

These notes are to a accompany a series of 8, hopefully, lectures on the general subject of inference in high dimensional linear models. They will develop over the course of Lent Term 2017. The basic data structure will be as follows. We have measurements $Y_1, \ldots, Y_n$ of some quantity which I will call the response. Associated with Y_i we have measurements $X_{i1}, \ldots, X_{ip}$ of some other quantities which I will probably call covariates, predictors, or features; any use I may happen to make of the last of these terms will be, or at least seem to be, forced. The high dimensional part will concern situations where p is large – typically larger than n but in any case substantial compared to n.

Some questions of interest to me include:

- In what scientific contexts is it important to provide inference for the parameters in a linear model?
- When we do model selection followed by inference how do we select a target of inference?
- How much trade-off must there be between model selection and inference?
- To what extent does large sample theory provide useful guidance in these problems?
- Do we want conditional or unconditional inference?

1.1. Motivating Analysis of Riboflavin Data.

I am going to use some data described in Bühlmann et al. [2014] to illustrate the sort of problem I intend to talk about for the next 8 lectures. In the example the response variable, Y, is the (base 2 logarithm of) production of riboflavin by a bacterium called *Bacillus subtilis*. The covariates are logarithms of normalized expression levels for $p = 4088$ protein coding genes. A total of $n = 71$ bacterial samples were analyzed.

The idea is that some small number of genes control the production of riboflavin. The expression data measures the extent to which a gene is 'switched-on'; for a gene which influences the production of riboflavin there ought to be a correlation how switched-on the gene is and the actual production of riboflavin.

I am going to pretend that we have a sample of n independent and identically distributed vectors $(Y_i, X_{i1}, \ldots, X_{ip})$. I will start with the basic question of whether or not there is any relationship between any of the genes and riboflavin. We will need some notation.

As usual we will stack the responses into a 71 dimensional vector $\boldsymbol{Y}$ and the covariate values into a 71×4088 matrix, denoted $\boldsymbol{X}$ with j^{th} column $\boldsymbol{X}_j$. We will write X_{ij} for the ij^{th} entry and $\boldsymbol{X}_A$ for the submatrix of $\boldsymbol{X}$ with columns whose indices j belong to $A \subset \{1, \ldots, p\}$.

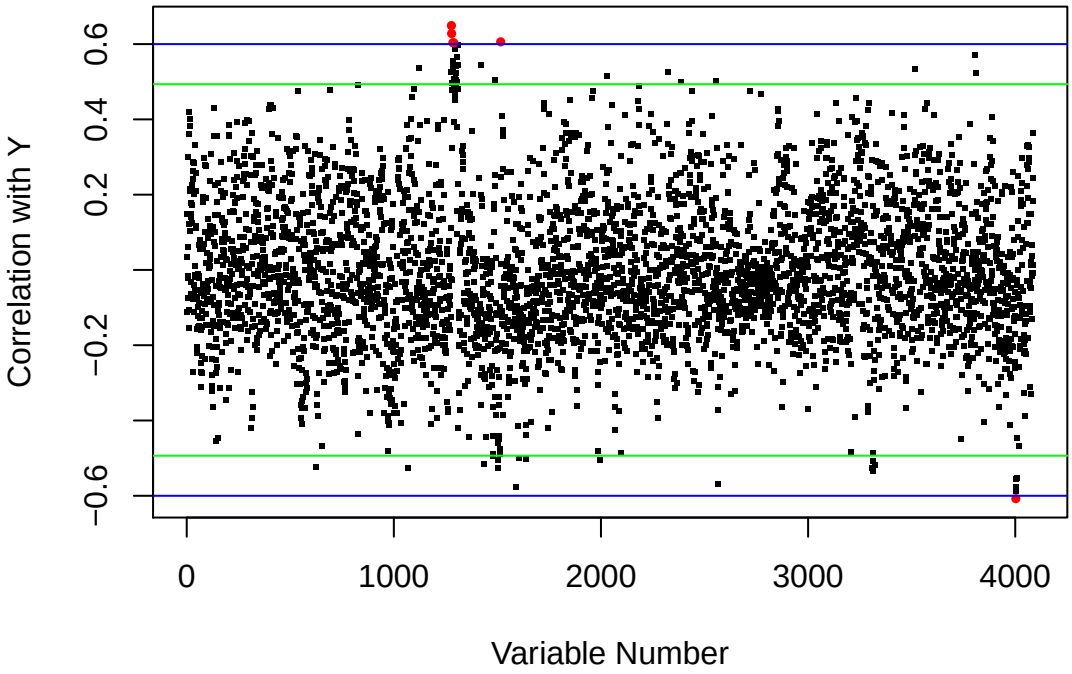
Global null hypothesis: We begin by considering the hypothesis, H_0, that Y is independent of the set of covariates. I will replace that strong null hypothesis with the weaker null hypothesis of pairwise independence. For each j we have a test statistic T_j for the null hypothesis, H_j, that Y is independent of X_j, the j^{th} covariate. Then we test the global hypothesis that H_j is true for every j. If we reject this hypothesis then of course we reject the original hypothesis of independence but there do exist (exotic) joint laws for Y and the set of covariates under which Y is independent of each subset of fewer than k (with $k < p$) of the covariates but not independent of all p. As in virtually all testing problems there is no uniformly most powerful test so we must choose where to focus our test — which alternatives we want good power for.

Even if we accept this strategy there are many tests of bivariate independence to choose from. I am simply going to use the ordinary Pearson correlation coefficient r_j between Y and the j^{th} covariate. Figure 1.1 is a plot of r_j against the index j running from 1 to 4088. I have highlighted with big red dots those points with $|r_j| > 0.6$ — just a round number chosen so that there would not be too many dots. Notice that 4 of the red dots are very close together.

Now I turn these 4088 correlations into a single test statistic by taking $\max_i\{|r_i|\}$. I computed a P-value by a variety of methods: Bonferroni correction of 1 at a time P-values from t-statistics; parametric bootstrap, taking the covariates as fixed and generating Gaussian Ys; nonparametric bootstrap, resampling Ys with replacement independently of the covariates; permutation test, where I randomly permute the Ys before computing the correlations.

The largest absolute value of a t statistic is 5.4325 for variable 1278 which has the name `YXLD_at`. All the methods I tried attached very small P-values to this test statistic as a test of the hypothesis that all 4088 correlation coefficients are 0. For the 3 simulation methods I generated 50,000 new values of Y by each method and recomputed the maximal absolute correlation. I never saw any statistic values as large at 5.4325. The parametric bootstrap and bootstrap methods each produced

FIGURE 1. Plot of the correlation of the ith covariate with Y against the index i from 1 to $p = 4088$ for the `riboflavin` data. Red dots indicate correlations larger than 0.6 in absolute value.

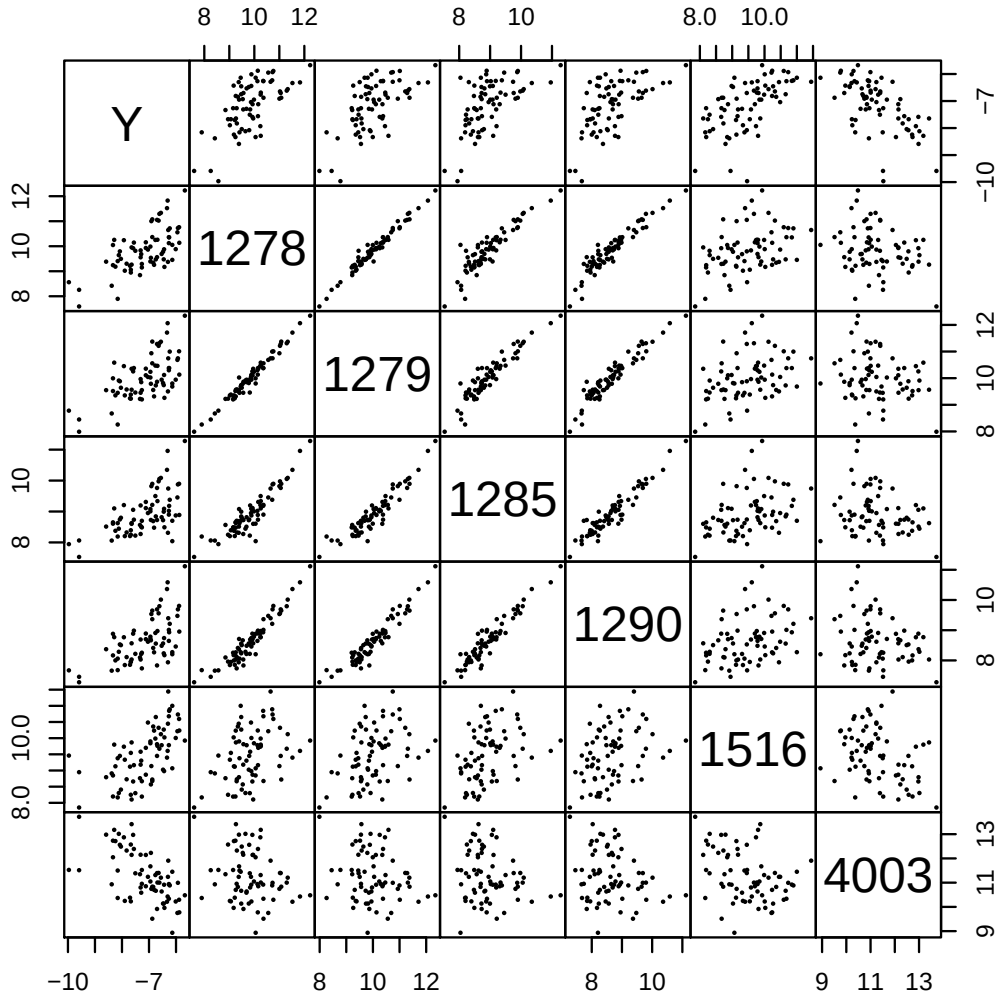


a largest absolute t statistic around 5.13 while the permutation test managed a 5.24.

The uncorrected P-value for the t statistic for variable 1278 would be 7.8×10^{-7}; after correction by multiplying by 4088 I get $P = 0.0032$ suggesting pretty strongly that at least one of these covariates is related to Y. But the Bonferroni correction is really quite conservative here. There are lots of strong correlations among the t-statistics because there are some very strong correlations among the covariates. Figure 1.1 shows all the pairwise scatterplots among the top 6 variables.

Remark: *An* exact *P-value is a random variable p which has, under some null hypothesis, a Uniform[0,1] distribution. I call p a* conservative *P-value if $P(p \leq u) \leq u$ for all $u \in [0,1]$ and the inequality is strict for some u. If $p_1, \ldots, p_m$ are*

FIGURE 2. Pairwise scatterplots of Y and the 6 covariates whose estimated correlation coefficients with Y are more than 0.6 in absolute value.



any m exact P-values (with any joint law whatsoever) then

$$P\left(\exists j : mp_j \le u\right) = P\left(m \min_{1\le j\le m}\{p_j\} \le u\right) \le \sum_{j=1}^{k} P(p_j \le u/m) = mu/m = u$$

so

$$p_{Bon} = m \min_{1\le j\le m}\{p_j\}$$

is a conservative P-value. Of course if each p_j is conservative then the conclusion still holds; the first equality just becomes an inequality.

The proof just uses the Bonferroni inequality

$$P\left(\bigcup_{i=1}^{m}\{p_j \le u/m\}\right) \le \sum_{i=1}^{m} P\left(p_j \le u/m\right)$$

If the events indicated have substantial overlaps (say because some p_j are strongly correlated with others) then the right hand side can be much larger than the left; we say Bonferroni can be very conservative.

Some commentary after seeing these plots and these statistics.

Comment 1: There is no reasonable way the response is independent of the predictors.

Comment 2: I find it hard to believe that we are confident that variable 1278 is the correct gene; distinguishing it from variable 1279 would appear to be very hard.

Here is a small easy study. Consider regressing $\boldsymbol{Y}$ on two columns $\boldsymbol{U}, \boldsymbol{V}$ with $\boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{V}^\top\boldsymbol{V} = 1$ and $\boldsymbol{U}^\top\boldsymbol{V} = 1 - \epsilon$. Generate the Y_i independently from a normal distribution with mean βU_i and variance 1. Thus the true model is

$$\boldsymbol{Y} = \boldsymbol{U}\beta + \epsilon$$

with $N(0, 1)$ errors. We will consider fitting three regression models

$$Y_i = \alpha_1 U_i + \epsilon_i,$$

$$Y_i = \alpha_2 V_i + \epsilon_i,$$

and

$$\boldsymbol{Y} = \boldsymbol{U}\beta_1 + \boldsymbol{V}\beta_2 + \epsilon$$

The middle model is wrong in the sense that the errors in that model do not have mean 0:

$$\mathrm{E}(Y_i) = \beta U_i = \alpha V_i$$

is not true for any choice of α; if it were our conditions would guarantee $\boldsymbol{U}^\top\boldsymbol{V} = \pm 1$.

When we regress $\boldsymbol{Y}$ on $\boldsymbol{U}$ without an intercept we get a fitted slope $\hat{\alpha}_1 = \boldsymbol{U}^\top\boldsymbol{Y}$ with mean β and variance 1 while if we regress $\boldsymbol{Y}$ on $\boldsymbol{V}$ without an intercept we

get fitted slope $\hat{\alpha}_2 = \boldsymbol{V}^\top \boldsymbol{Y}$ with mean $(1-\epsilon)\beta$. The covariance between these two estimates is

$$\mathrm{Cov}(\boldsymbol{U}^\top \boldsymbol{Y}, \boldsymbol{Y}^\top \boldsymbol{U}) = \boldsymbol{U}^\top \boldsymbol{V} = 1 - \epsilon.$$

Since $\boldsymbol{Y}$ has a multivariate normal distribution the pair $(\hat{\alpha}_1, \hat{\alpha}_2)$ has a bivariate normal distribution with the given means and variance-covariance.

Now consider the sort of selection algorithm I am suggesting above where we pick the covariate with the highest absolute correlation with $\boldsymbol{Y}$ as our preferred predictor. This is what I am doing when I pick out variable 1278. In the example I get the right variable if $|\hat{\alpha}_1| > |\hat{\alpha}_2|$ so I will compute this probability in the limit as $\epsilon \to 0$. I will prove this probability is $1/2$.

The probability I want is

$$\begin{aligned}\pi_\epsilon \equiv & P(0 < \hat{\alpha}_2 < \hat{\alpha}_1) + P(0 < -\hat{\alpha}_2 < -\hat{\alpha}_1) \\ & + P(0 < -\hat{\alpha}_2 < \hat{\alpha}_1) + P(0 < \hat{\alpha}_2 < -\hat{\alpha}_1).\end{aligned}$$

Let

$$\hat{\delta} = \frac{\hat{\alpha}_1 - \hat{\alpha}_2}{\sqrt{2\epsilon}}$$

Then the joint distribution of $\hat{\delta}$ and $\hat{\alpha}_2$ is bivariate normal with mean vector $(\beta\sqrt{\epsilon/2}, \beta(1-\epsilon)$, both variances equal to 1, and covariance $-\sqrt{\epsilon/2}$. As $\epsilon \to 0$ this joint distribution then converges to bivariate normal with identity covariance and means 0 and β. Rewrite the events of interest in terms of $\hat{\alpha}_2$ and δ to get

$$\begin{aligned}\pi_\epsilon = & P(\hat{\alpha}_2 > 0, \delta > 0) + P(\hat{\alpha}_2 < 0, \delta < 0) \\ & + P(0 < -\hat{\alpha}_2 < \hat{\alpha}_2 + \sqrt{2\epsilon}\delta) + P(0 < \hat{\alpha}_2 < -\sqrt{2\epsilon}\delta - \hat{\alpha}_2).\end{aligned}$$

In the limit the first two probabilities involve intersections of independent events so the first two terms converge to

$$\frac{1}{2}P(N(\beta, 1) > 0) + \frac{1}{2}P(N(\beta, 1) < 0) = \frac{1}{2}.$$

In the limit $\epsilon \to 0$ the other two terms become

$$P(0 < -\hat{\alpha}_2 < \hat{\alpha}_2) + P(0 < \hat{\alpha}_2 < -\hat{\alpha}_2) = 0$$

because the events indicated are empty. So $\lim_{\epsilon \to 0} \pi_\epsilon = 1/2$.

Remark: *if we regress* $\boldsymbol{Y}$ *on both* $\boldsymbol{U}$ *and* $\boldsymbol{V}$ *we get* $\tilde{\beta}_1, \tilde{\beta}_2$ *with a bivariate normal distribution with mean* $\beta, 0$ *and variance covariance matrix*

$$\frac{1}{2\epsilon - \epsilon^2} \begin{bmatrix} 1 & -(1-\epsilon) \\ -(1-\epsilon) & 1 \end{bmatrix}$$

which is, of course, huge for small ϵ*. Both variances are effectively* $1/(2\epsilon)$ *and the correlation converges to* -1*.*

For the data at hand think of $\boldsymbol{U}$ as column 1278 and $\boldsymbol{V}$ as column 1279. Take β to be the slope of $\boldsymbol{Y}$ regressed on variable 1278 (ignoring the selection problems these lectures are actually about) and simulate new vectors $\boldsymbol{Y}$ as described above. The correlation between $\boldsymbol{U}$ and $\boldsymbol{V}$ is 0.9845 so $\epsilon = 0.0155$. For these settings it is easy to check that the probability that the correlation with variable 1279 will be larger in absolute value than the correlation with variable 1278 is close to 1/2. In other words – for the data at hand the argument above is applicable.

When I discuss extreme value theory I hope I will deal more clearly with the probability of this event intersected with the event that the variable 1278 produces the largest correlation. For the moment I will just say the answer is essentially 1/2 under the (false, I believe) hypothesis that variable 1278 is the only variable needed to predict Y. NOTE: quite a different picture emerges if we allow for selection and take a substantially smaller value of β. More about this later.

Comment 3: I also don't believe that there is clear evidence about the number of non-zero predictors.

Bühlmann et al. [2014] use a variety of methods on the Riboflavin data. One finds no important predictors. One finds exactly variable 4003. One *marginal screening* method (roughly trying to find which predictors have unadjusted correlations with Y which could not credibly be 0) finds 53 genes when controlling the family wise (Type I) error rate at 0.05. Another, controlling the False Discovery Rate at 10% finds 375 genes. The differences between these methods reflect, principally, the different error rates each is trying to control.

More than one variable needed?

The central difficulty surrounding hypothesis testing arrives at this stage. We are now sure that at least one variable is related to the production of riboflavin. I want to test the hypothesis that none of the others is, adjusted for the one we have

found. But describing the problem that way assumes more than I have achieved. The P-value I computed does not attach to the hypothesis that $\beta_{1278} = 0$. Instead I have rejected the null hypothesis that all β_j are 0 and that is far from implying that $\beta_{1278} \neq 0$. The multi-sample splitting method of Bühlmann et al. [2014] splits the data set at random, selects a model based on one half, then uses the other half to test the hypotheses $H_{0j} : \beta_j = 0$ for each variable included in the model. Then it computes a Bonferroni adjusted P-value for that split. The process is repeated and the P values are aggregated (carefully) to control the family wise error rate

$$P(\text{Any true null hypothesis is rejected}) \leq 0.05.$$

Bühlmann et al. [2014] indicate that they found exactly 1 significant variable this way. Using `multi.split` from the `R` package `hdi` I find variable #4003.

So taking note of the obvious difficulty I go on to ask: is variable #1278 enough? Is variable #4003 enough? I need a model. I want to test the hypothesis that given X_{1278} the response Y is independent of all the other X_j. Again I will replace that with the hypothesis that each other X_j is conditionally uncorrelated with Y given X_{1278}. But this requires me to be able to condition on X_{1278} and I don't know how to do that without assumptions. So finally I assume that $(Y, X_1, \ldots, X_p)$ have a multivariate normal distribution. I regress each X_j on X_{1278} and compute the residuals. I do the same for Y. Now I have a new data set with say Y^* and X_j^* and compute 4087 correlation coefficients (or equivalently 4087 t-statistics). I get P values by bootstrapping the Y^* or permuting the Y^*. Ignoring estimation error the resampled Y^* variable is independent of the X^* variables. I find the correlation is maximized for X_{4002} and the associated P-values are estimated at 0.00052 for the bootstrap and 0.00077 for the permutation scheme. Notice that I get the variable right next door to X_{4003}. These two variables are strongly correlated and although the unadjusted correlation of X_{4003} with Y is slightly larger than that of X_{4002} with Y, this ordering is reversed after adjusting for X_{1278}.

I repeated the exercise removing the effects of X_{1278} and X_{4002} on Y and on all the other X_j and was no longer able to reject the null that all the remaining β_j are 0. Of course, not rejecting a null is a far cry from asserting its truth. I also repeated the second step of this exercise starting with variable X_{4003} (the one picked by `multi-split`). Again I found another variable was needed. The most likely candidate was X_{1278}.

My take is that there is reasonably strong evidence for the existence of more than 1 important predictor but:

- I would certainly do follow up experimental work with these genes and all those highly correlated with them.
- I think the evidence that 1278 and 4002 are the important predictors is very weak. But I suspect that one of 1278 and the things it is strongly correlated to, together with 4003 or the things it is strongly connected to, are needed. I don't know much about procedures which automatically produce groups or clusters of covariates where you try to control, for each cluster, the error rate of the statement "at least one of the variables in this cluster has a non-zero coefficient in the full model". I think procedures of that sort might be quite desirable.
- We have no clear idea what the evidence is about the size of the effects.
- Suppose I wanted to summarize my results by fitting some linear model of Y on some or all of the X_j. Should I offer confidence intervals for 4088 β_j in a regression of Y on all 4088 predictors? Should I regress Y on some subset of the 4088 – say just $\{1278, 4002\}$ and give confidence intervals for the slopes in that regression?
- I am not sure the β_j are of any real scientific interest given the pre-processing of the gene expression data.

More or less the end of what I said in Lecture 1.

1.2. **Some inference and modelling issues.** I hope the example has shown that there are some important issues to face up to. We are going to focus on a regression model of the form

$$\boldsymbol{Y} = \beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{1}$$

where we assume that conditional on $\boldsymbol{X}$ the entries in $\boldsymbol{\epsilon}$ are independent and identically distributed with mean 0 and variance σ^2. This situation arises in at least two ways:

(1) The entries in the design matrix $\boldsymbol{X}$ are actually controlled by an experimenter / data collector. In compressed sensing applications, for instance, these entries code up some expansion of some 'image' in terms of some set

of basis functions like wavelets or whatever. (I am not going to deal explicitly with any such problem but will talk about at least one deterministic design.)

(2) The vectors $(Y_i, X_{i1}, \ldots, X_{ip})$ are independent and identically distributed and the conditional expectation of Y_i given the rest is linear with homoscedastic errors. Essentially: the data are jointly multivariate normal and we have an iid sample of size n. In this case our analysis will be *conditional* on the design in the beginning at least.

The preliminary analysis I did above was focused on the second of these ideas. But I want to point out two things.

First is nature of the response. Here are the first few sorted values of 10000×2^Y.

```
> cat(10000*sort(2^y))
10 13 13 26 30 31 31 32 33 35 35
```

You see that there is considerable discreteness in Y itself and this may be worth remembering when we start to throw around assumptions like they were candy.

Second the rows of the data matrix `riboflavin` in `R` have names: the first three observations are called

```
b_Fbat107PT24.CEL
b_Fbat107PT30.CEL
b_Fbat107PT48.CEL
```

I hope the names don't mean the rows shouldn't be thought of as an iid sample (and apologize for the double negative).

2. Model selection by the LASSO

Traditionally we fit the model

$$\boldsymbol{Y} = \beta_0 \mathbf{1} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

(where $\mathbf{1}$ is a vector with all entries equal to 1) by ordinary least squares minimizing the Error Sum of Squares

$$||\boldsymbol{Y} - \boldsymbol{X}^*\boldsymbol{\beta}^*||^2$$

where $\boldsymbol{X}^*$ is the matrix with a column of 1s followed by $\boldsymbol{X}$ and $\boldsymbol{\beta}^*$ is the concatenation of β_0 and $\boldsymbol{\beta}$. This smooth function of $\boldsymbol{\beta}^*$ has gradient

$$-2\left\{\boldsymbol{X}^{*\top}\boldsymbol{X}^*\boldsymbol{\beta}^* - \boldsymbol{X}^{*\top}\boldsymbol{Y}\right\}$$

and is minimized at the least squares estimates

$$\hat{\boldsymbol{\beta}}^* = \left\{ \boldsymbol{X}^{*\top} \boldsymbol{X}^* \right\}^{-1} \boldsymbol{X}^{*\top} \boldsymbol{Y}.$$

When p exceeds $n-1$ however the matrix $\boldsymbol{X}^{*\top}\boldsymbol{X}^*$ must be singular and this method fails. We focus on situations where p is large from now on.

In general $\boldsymbol{X}^*\boldsymbol{\beta}$ is a vector in the column space of $\boldsymbol{X}^*$; any vector in that column space can be realized in this way. When $\boldsymbol{X}^{*\top}\boldsymbol{X}^*$ is singular there is nevertheless a unique vector $\hat{\boldsymbol{v}}$ in the column space of $\boldsymbol{X}^*$ minimizing

$$||\boldsymbol{Y} - \boldsymbol{v}||^2$$

over all $\boldsymbol{v}$ in the column space of $\boldsymbol{X}^*$. But there is not a unique vector β for which $\boldsymbol{v} = \boldsymbol{X}\boldsymbol{\beta}$.

One way to describe the problem is to say that the map

$$\boldsymbol{\beta} \to ||\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}||^2$$

is convex (its second derivative matrix is non-negative definite) but not strictly convex. If the rank of $\boldsymbol{X}$ is less than the number of columns of $\boldsymbol{X}$ then the null space of $\boldsymbol{X}$ is non-empty; there is a non-trivial subspace of vectors $\boldsymbol{\theta}$ with $\boldsymbol{X}\boldsymbol{\theta} = \boldsymbol{0}$. For any such θ and any $\boldsymbol{\beta}$ we see that

$$t \to ||\boldsymbol{Y} - \boldsymbol{X}\left(\boldsymbol{\beta} + t\boldsymbol{\theta}\right)||^2$$

is constant.

It turns out, however, that there are many (possibly *ad hoc* in flavour) ways to modify the error sum of squares criterion to restore strict convexity, or at least uniqueness of solutions, (except perhaps for truly pathological design matrices). The general form of a penalized error sum of squares is

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

Procedures in this class includes Ridge regression where the penalty is

$$\lambda \sum_i \beta_i^2$$

Smoothly Clipped Absolute Deviation (SCAD) which I won't define and others. I am going to focus on Least Absolute Shrinkage and Selection Operator (LASSO) because it is the only one I know even a little about.

For a given $\lambda > 0$ the LASSO estimate of $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}}_\lambda$ minimizing the penalized error sum of squares:

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

Notice that this function (and so its minimizer) depends on the data $\boldsymbol{Y}$ only via $\boldsymbol{U} = \boldsymbol{X}^\top \boldsymbol{Y}$.

I am not going to discuss uniqueness of the value of β which minimizes J_λ. The issue is studied carefully in Tibshirani [2013].

2.1. **Scaling, intercepts.** I think most scientists would regard this definition with suspicion. The columns of $\boldsymbol{X}$ are different co-variates and in most regression problems different columns will be measured in different units. Suppose for instance that Y is weight in kilograms of a person, X_1 is height in centimetres, and X_2 is age in years. If we wrote down the (silly) model

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i$$

then Y_i, α, and ϵ_i must all be measured in kilograms. The term $\beta_1 X_{1i}$ will be in kilograms as well provided β_1 is in kilograms per centimetre. Similarly β_2 has units kilograms per year. The error sum of squares has units kilograms squared. But the penalty term adds kilograms per year to kilograms per centimetre and multiplies by λ so we are adding apples to oranges; you should not do that. If the intercept α is included in the penalty then that term has units kilograms multiplied by the units of Y.

There are some natural ways out:

- Sometimes (like the riboflavin example) all the columns of $\boldsymbol{X}$ other than the column of 1s are measured in the same units. In this case the β all have units given by units of Y divided by units of an X and λ must have units of X times units of Y.
- In the penalty multiply any β_i by an estimate of scale for the variable X_i; then λ has units of Y.

- Don't shrink the intercept. This is most easily handled by estimating α by $\bar{Y}$, the mean of the responses and then centring $\boldsymbol{Y}$ and each column of $\boldsymbol{X}$ by subtracting means.
- Scale $\boldsymbol{X}$ (after centring) so that $\boldsymbol{R} \equiv \boldsymbol{X}^\top \boldsymbol{X}$ has a constant on the diagonal. I will make sure that constant is 1 so that $\boldsymbol{R}$ is a correlation matrix. Another common choice is to make the constant n so that $\boldsymbol{R}/n$ is a correlation matrix.
- People have suggested replacing the error sum of squares with its square root giving rise to the square root LASSO which minimizes

$$\frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\| + \lambda \sum_i |\beta_i|;$$

see Belloni et al. [2011]. If the X variables are all in the same units then the units of λ are units of X. If we have standardized the Xs, making them unitless then λ is also unitless. This is the potential advantage of the square root lasso; it offers the possibility of picking λ depending only on n and p and not on details of $\boldsymbol{X}$, perhaps.

- The *Scaled Lasso* (see Antoniadis [2010], Sun and Zhang [2010], and Sun and Zhang [2012]) estimates σ as well as $\boldsymbol{\beta}$ by minimizing

$$\frac{\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda \sum_i |\beta_i|$$

over both σ and $\boldsymbol{\beta}$. For any given $\boldsymbol{\beta}$ we can compute the minimizer $\tilde{\sigma}(\boldsymbol{\beta})$ via

$$\tilde{\sigma}^2(\boldsymbol{\beta}) = \frac{\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2}{n}.$$

Plugging in this value of σ we get the profile function

$$\frac{\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|}{n} + \lambda \sum_i |\beta_i|$$

which is the square root LASSO objective function (up to the scaling factor n in the denominator). The comments on units of measurement I gave for the Square Root LASSO apply here too.

When people work with the iid sampling model they often use a slightly different formulation. Like us they centre the columns of $\boldsymbol{X}$. But then they divide the error

sum of squares by the sample size n and minimize

$$\frac{1}{n}\|\boldsymbol{Y} - \boldsymbol{X}\beta\|^2 + \gamma \sum |\beta_i|.$$

This means that γ corresponds to $2\lambda/n$ in my scaling above. In the iid sampling context the matrix $\boldsymbol{X}^\top\boldsymbol{X}$ grows like n because with p fixed

$$\lim_{n\to\infty} \frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X} = \mathrm{Var}(X),$$

the population variance-covariance matrix of the covariates. If we have shave been normalized to have length n then the limit is the correlation matrix of X. If we apply our scaling so that $\boldsymbol{X}^\top\boldsymbol{X}$ has 1 on the diagonal then we have effectively divided each column by the standard deviation of that covariate multiplied by $\sqrt{n}$. This means that the corresponding entry in β has been multiplied by the same quantity. Thus in our formulation β effectively grows with n, like $\sqrt{n}$.

Assumption summary: From now on until I say otherwise I assume that $\boldsymbol{Y}$ and the columns of $\boldsymbol{X}$ have been centred and the columns of $\boldsymbol{X}$ have been standardized to have unit length. Thus $\boldsymbol{X}^\top\boldsymbol{X}$ has each diagonal entry equal to 1; it is a correlation matrix.

2.2. **Asymptotic Tests following Lockhart et al. [2014].** In order to actually use the LASSO, or any other penalized method, you have to specify λ. Many suggestions have been made but I am not going to discuss any of them. Instead I am going to describe a technique which considers the way the estimates depend on λ. That is, I am going to think about the fit as a function of λ. I will start out with λ very large and show you that for all sufficiently large λ the estimated vector $\hat{\boldsymbol{\beta}}_\lambda$ is $\mathbf{0}$. I am going to compute the infimum of that set of λ values explicitly, show that the estimate is continuous and piecewise linear in λ and show you how to compute sequentially the places where there are corners.

Here is a brief summary of our strategy which introduces some notation:

- 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, which I will denote by λ_1, variable J_1 enters our model; that is, its estimate becomes non-zero. (This value is a random variable of course.)

TABLE 1. For the riboflavin data this table shows the first 10 knots on the LASSO path. At each of the first 9 knots the active set is enlarge by the addition of the variable indicated. At $\lambda_{10} = 2.409$ variable 1588 leaves the model.

Knot	Knot value	Variable	What happened
λ_1	5.000214	1278	Added
λ_2	4.567995	4003	Added
λ_3	4.387905	1516	Added
λ_4	3.863533	2564	Added
λ_5	3.285314	1588	Added
λ_6	2.963925	624	Added
λ_7	2.960060	1312	Added
λ_8	2.942163	1502	Added
λ_9	2.424337	1639	Added
λ_{10}	2.408743	1588	Deleted

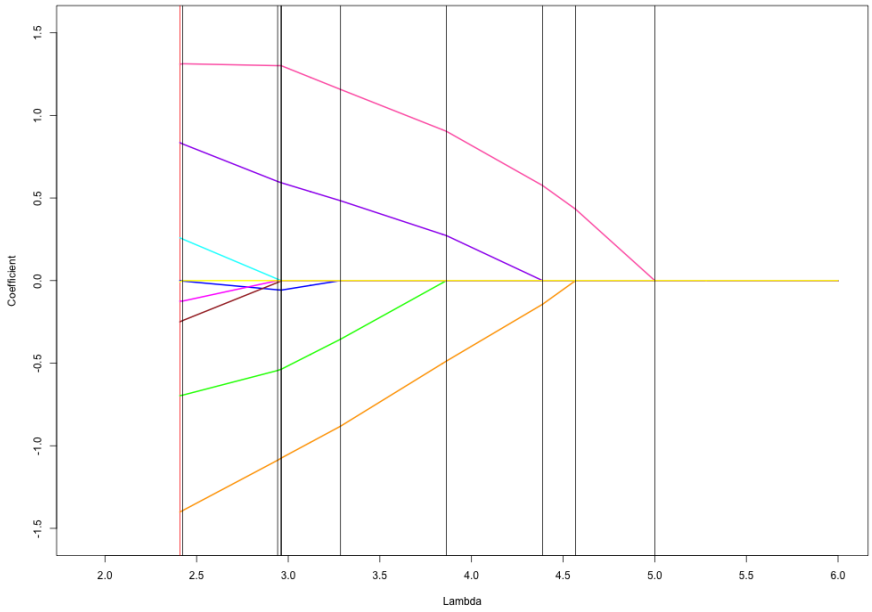
- For λ slightly smaller than λ_1 only $\hat{\beta}_{J_1}$ is non-zero.
- As we shrink λ new variables enter (or possibly leave) at knots

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

- ith variable entering is J_i with sign $S_i \in \{\pm 1\}$; this notation will become unsatisfactory when we look carefully at variables which leave the model.
- As λ goes from λ_i to λ_{i+1}, $\hat{\beta}_{J_i}(\lambda)$ grows (linearly).

For the riboflavin data after centring the columns of $\boldsymbol{X}$ and standardizing each column to have unit length we find the first 10 knots, $\lambda_1, \ldots, \lambda_{10}$ and corresponding index numbers and sign are as in Table 2.2. In Figure 2.2 I plot the estimates of the 9 coefficients involved against λ between $\lambda = \lambda_{10}$ and $\lambda = 6$. For $\lambda < \lambda_{10}$ the picture becomes quite complex; for $\lambda > \lambda_1$ we are just plotting 0. At $\lambda = \lambda_{10}$ the LASSO estimate of β_{1588} becomes 0 and that variable leaves the model Between λ_{10} and $\lambda_{11} = 2.213$ there are only 9 non-zero estimated slopes. At $\lambda_{11} = 2.213$ variable 1297 is added.

Now I show you in Figure 2.2 a frame from a movie. It shows the values of the 10 entries for $\hat{\beta}_{\lambda j}$ for j as in Table 2.2 plotted against $\lambda \in [\lambda_{10}, 6]$. The movie itself, which simply steps λ down from the right by small increments is available here. At each knot in the table you see the value of the corresponding estimated coefficient is 0 to the right and changes linearly to the left. The slopes of all these



lines change each time a variable enters a model; this is natural because now we are adjusting the slopes of each variable on a different set of covariates.

One important point is what happens with variable 1588. That variable enters the model at knot λ_5. At knot 6 or 7 the estimate for this coefficient switches from moving away from 0 (as λ decreases) to moving towards 0. Indeed at knot 10 this estimate hits 0. No variable enters at λ_{10}.

We will use the following jargon. The term *active set* refers to the set of j for which the jth coefficient is not 0. We will speak of the *true* active set as $A_0 \equiv \{j : \beta_{0j} \neq 0\}$ where the subscript 0 indicates the true parameter vector. For a given value of λ we will have an *estimated* active set

$$\hat{A}_\lambda = \{j : \hat{\beta}_{\lambda j} \neq 0\}.$$

For clarity here are some examples. For $\lambda \geq \lambda_1$ we have $\hat{A}_\lambda = \emptyset$. For $\lambda_2 \leq \lambda < \lambda_1$ we have $\hat{A}_\lambda = \{1278\}$. Finally for $\lambda_{11} \leq \lambda < \lambda_{10}$ the estimated active set consists of all the variables in Table 2.2 except 1588.

2.3. **Tests for the selected variable.** I now want to discuss our strategy for answering the question: Do we need these variables in our model? I begin by considering a test of the hypothesis $\boldsymbol{\beta} = \mathbf{0}$. We will want to understand, however,

the relation between this classical hypothesis and the *random* hypothesis $\beta_{J_1} = 0$. Our strategy is to measure the improvement of the fit when we add variable J_1 to the model using the change in covariance between the predictor $\boldsymbol{X}\hat{\boldsymbol{\beta}}(\lambda)$ and $\boldsymbol{Y}$ as λ varies between λ_1 and λ_2. This change scales with ϵ so we will scale the change in covariance by an estimate of the error variance σ^2. Lockhart et al. [2014] mostly consider a fictitious universe in which σ is known.

An aside on the nature of the model selection problem

We cannot simply look at the t statistic in the fit of Y against X_{1278} or at corresponding F tests when we consider more variables. Suppose we regress log riboflavin production on variables 1278, 4003, 1516, 2564, 1588; these are the first 5 variables which come into the model in Table 2.2. The usual overall F test gives a P-value of $P = 2.2 \times 10^{-16}$. Individual t-test P-values: 4×10^{-5}, 5×10^{-6}, 4×10^{-3}, 1×10^{-4} and 0.34.

We have already seen, however, the impact of cherry picking and discussed adjusted P-values. There are 9.5×10^15 possible regressions of Y on 5 of our 4088 covariates. So the Bonferroni corrected overall F-test P-value is 1 (the product $2.2 \times 10^{-16} \times 9.5 \times 10^{15} > 1$, that is).

The test statistic from Lockhart et al. [2014] for the first variable is

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

The word "or" reflects uncertainty about how to estimate σ^2. For the two choices we usually suggest we get a P-value which is either 3.7×10^{-11} or 0.078. That is a big range. Estimation of σ is crucial and hard, I think. I now turn to the details of our suggestion.

I am going to work my way through the Karush-Kuhn-Tucker conditions for the LASSO fit. My presentation will be elementary because our objective function J_λ is nearly differentiable and it is easy to say where it is not. Thus I will just discuss the components of the gradient vector. At values of $\boldsymbol{\beta}$ for which some component of the gradient is not defined I will just write down left and right derivatives.

At $\boldsymbol{\beta}^*$ these derivatives take one of three forms depending on the value of β_j^*.

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

$$\left(\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}^*\right)_j - U_j + \lambda = \boldsymbol{X}_j^\top \boldsymbol{X}\boldsymbol{\beta}^* - U_j + \lambda$$

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

$$\boldsymbol{X}_j^\top \boldsymbol{X}\boldsymbol{\beta}^* - U_j - \lambda$$

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

What, then, are the Karush-Kuhn-Tucker conditions? They simply say that at a solution the derivative with respect to β_j must be 0 for each non zero component and that the left and right derivatives with respect to β_j must be on opposite sides of 0 for the components j which are 0. To be precise, fix some $\lambda > 0$. The estimate $\hat{\boldsymbol{\beta}}_\lambda$ is the vector $\boldsymbol{\beta}^*$ if:

$$\begin{aligned}
\beta_j^* \neq 0 \Rightarrow \quad & \left.\frac{\partial J(\boldsymbol{\beta})}{\partial \beta_i}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} = 0 \textbf{ and} \\
\beta_j^* = 0 \Rightarrow \quad & \left.\frac{\partial J(\boldsymbol{\beta}-)}{\partial \beta_i}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \leq 0 \textbf{ and} \\
\beta_j^* = 0 \Rightarrow \quad & \left.\frac{\partial J(\boldsymbol{\beta}+)}{\partial \beta_i}\right|_{\boldsymbol{\beta}=\boldsymbol{\beta}^*} \geq 0.
\end{aligned}$$

Here $\boldsymbol{\beta}\pm$ indicate a right $(+)$ or left $(-)$ partial derivative. As I said the right and left derivatives differ, when $\beta_j^* = 0$, by 2λ. With the formulas for derivatives as above we may write the inequalities in a form which can be quite useful for theoretical purposes. The vector $\boldsymbol{\beta}^*$ is a minimizer of J_λ if

$$\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{\beta}^* - \boldsymbol{U} + \lambda\boldsymbol{\kappa} = \boldsymbol{0}$$

for a vector $\boldsymbol{\kappa}$ with

$$\beta_j^* \neq 0 \Rightarrow \kappa_j = \text{sign}\beta_j^*$$

and

$$\beta_j^* = 0 \Rightarrow |\kappa_j| \leq 1.$$

We usually write this as

$$\|\boldsymbol{\kappa}\|_\infty \leq 1$$

and

$$\beta_j^* \neq 0 \Rightarrow \kappa_j - \text{sign}\beta_j^*.$$

Compactly, let S_i be the sign of β_j^* and $A = \{i : \beta_j^* \neq 0\}$ and $\boldsymbol{S}_A$ the vector of S_i for $i \in A$. Then

$$\boldsymbol{X}\boldsymbol{\beta}^* = \boldsymbol{X}_A\boldsymbol{\beta}_A^*$$

and

$$\boldsymbol{X}_A^\top \boldsymbol{X}_A\boldsymbol{\beta}_A^* = \boldsymbol{X}_A^\top \boldsymbol{Y} - \boldsymbol{S}_A\lambda.$$

Consider now the simplest case. When is $\boldsymbol{\beta}^* = 0$? For this value we must have that for all j

$$-U_j - \lambda \leq 0 \text{ and } -U_j + \lambda \geq 0$$

or

$$|U_j| \leq \lambda.$$

Thus

$$\lambda_1 = \max_j \{|U_j|\}.$$

Now I turn my attention to finding λ_2. First I claim that 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

$$|\boldsymbol{X}_i^\top \boldsymbol{Y}| - |\boldsymbol{X}_j^\top \boldsymbol{Y}| = \left|(\boldsymbol{X}_i \pm \boldsymbol{X}_j)^\top \boldsymbol{Y}\right| = 0$$

which won't happen for absolutely continuous errors unless there is a choice of signs making

$$\boldsymbol{X}_i \pm \boldsymbol{X}_j = 0$$

If the matrix $\boldsymbol{X}$ has columns in *general position* then this does not happen for any pair $i \neq j$; the technical meaning of general position is discussed in Tibshirani [2013]. In general it means that for no $k < n$ can you write a column of $\boldsymbol{X}$ or its negative as a convex combination of k other columns (permitting you to change the sign of those other columns). A design matrix with two identical columns or one column exactly equal to minus the other is a very doubtful design.

More or less the end of what I said in Lecture 2

Recall $\lambda_1 = \max_i\{|U_i|\}$. Use J_1 for the maximizing index and S_1 for the sign of U_{J_1}. For $\lambda > \lambda_1$ we have shown that $\hat{\boldsymbol{\beta}}_\lambda = 0$. I claim there is a $\lambda_2 < \lambda_1$ such that for all $\lambda_2 \leq \lambda \leq \lambda_1$ we have

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

Proof: We will check to see that this $\boldsymbol{\beta}^*$ satisfies the conditions. Remember that

$$\lambda_1 = \max_i \{|U_i|\} = S_1 U_{J_1}.$$

For $\lambda < \lambda_1$ we see that $U_{J_1} - S_1\lambda \neq 0$ so the relevant KKT condition is given by $A = \{J_1\}$ and the equation

$$\boldsymbol{X}_A^\top \boldsymbol{X}_A \boldsymbol{\beta}_A - U_{J_1} + S_1\lambda = 0.$$

Since $\boldsymbol{X}^\top\boldsymbol{X}$ is a correlation matrix and A has only a single column this reduces to $(U_{J_1} - S_1\lambda) - U_{J_1} + S_1\lambda = 0$ which is trivial.

For $j \neq J_1$ the KKT condition is

$$\boldsymbol{X}_j^\top \boldsymbol{X}_A(U_1 - S_1\lambda) - U_j - \lambda < 0 < \boldsymbol{X}_j^\top \boldsymbol{X}_A \boldsymbol{X}_j^\top \boldsymbol{X}_A \boldsymbol{\beta}_A - U_j + \lambda.$$

Write ρ_{jk} for the jk^{th} entry in $\boldsymbol{X}^\top\boldsymbol{X}$; the choice of the letter ρ is to remind you that $\boldsymbol{X}^\top\boldsymbol{X}$ is a correlation matrix and every off diagonal entry lies in $[-1, 1]$. Note that

$$\text{Cov}(U_j, U_k) = \text{Corr}(U_j, U_k) = \rho_{jk}.$$

Then the left and right derivatives are on opposite sides of 0 if

$$-\lambda(1 + \rho_{jJ_1}S_1) \leq U_j - \rho_{jJ_1}U_{J_1} \leq \lambda(1 - \rho_{jJ_1}S_1).$$

I want to divide through by the quantities multiplying λ but I don't want to divide by 0 and I want to remember that if I divide by a negative number the direction of the inequalities would change. Since $|\rho_{jJ_1}| \leq 1$ we can divide by 0 only if $\rho_{jJ_1} \in \{-1, 1\}$. But that would mean that columns j and J_1 were perfectly correlated and, in view of our scaling, contradict our general position assumption. Notice too that $|S_1\rho_{jJ_1}| < 1$ so we will not be dividing by a negative number. We learn that if, for each $j \neq J_1$ we have

$$\max\left\{\frac{U_j - \rho_{jJ_1}U_{J_1}}{1 - \rho_{jJ_1}S_1}, \frac{-(U_j - \rho_{jJ_1}U_{J_1})}{1 + \rho_{jJ_1}S_1}\right\} < \lambda$$

then $\hat{\beta}_{\lambda j} = 0$ for $j \neq J_1$. Thus 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, as claimed,

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

Use J_2 for the maximizing value of j and S_2 for the choice of s in the definition of λ_2. Notice that S_2 will be the sign of the term $U_j - \rho_{jJ_1}U_{J_1}$ in the numerator.

Notice too that this quantity is the residual when U_j is regressed on U_{J_1} (and J_1 is treated as non-random).

Now I describe the tests of Lockhart et al. [2014]. They compared two fits at $\lambda = \lambda_2$ to get a test of the global null $\boldsymbol{\beta} = 0$. The two fits compare the active set at λ just larger than λ_1 and at λ just smaller than λ_1.

For $\lambda > \lambda_1$ the active set is empty. For this active set the LASSO fitted predictor is 0 at every λ including $\lambda = \lambda_2$; the covariance with $\boldsymbol{Y}$ is 0. For $\lambda_2 < \lambda < \lambda_1$ the active set is just $\{J_1\}$. At $\lambda = \lambda_2$ the fitted predictor is $X\hat{\boldsymbol{\beta}}_{\lambda_2}$ (which is column J_1 of $\boldsymbol{X}$ multiplied by $\hat{\beta}_{\lambda_2 J_1}$) and the "covariance" is

$$\boldsymbol{Y}^\top \boldsymbol{X} \hat{\beta}_{\lambda_2}.$$

The change in covariance then becomes

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

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.

2.4. **Toy example: orthogonal design, global null hypothesis true.** Approximate theory usually depends on limits. When I was a child we did limit theory by fixing the parameter vector $\boldsymbol{\beta}$ and so also fixing p. Then we would take a limit as $n \to \infty$ and tell the story that we were describing what would happen if we continued collecting data. Here, however, our focus is on big p. I will start with an example which can be worked out in considerable detail using extreme value theory. So now consider an orthogonal design where $\boldsymbol{X}^\top \boldsymbol{X} = \mathbf{I}$. Fix $\sigma = 1$ known. Under these assumptions the entries $U_1, \ldots, U_p$ of $\boldsymbol{U}$ are iid N(0,1). Our statistic for $i = 1$ boils down to

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

where the square brackets in the subscript denote descending order of absolute values. Thus we are studying extreme order statistics and this is an extreme value problem.

What does extreme value theory tell us? Suppose $X_1, \ldots, X_n$ are iid with continuous cdf F. The cdf of $X_{(n)} = \max\{X_i; 1 \le i \le n\}$ is $F^n(x)$ and the cdf of

$$\frac{X_{(n)} - a_n}{b_n}$$

is

$$F^n(a_n + b_n x)$$

If this sequence of distribution functions converges to a distribution $G(x)$ then the Fisher-Tippett theorem (whose final form is due to Gnedenko) says that G must be, up to a location-scale transformation one of three possibilities: Weibull, Pareto or Gumbel. In the case at hand F is the cumulative distribution function of a χ_1 random variable (the square root of a χ^2_1 variate or the absolute value of a standard normal variate). That is

$$F(x) = \max\{2\Phi(x) - 1, 0\}.$$

For this distribution the choices

$$a_n = \sqrt{2 \log n}$$

and

$$b_n = a_n - \frac{\log\log n + \log \pi}{2a_n}$$

work and the limit distribution is the standard Gumbel law

$$G(x) = \exp(-\exp(-x)).$$

Weissman [1978] extends these conclusion to the joint law of the k largest order statistics via a Poisson process approximation. Let $N(x)$ be the number of X_i which are at least x. Then $N(x)$ has a Binomial$(n, 1 - F(x))$ distribution and $N(a_n + b_n x)$ has a Binomial$(n, 1 - F(a_n + b_n x))$ distribution. The condition

$$F^n(a_n + b_n x) \to G(x)$$

for all x is equivalent to

$$n(1 - F(a_n + b_n x)) \to G(x) \tag{2}$$

for all x and then the sequence of counting processes M_n defined by

$$M_n[x, \infty) = N(a_n + b_n x)$$

converges weakly to a Poisson process with intensity $G'(x)$. That is, whenever $x_1 < \cdots < x_k$ we have

$$M_n[x_1, x_2), \ldots, M_n[x_{k-1}, x_k), M_n[x_k, \infty) \Rightarrow M[x_1, x_2), \ldots, M[x_{k-1}, x_k), M[x_k, \infty)$$

where M is an inhomogeneous Poisson Process on the line with intensity $G'(x)$. The canonical theoretical choice is $1 - F(b_n) = 1/n$ and $a_n = nf(b_n)$ but there are many asymptotically equivalent choices. To be explicit I will choose, as above,

$$a_n = \sqrt{2 \log n}$$

and

$$b_n = a_n - \frac{\log \log n + \log \pi}{2a_n}.$$

With these choices it is easy to check that (2) holds. Now fix k and real numbers

$$w_k < \cdots < w_1$$

and consider the event

$$w_k \leq a_n(|U_{[k]}| - b_n) < w_{k-1} \leq a_n(|U_{[k-1]}| - b_n) < \cdots < w_1 \leq a_n(|U_{[1]}| - b_n).$$

This is the event

$$M[w_k, w_{k-1}) \geq 1, M[w_{k-1}, w_{k-2}) = 1, \ldots, M[w_1, \infty) = 1.$$

The Poisson approximation to the probability of this event, which is valid for each integer k fixed, is used by Weissman [1978] to deduce that

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

converges in distribution to $(W_1, \ldots, W_k)$ with joint density

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

(Notice that the probability of the first event can be computed from the joint cdf of the k variables in question. That joint cdf then is seen to converge to a limit whose density I have just given.)

This conclusion has several implications. Notice that $b_n/a_n \to 1$ as $n \to \infty$. Then

$$a_n(|U_{[1]}| - |U_{[2]}|) \overset{d}{\rightsquigarrow} \text{Exponential}(1).$$

Moreover dividing the convergent quantities by a_n shows that $|U_{[1]}|/a_n \to 1$ so

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

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.

2.5. **General design, global null hypothesis true.** We now turn to the problem of a general $\boldsymbol{X}^\top \boldsymbol{X}$ (subject still to being a correlation matrix). I want to show that for $x \geq 0$ we have (again for $\sigma = 1$)

$$\lim_{n\to\infty} P(\lambda_1(\lambda_1 - \lambda_2) > x) = e^{-x}.$$

The key step is to partition this event according to the values of J_1 and S_1. That is

$$\{\lambda_1(\lambda_1 - \lambda_2) > x\} = \bigcup_{1\leq j\leq p, s_1\in\{-1,1\}} \{J_1 = j, S_1 = s_1, \lambda_1(\lambda_1 - \lambda_2) > x\}.$$

This is a disjoint union. On the event $J_1 = j, S_1 = s_1$ we have

$$\lambda_1 = sU_j$$

and

$$\lambda_2 = \max_{k\neq j, s\in\{-1,1\}} \left\{ \frac{s(U_k - \rho_{kj}U_j)}{1 - s\rho_{kj}s_1} \right\}$$

Notice that U_j is independent of the vector $\boldsymbol{V}_j$ with entries

$$V_{jk} = U_k - \rho_{kj}U_j$$

for $k \neq j$ because all these variates are jointly normal and the covariance of U_j with $U_k - \rho_{kj}U_j$ is 0.

Let F_j be the distribution of

$$W_j \equiv \max_{k \neq j, s \in \{-1,1\}} \left\{ \frac{sV_{jk}}{1 - s\rho_{kj}s_1} \right\}$$

(where I am hiding the dependence of F_j on s_1 for convenience). Then:

$$\begin{aligned} P(T > x) &= \sum_{j,s_1} P(T > x, J_1 = j, S_1 = s_1) \\ &= \sum_{j,s_1} P(s_1 U_j(s_1 U_j - W_j) > x, J_1 = j, S_1 = s_1) \end{aligned}$$

Now I rewrite the event

$$\{J_1 = j, S_1 = s_1\} = \cap_{k \neq j, s} \{sU_k \leq s_1 U_j\}$$

But if $sU_k \leq s_1 U_j$ then

$$s(U_k - \rho_{kj}U_j) \leq s_1 U_j - s\rho_{kj}U_j = s_1 U_j(1 - s\rho_{kj}s_1)$$

So

$$\cap_{k \neq j, s} \{sU_k \leq s_1 U_j\} = \{W_j \leq s_1 U_j\}.$$

We get

$$P(T > x) = \sum_{j,s_1} P(s_1 U_j(s_1 U_j - W_j) > x, s_1 U_j > W_j).$$

Since U_j and W_j are independent the conditional law of U_j given $W_j = w$ is standard normal. So

$$P(s_1 U_j(s_1 U_j - W_j) > x, s_1 U_j > W_j) = \int P(Z(Z - w) > x, Z > w) F_j(dw).$$

The tail of the normal distribution is exponential in the following sense. Assume $Z \sim N(0,1)$ and $\mathrm{E}(Z) = 0$ and let $\lambda \to \infty$. Then we can use Mill's ratio to prove

$$\lim_{\lambda \to \infty} P(Z(Z - \lambda) > x | Z > \lambda) = e^{-x} \text{ for } x > 0. \tag{3}$$

In fact define

$$u(x, \ell) = \frac{\ell + \sqrt{\ell^2 + 4x}}{2}$$

then

$$P(Z(Z - \lambda) > x | Z > \lambda) = P(Z > u(x, \lambda) | Z > \lambda) = \frac{1 - \Phi(u(x, \lambda))}{1 - \Phi(\lambda)}.$$

The limit, as $\lambda \to \infty$, of this ratio is the same, by the Mill's ratio inequalities as the limit of

$$\frac{\lambda\phi(u(x,\lambda))}{u(x,\lambda)\phi(\lambda)}.$$

Standard calculus techniques finish the proof of (3).

Let

$$\Psi(\ell) = \sup_{\lambda \geq \ell} \left| P(Z(Z-\lambda) > x | Z > \lambda) - e^{-x} \right|$$

and notice that $\Psi(\ell)$ decreases to 0 as ℓ increases. Then for any $\ell > 0$ we have

$$\begin{aligned}
&|P(T > x) - e^{-x}| \\
&\quad = \left| \sum_{j,s_1} P(s_1 U_j (s_1 U_j - W_j) > x, s_1 U_j > W_j) - e^{-x} \right| \\
&\quad = \left| \sum_{j,s_1} \left\{ P(s_1 U_j (s_1 U_j - W_j) > x, s_1 U_j > W_j) - e^{-x} P(s_1 U_j > W_j) \right\} \right| \\
&\quad = \left| \sum_{j,s_1} \int P(s_1 U_j > w) \left\{ P(Z(Z-w) > x | Z > w) - e^{-x} \right\} F_j(dw) \right| \\
&\quad \leq \sum_{j,s_1} F_j(\ell) + \Psi(\ell) \sum_{j,s_1} P(s_1 U_j > W_j > \ell).
\end{aligned}$$

The second term is bounded by $\Psi(\ell)$ which goes to 0 for any sequence $\ell = \ell_n$ tending to infinity with n. To get a theorem we assume that for each fixed ℓ the first term goes to 0. That implies the existence of a sequence $\ell = \ell_n$ increasing to infinity for which the first term converges to 0 giving

$$\left| P(T > x) - e^{-x} \right| \to 0.$$

Our theorem is

Theorem 1. *Suppose that the W_j converge to ∞ in probability uniformly in the (fairly strong) sense that for each fixed w we have*

$$\mathrm{E}(\#\{j : W_j \leq w\}) = \sum_j P(W_j \leq w) \to 0.$$

Then

$$\lim_{n,p \to \infty} P(\lambda_1(\lambda_1 - \lambda_2) > x\sigma^2) = e^{-x}.$$

More or less the end of what I said in Lecture 3

The condition looks hard to check but is actually rather mild. It really imposes a condition on the correlation structure of the U_i. Many LASSO results impose conditions on the correlations which require there to be no correlations too close to 1. What is needed here, though, is that W_j is larger than the maximum of a large number of independent normal variables whose variance is bounded below by some $\delta > 0$.

Theorem 2. *Suppose there is a $\delta > 0$ and for each p such that for each j there is a set of indices S_p of cardinality at least $d_p + 1$ with $j \in S_p$ and such that for each $i \in S_p, i \neq j$ we have*

$$\mathrm{Var}(U_i | U_k, k \in S_p, k \neq i) \geq \delta^2.$$

If

$$\frac{\log p}{d_p} \to 0$$

then the condition

$$\mathrm{E}(\#\{j : W_j \leq w\}) = \sum_j P(W_j \leq w) \to 0.$$

of the previous theorem holds.

Proof: We will find $\rho < 1$ such that

$$P(W_j \leq w) \leq \rho^{d_p}$$

for all j. The desired conclusion then follows easily since the sum is at most $p\rho^{d_p}$.

Fix j and let S_p be the corresponding set of indices. Then

$$\begin{aligned} W_j &= \max_{k \neq j, s \in \{-1,1\}} \left\{ \frac{s(U_k - \rho_{kj} U_j)}{1 - s\rho_{kj} s_1} \right\} \\ &\geq \max_{k \neq j} \left\{ \frac{|U_k - \rho_{kj} U_j|}{2} \right\} \\ &\geq \max_{k \in S_p, k \neq j} |V_k| \end{aligned}$$

where V_k is temporary shorthand for $(U_k - \rho_{kj} U_j)/2$. There are $m \equiv |S_p|$ different V_k with $m \geq d_p$ and I will simplify the notation by labelling them as $V_1, \ldots, V_m$. I now want to bound

$$\begin{aligned} P(W_j \leq w) &\leq P(|V_1| \leq w, \ldots, |V_m| \leq w) \\ &= P(|V_1| \leq w, \ldots, |V_{m-1}| \leq w) P(|V_m| \leq w, | \, |V_1| \leq w, \ldots, |V_{m-1}| \leq w). \end{aligned}$$

First I note that if G is a $N(\mu, \sigma^2)$ random variable then

$$P(|G| \le w) \le P(|G - \mu| \le w) = \Phi(w/\sigma) - \Phi(-w/\sigma).$$

For each k the variables $(U_j, V_1, \ldots, V_k)$ are multivariate normal so that the conditional law of V_k given U_j and $V_1, \ldots, V_{k-1}$ is normal with variance

$$\begin{aligned}\mathrm{Var}(U_k \mid U_j, V_1, \ldots, V_{k-1}) &= \mathrm{Var}(U_k \mid U_j, U_1, \ldots, U_{k-1}) \\ &\ge \mathrm{Var}(U_k \mid U_i, i \in S, i \ne k) \\ &\ge \delta^2.\end{aligned}$$

The first equality arises because $U_j, V_1, \ldots, V_{k-1}$ is one to one with $U_j, U_1, \ldots, U_{k-1}$. (Notationally, of course, I am temporarily labelling the entries in U so that $j > k$.) The inequality on the next line arises because adding variables to a conditioning set always decreases the variance. Remember

$$\mathrm{Var}(A|C) = \mathrm{E}(\mathrm{Var}(A|B,C)|C) + \mathrm{Var}(\mathrm{E}(A|B,C)|C) \ge \mathrm{E}(\mathrm{Var}(A|B,C)|C).$$

For jointly Gaussian variates that inner conditional variance is not random, of course, so we get

$$\mathrm{Var}(A|C) \ge \mathrm{Var}(A|B,C).$$

It follows that

$$P(|V_k| \le w | V_1, \ldots, V_k) \ge \rho \equiv \Phi(w/\delta) - \Phi(-w/\delta)$$

Then

$$\begin{aligned}P(|V_m| \le w \,||V_1| \le w, \ldots, |V_{m-1}| \le w) \\ &= \mathrm{E}\left[P(|V_k| \le w \mid V_1, \ldots, V_k) 1(|V_1| \le w, \ldots, |V_{m-1}| \le w)\right] \\ &\le \rho P(|V_1| \le w, \ldots, |V_{m-1}| \le w).\end{aligned}$$

We find inductively that

$$P(W_j \le w) \le P(|V_1| \le w)\rho^m \le \rho^{d_p}.$$

In the last inequality we used $P(|V_1| \le w) \le \rho$ which may be checked by conditioning on U_j. •

One of our findings is that this is a better approximation than usual extreme value theory. I won't talk about it in class but here is some commentary.

There are two natural ways to plot the quality of this approximation. The first, in Figure 2.5, plots the approximate P-value

$$P(Z(Z-\lambda) > v | Z > \lambda) \approx e^{-v}$$

against the exact P-value

$$P(Z(Z-\lambda) > v | Z > \lambda) = P(Z > u(v,\lambda) | Z > \lambda) = \frac{1-\Phi\{u(v,\lambda)\}}{1-\Phi(\lambda)}$$

for the values $\lambda = 2, 3, 4, 5, 6$. It will be seen that the plots lie very close to the line $y = x$. A less favourable view focuses on the quality of the approximation when the P-value is low. In Figure 2.5 I plot the ratio

$$\frac{e^{-v}}{P(Z > u(v,\lambda) | Z > \lambda)}$$

against $P(Z > u(v,\lambda) | Z > \lambda)$ for the same set of λ values. It may be worth saying that $\lambda = 6$ is very very far in the normal tail; we are conditioning on an event of probability 2×10^{-9}.

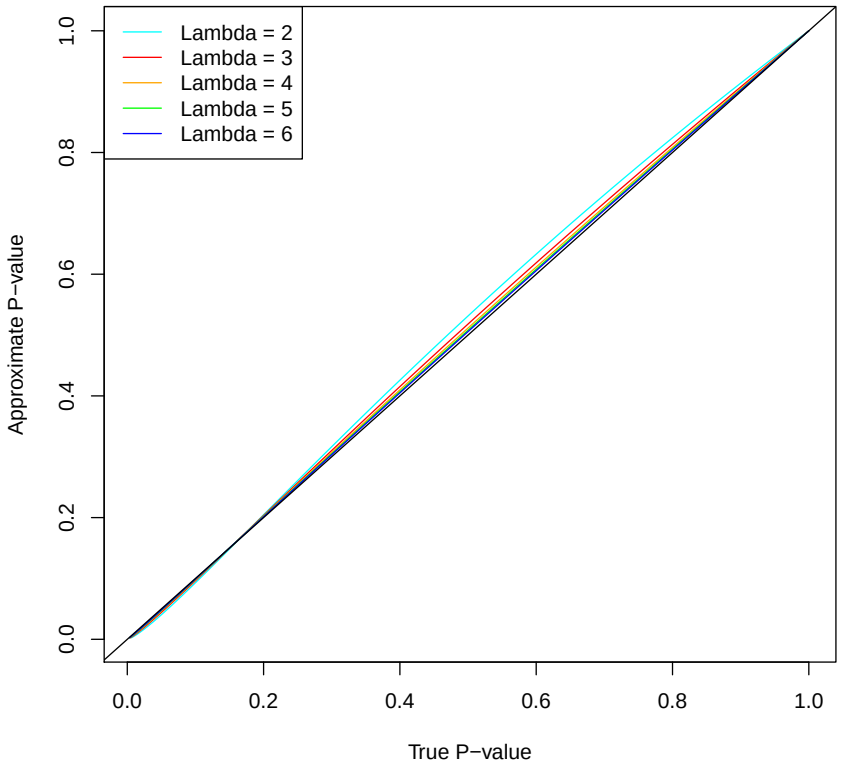
2.6. **Extensions and criticism.** Lockhart et al. [2014] extend the ideas to try to handle the case where there is some true active set A_0. They imagine following the LASSO (LARS) path down to the kth knot and assume that a variable, say variable J_k joins the active set at λ_k. Suppose A_{k-1} is the active set not including this variable and A_k is $A_{k-1} \cup \{J_k\}$. They try to test the null hypothesis that the active set A_{k-1} includes A_0. The active set A_{k-1} is random so there is some controversy over whether you can call it a hypothesis if the null hypothesis tested is random.

The test statistic compares two fitted values but now I have to be more careful. In class I said we compared two fits at λ_1 and at λ_2 but this was only right for the first knot because one of the fits was the same at λ_1 as it was at λ_2 (the fit with no predictors at all). In general we actually compare two fits at the next knot, λ_{k+1}. One fit uses the LASSO at λ_{k+1} and gives the fitted vector

$$X\hat{\boldsymbol{\beta}}_{\lambda_{k+1}} = \boldsymbol{X}_{A_k} bbeta_{\lambda_{k+1}, A_k}$$

That fit uses the larger set of predictors including J_k. The other fit uses *only* the predictors in A_{k-1} but fits the LASSO at λ_{k+1} using this restricted set of predictors.

FIGURE 3. Plot of approximate approximate tail probability $P(Z(Z-\lambda) > x | Z > \lambda) \approx \exp(-x)$ against the true tail probability $P(Z(Z-\lambda) > x | Z > \lambda) = P(Z > u(x,\lambda) | Z > \lambda)$ where Z is standard normal for values of $\lambda \in \{2,3,4,5,6\}$



Define $\tilde{\boldsymbol{\beta}}_{\lambda,A}$ to minimize

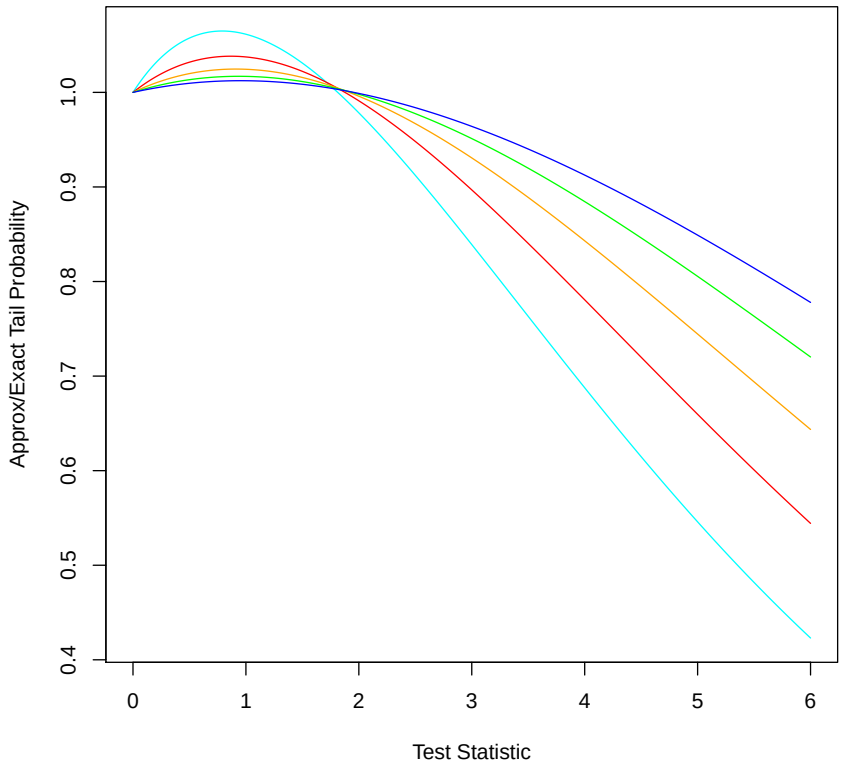
$$J_{\lambda,A}(\boldsymbol{\beta}_A) = \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}_A\boldsymbol{\beta}_A\| + \lambda \sum_{j \in A} |\beta_j|.$$

The test statistic is again the change in scaled covariance

$$T \equiv \frac{\boldsymbol{Y}^\top \boldsymbol{X}_{A_k} \hat{\beta}_{\lambda_{k+1},A_k} - \boldsymbol{Y}^\top \boldsymbol{X}_{A_{k-1}} \tilde{\beta}_{\lambda_{k+1},A_{k-1}}}{\hat{\sigma}^2}.$$

Notice that in the global null hypothesis the active set A_{k-1} is actually empty so the fitted value is 0 whether I fit at λ_1 or at λ_2. Other choices seem possible, of course, but we did not fully analyse them all. For instance we could compare, as

FIGURE 4. Ratio, as a function of x, of approximate tail probabilities $P(Z(Z-\lambda) > x | Z > \lambda) \approx \exp(-x)$ divided by true tail probabilities $P(Z(Z-\lambda) > x | Z > \lambda) = P(u(Z,\lambda) > x)$ where Z is standard normal and λ is 2, 3, 4, 5, and 6.



I suggested earlier the fit using A_{k-1} at λ_k to the fit using A_k at λ_{k+1}. I remark that the fit at λ_k can be made including J_k (as indicated in the formula) or not since the variable being added at λ_k has coefficient 0 at that exact λ value. The theorem is that T is stochastically smaller than a standard exponential variable under that null hypothesis and some strong assumptions. The most important of these assumptions is that A_{k-1} is nearly deterministic — there is a set A^* of indices which includes A_0 and has $P(A_{k-1} = A^*) \to 1$ as $n, p \to \infty$. In the LASSO literature literature there are many papers giving conditions under which the first k variables in the model are exactly the k truly active variables; these conditions would be sufficient for the theory here.

The conclusions are weaker than in the general case. Several problems arise:

- When I was computing λ_2 I began with the assertion that for all the coefficients $\beta_j, j \neq J_1$ to be 0 the left and right derivatives of the penalty with respect to each such j had to be on opposite sides of 0 at the estimate. Suppose I have computed knots $\lambda_1 > \cdots > \lambda_k$ and that A is the active set for λ just less than λ_k. For j not in the active set the inequalities mentioned are

$$\boldsymbol{X}_j^\top \boldsymbol{X}_A \left(\boldsymbol{X}_A^\top \boldsymbol{X}_A\right)^{-1} (\boldsymbol{U}_A - \lambda S_A) - U_j - \lambda \leq 0 \leq \boldsymbol{X}_j^\top \boldsymbol{X}_A \left(\boldsymbol{X}_A^\top \boldsymbol{X}_A\right)^{-1} (\boldsymbol{U}_A - \lambda S_A) - U_j + \lambda.$$

  I then rewrite them as

$$-\lambda(1 + \boldsymbol{X}_j^\top \boldsymbol{X}_A \left(\boldsymbol{X}_A^\top \boldsymbol{X}_A\right)^{-1} S_A) \leq U_j - \boldsymbol{X}_j^\top \boldsymbol{X}_A \left(\boldsymbol{X}_A^\top \boldsymbol{X}_A\right)^{-1} \boldsymbol{U}_A \leq \lambda(1 - \boldsymbol{X}_j^\top \boldsymbol{X}_A \left(\boldsymbol{X}_A^\top \boldsymbol{X}_A\right)^{-1} S_A)$$

  If G_1, G_2 are jointly Gaussian with mean 0 and partitioned covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

  then the (true) residual, $G_1 - \mathrm{E}(G_1|G_2)$ when G_1 is regressed on G_2 is

$$G_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} G_2$$

  with variance

$$\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}.$$

  So you see that the central term in the chain of inequalities is just such a residual. When I was studying the global null hypothesis I divided through by the coefficient of λ on each side of this equation to get a lower bound for λ. In general, however, the term

$$1 \pm \boldsymbol{X}_j^\top \boldsymbol{X}_A \left(\boldsymbol{X}_A^\top \boldsymbol{X}_A\right)^{-1} S_A$$

  could be negative for one of the two sign choices. Dividing through gives an upper bound for λ, not a lower bound. This complication is dealt with at length in Lockhart et al. [2014] but it will be seen that it is assumed out of existence in the large sample theory given there.

- When computing the next knot, λ_{k+1} it may happen, as in the Riboflavin data set at knot 10, that rather than adding a variable, a variable must be deleted. If A_k is the active set for λ just smaller than λ_k then the estimated coefficients at such a λ are

$$\left(\boldsymbol{X}_A^\top \boldsymbol{X}_A\right)^{-1}\left(\boldsymbol{U}_A - \lambda \boldsymbol{S}_A\right) = \left(\boldsymbol{X}_A^\top \boldsymbol{X}_A\right)^{-1}\boldsymbol{U}_A - \lambda\left(\boldsymbol{X}_A^\top \boldsymbol{X}_A\right)^{-1}\boldsymbol{S}_A.$$

  It is possible that for some index $j \in A$ the jth component of the first term has the opposite sign to the jth component. In this case that component of the estimate is moving towards 0, not away, as λ shrinks. The estimate is linear in λ so set it equal to 0 and solve. Now for each j in the active set you have computed a value, say λ_j^{del} at which the coefficient would be 0. If this is more than λ_k ignore it. Let λ^{del} be the maximum of all those λ_j^{del} which are less than λ_k. The theory in the paper makes a very strong assumption that prevents (with probability tending to 1) any deletion before two more additions to A under the null hypothesis.
- The estimate $\tilde{\beta}_A$ considers only variables in A and not, in particular, the variable J_k which joined at λ_k. Because J_k is not included in the active set, when computing $\tilde{\beta}_A$ it could happen that one of the coefficients in the LASSO solution using only the variables in A hits 0 in the interval $\lambda_{k+1} < \lambda < \lambda_k$. Again the paper makes the assumption that the chance this happens is negligible in the limit
- The actual test statistic takes the form

$$\frac{C\lambda_k(\lambda_k - \lambda_{k+1})}{\hat{\sigma}^2}.$$

  The number C is computable from the values of J_k, S_k, and the design matrix $\boldsymbol{X}$ (using only the columns in the active set after adding J_k). Conditional on these values the variable $\sqrt{C}\lambda_k$ is standard normal but the statistic does not then have the conditional form $Z(Z-M)$ given $Z > M$ with Z standard normal. The result is that the statistic is actually stochastically smaller than exponential; this should be expected to result in diminished power.

In reading papers it is important to look for things that might be regarded as weaknesses. Lockhart et al. [2014] has a few, in my view:

- Its handling of estimation of σ is unconvincing.

- The paper illustrates the mechanics of computing P-values with this exponential distribution using some prostate cancer data in which $n = 67$ and $p = 8$. It is important to remember that the theory developed is making an approximation to the tails of a normal distribution. For the global null in this problem we are looking at the largest of 8 Gaussian's and making an extreme value computation. Our limit theory requires $p \to \infty$.
- The use of conservative limits is quite unattractive. Suppose you have a test statistic, T, which you pretend has an exponential distribution with mean 1 under some null hypothesis. If the distribution is exponential but the real mean is $1/2$ and we observe $T = 2.5$ then our computed P-value using the standard exponential is 0.082 while if we use the correct exponential mean we get a P-value of 0.0067. In such a case we would be giving away a lot of power.

3. Conditional Inference

The weaknesses I enumerated above of Lockhart et al. [2014] led to a number of papers from my collaborators. In particular they noticed the following key point. We are approximating $P(Z(Z-M) > x|Z > M) \approx \exp(-x)$. But we can compute exactly $P(Z(Z-M) > x|Z > M, M = m)$ so if we could just condition on the value of M we would not need to rely on the exponential limit.

Tibshirani et al. [2016] implements this conditioning idea.

3.1. Global null test.

Here is the simplest version. Consider a general design with the standardization as before; assume $\sigma = 1$ is known. At $\lambda = \lambda_1$ variable J_1 is entered into the model with sign S_1. At $\lambda = \lambda_2 < \lambda_1$ a second variable enters the model We begin with testing the global null hypothesis $\boldsymbol{\beta} = 0$ using our test statistic

$$T = \lambda_1(\lambda_1 - \lambda_2).$$

If we observe $T = t_{\text{obs}}$ then the P-value is naturally

$$P_{H_o}(T > t_{\text{obs}})$$

This probability was approximated above. We are treating $\boldsymbol{X}$ as fixed (our model holds conditionally on $\boldsymbol{X}$) so the P-value depends on $\boldsymbol{X}$ in principle and is hard

to compute. We could, however, try to compute a conditional P-value

$$p_{\text{cond}} = P_{H_o}(T > t_{\text{obs}} \mid J_1 = j_{1,\text{obs}}, S_1 = s_{1,\text{obs}})$$

which is also hard. On the event $J_1 = j_1, S_1 = s_1$ however we have

$$\lambda_2 = \max_{j \neq j_1, s \in \{-1,1\}} \left\{ \frac{s\,(U_j - \rho_{jj_1} U_{j_1})}{1 - s\rho_{jj_1} s_1} \right\}$$

which is independent of U_{j_1}. Thus when the null hypothesis $\boldsymbol{\beta} = 0$ holds we have

$$P(S_1 U_{J_1} > z \mid J_1 = j_1, S_1 = s_1, \lambda_2 = \ell) = P(Z > z | Z > \ell)$$

where Z is standard normal. Write this as

$$P(S_1 U_{J_1} > z \mid J_1 = j_1, S_1 = s_1, \lambda_2 = \ell) = \frac{1 - \Phi(z)}{1 - \Phi(\ell)}$$

where Φ is the standard normal cumulative. So given $J_1 = j_1, S_1 = s_1, \lambda_2 = \ell$ the random variable

$$p = \frac{1 - \Phi(s_1 U_{j_1})}{1 - \Phi(\ell)} = \frac{1 - \Phi(\lambda_1)}{1 - \Phi(\lambda_2)}$$

has a uniform distribution on the unit interval. Since that conditional distribution is free of J_1, S_1, and λ_2 we see that p is a valid unconditional P-value for our global null hypothesis.

Theorem 3. *Suppose $\boldsymbol{Y}$ is an n-dimensional response vector and $\boldsymbol{X}$ is an $n \times p$ non-random design matrix. Assume that*

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

where the entries in $\boldsymbol{\epsilon}$ are iid standard normal, the columns of $\boldsymbol{X}$ have been centred and standardized and are in standard position. Under the null hypothesis the quantity

$$p = \frac{1 - \Phi(\lambda_1)}{1 - \Phi(\lambda_2)}$$

has a Uniform[0,1] distribution.

The P-value p is a test statistic which is its own P-value; a one sided test rejects at level α if $p < \alpha$. Tibshirani et al. [2016] call this a *spacings* test.

Notice the assumptions: centring, scaling and general position. This new procedure appears to be unrelated to the earlier test statistic $T = \lambda_1(\lambda_1 - \lambda_2)$ but given $J_1 = j_1$, $S_1 = s_1$ and $\lambda_2 = \ell$ the event $T > x$ is equivalent to $\lambda_1 > u(\lambda_2, x)$

where as before

$$u(x,\ell) = \frac{\ell + \sqrt{\ell^2 + 4x}}{2}.$$

So a transformation of T by its conditional survival function is equivalent to a transformation of λ_1 by its conditional survival function (given J_1, S_1 and λ_2).

Also, when λ_2 is large and p is not too close to 0 the value of p must be well approximated by $\exp(-\lambda_1(\lambda_1-\lambda_2))$, the P-value from the previous covariance test.

3.2. **Corresponding confidence intervals.** Now consider the question of confidence intervals. The idea is simple. Suppose only variable 1 has a non-zero coefficient. If we regress $\boldsymbol{Y}$ on the "correct" covariate $\boldsymbol{X}_1$ then the least squares estimate of β_1 is simply $U_1 = \boldsymbol{X}_1^\top \boldsymbol{Y}$ after our scaling. In the model selection world, however, we will likely only be interested in this estimate if we select variable 1 by the LASSO – that is, we find $J_1 = 1$. So now I will imagine doing the following. Run the LASSO for one step and find variable J_1. Assume that variable 1 is the only variable with a non-zero regression coefficient, $\beta_1 = \mathrm{E}(U_1) = \mathrm{E}(\boldsymbol{X}_1^\top \boldsymbol{Y})$.

When I actually run LASSO, however, I will get variable J_1 entering with sign S_1. If I see $J_1 = j_1$ I will be hoping that j_1 is an active variable (in fact I rather hope it is the only active variable). Then I will want a confidence interval for β_{j_1}. If j_1 were the unique active variable then I would be interested in the mean of $\boldsymbol{X}_{j_1}^\top \boldsymbol{Y}$ which would be β_1. So I imagine that having observed $J_1 = j_1$ I will consider getting a confidence interval for $\mathrm{E}(\boldsymbol{X}_{j_1}^\top \boldsymbol{Y}) = \boldsymbol{X}_{j_1}^\top \mathrm{E}(\boldsymbol{Y})$.

In the situation I described where only variable 1 is active then the expected value of $\boldsymbol{Y}$ is simply $\beta_1 \boldsymbol{X}_1$ so I will look for a confidence interval for

$$\boldsymbol{X}_{j_1}^\top \boldsymbol{X}_1 \beta_1 = \rho_{j_1,1}\beta_1.$$

This quantity is the regression of the true mean of $\boldsymbol{Y}$ on $\boldsymbol{X}_{j_1}$. I acknowledge that some people will feel let-down. We are not getting a confidence interval for β_1. But if your model selection picked variable 12 there is no way you can maintain that you are interested in the coefficient of a variable you have eliminated from your model. Instead we just say – I am making the best linear approximation I can to predicting $\boldsymbol{Y}$ from $\boldsymbol{X}_{j_1}$ by predicting its mean as well as I can.

Theorem 4. *Suppose $\boldsymbol{Y}$ is an n-dimensional response vector with independent normally distributed entries (given $\boldsymbol{X}$) with variance 1 (given $\boldsymbol{X}$) and $\mathrm{E}(Y_i \mid \boldsymbol{X}) = \theta_i$. Define $\psi_j = \boldsymbol{X}_j \boldsymbol{\theta}$. Assume that the columns of $\boldsymbol{X}$ have been centred and*

standardized and are in standard position. Then

$$P(\lambda_1 > x | J_1 = j_1, S_1 = s_1, \{U_j - \rho_{jj_1} U_{j_1}; j \neq j_1\}) =$$
$$P(s_1(Z + \psi_j) > x | s_1(Z + \psi_j) > \ell) = \frac{1 - \Phi(x - s_1\psi_j)}{1 - \Phi(\ell - s_1\psi_j)}$$

where $\ell = \lambda_2$ *(computed from* $\{U_j - \rho_{jj_1} U_{j_1}; j \neq j_1\}$*). Thus*

$$\frac{1 - \Phi(\lambda_1 - S_1\psi_{J_1})}{1 - \Phi(\lambda_2 - S_1\psi_{J_1})}$$

is a pivot; it has a Uniform[0,1] distribution.

So if we solve the inequalities

$$\frac{\alpha}{2} \leq \frac{1 - \Phi(\lambda_1 - S_1\psi_{J_1})}{1 - \Phi(\lambda_2 - S_1\psi_{J_1})} \leq 1 - \frac{\alpha}{2}$$

to get

$$c_L \leq \psi_{J_1} \leq c_U$$

then (c_L, c_U) is a level $1 - \alpha$ confidence interval for Ψ_{J_1}. Notice that the target of the interval is random. As a result you might feel it is not a confidence interval because Ψ_{J_1} is not a parameter. But I argue it plays the role of a confidence interval in a perfectly natural way.

More or less the end of what I said in Lecture 4
From here on the notes are quite raw.

3.3. **Forward Stepwise Algorithm, General Step.** The forward stepwise method for variable selection selects variables to add to the model sequentially. It begins by regressing $\boldsymbol{Y}$ on each column $\boldsymbol{X}_j$ of $\boldsymbol{X}$ and selecting variable $J_1 = j_1$ if the that variable minimizes the Error Sum of Squares — or equivalently maximizes the Regression Sum of Squares, the squared length of the fitted vector. Then a second variable J_2 is added to minimize the Error Sum of Squares over all two variable models including variable J_1. This procedure continues. I will discuss this procedure under the condition that each model contains an intercept term.

The result is that I may assume that the columns of $\boldsymbol{X}$ have been centred; remember fitted values depend only on the column space of the design matrix. Similarly fitted values and the ESS are unaffected if the columns of $\boldsymbol{X}$ are rescaled so I will again take $\boldsymbol{X}$ to be a correlation matrix.

If we regress $\boldsymbol{Y}$ on $\boldsymbol{X}_j$ the centred fitted vector is then

$$\boldsymbol{X}_j \boldsymbol{X}_j^\top \boldsymbol{Y} = \boldsymbol{X}_j U_j.$$

The squared length of this vector is

$$U_j^2 = \left(\boldsymbol{X}_j^\top \boldsymbol{Y}\right)^2$$

The variable $J_1 = j_1$ maximizes this length if

$$s_1 U_{j_1} > s U_k$$

for $s_1 = sign(U_{j_1})$ and all $k \neq j_1$ and signs $s \in \{-1, 1\}$. This is a set of $2(p-1)$ inequalities:

$$\begin{aligned}
(s_1 \boldsymbol{X}_{j_1} + \boldsymbol{X}_1)^\top \boldsymbol{Y} &> 0 \\
(s_1 \boldsymbol{X}_{j_1} - \boldsymbol{X}_1)^\top \boldsymbol{Y} &> 0 \\
&\vdots \\
(s_1 \boldsymbol{X}_{j_1} - \boldsymbol{X}_p)^\top \boldsymbol{Y} &> 0
\end{aligned}$$

where we omit the two rows where the second index would be j_1. Thus the event $J_1 = j_1$ and $S_1 = s_1$ is the event

$$\Gamma_1 \boldsymbol{Y} \geq 0$$

where Γ_1 is a $2p-2 \times n$ matrix with rows $(s_1 \boldsymbol{X}_{j_1} \pm \boldsymbol{X}_k)^\top$. The matrix Γ_1 depends on j_1 and s_1 though the notation hides that. The matrix is not random.

Now when we add a second variable to the model the new fitted vector for the model including the variable $J_1 = j_1$ and some variable $k \neq j_1$ is the projection of $\boldsymbol{Y}$ on the linear span of $\boldsymbol{X}_{j_1}$ and $\boldsymbol{X}_k$. This span is unaffected if we replace $\boldsymbol{X}_k$ by its standardized residual when we regress $\boldsymbol{X}_k$ on $\boldsymbol{X}_{j_1}$; for clarity I replace $\boldsymbol{X}_k$ by

$$\boldsymbol{X}_k^* = (\boldsymbol{X}_k - \boldsymbol{X}_k^\top \boldsymbol{X}_{j_1} \boldsymbol{X}_{j_1})/c$$

and $c = \sqrt{1 - \rho_{k,j_1}^2}$ is the length of the vector $\boldsymbol{X}_k - \boldsymbol{X}_k^\top \boldsymbol{X}_{j_1} \boldsymbol{X}_j$. The change in the Error Sum of Squares due to adding k is just the squared length of

$$\boldsymbol{X}_k^* \boldsymbol{Y}$$

Pick $J_2 = j_2$ which maximizes this squared length and let $S_2 = s_2$ be the sign of $\boldsymbol{X}_{j_2}^* \boldsymbol{Y}$. The condition that j_2 is the maximizer with sign s_2 is just

$$s_2 \frac{\left(\boldsymbol{X}_{j_2} - \boldsymbol{X}_{j_2}^\top \boldsymbol{X}_{j_1} \boldsymbol{X}_{j_1}\right)^\top \boldsymbol{Y}}{\sqrt{1 - \rho_{j_2,j_1}^2}} > s \frac{\left(\boldsymbol{X}_k - \boldsymbol{X}_k^\top \boldsymbol{X}_{j_1} \boldsymbol{X}_{j_1}\right)^\top \boldsymbol{Y}}{\sqrt{1 - \rho_{k,j_1}^2}}$$

for all $k \notin \{j_1, j_2\}$ and $s \in \{-1, 1\}$.

Again we may write the event $J_1 = j_1, S_1 = s_1, J_2 = j_2, S_2 = s_2$ in the form

$$\Gamma_2 \boldsymbol{Y} > 0$$

where the matrix Γ_2 adds $2p - 4$ rows to Γ_1. This process may be continued to give explicit matrices Γ_k for which the event

$$J_1 = j_1, S_1 = s_1, \ldots, J_k = j_k, S_k = s_k$$

is exactly the event

$$\Gamma_k \boldsymbol{Y} > 0.$$

Definition: If Γ is a $q \times n$ matrix and $\boldsymbol{u}$ is a q-vector then the

$$\{\boldsymbol{x} \in \mathbb{R}^q : \Gamma \boldsymbol{x} \geq \boldsymbol{u}\}$$

is a *polytope* which I am going to call a *polyhedron*. The inequality means that each entry in $\Gamma \boldsymbol{x}$ is at least as big as the corresponding entry in $\boldsymbol{u}$.

Theorem 5 (Polyhedral Lemma)**.** *Suppose $\boldsymbol{Y}$ has a multivariate normal distribution in $\mathbb{R}^n$ with mean $\boldsymbol{\theta}$ and covariance matrix $\boldsymbol{\Sigma}$. Suppose that $\boldsymbol{\Gamma}$ is some $q \times n$ matrix with jth row $\boldsymbol{\gamma}_j^\top$. Let $\boldsymbol{v}$ be some n-vector. The covariance between $\boldsymbol{\Gamma Y}$ and $\boldsymbol{v}^\top \boldsymbol{Y}$ is*

$$\boldsymbol{\Gamma \Sigma v}.$$

Define

$$\rho_j = \boldsymbol{\gamma}_j^\top \boldsymbol{\Sigma v} (\boldsymbol{v}^t \boldsymbol{\Sigma v})^{-1}.$$

and

$$V^{\rm lo}(\boldsymbol{Y}) = \max_{j:\rho_j>0}\left\{\frac{u_j - \left(\boldsymbol{\gamma}_j^\top \boldsymbol{Y} - \rho_j \boldsymbol{v}^\top \boldsymbol{Y}\right)}{\rho_j}\right\}$$

$$V^{\rm hi}(\boldsymbol{Y}) = \min_{j:\rho_j<0}\left\{\frac{u_j - \left(\boldsymbol{\gamma}_j^\top \boldsymbol{Y} - \rho_j \boldsymbol{v}^\top \boldsymbol{Y}\right)}{\rho_j}\right\}$$

$$V^0(\boldsymbol{Y}) = \min_{j:\rho_j=0}\left\{\boldsymbol{\gamma}_j^\top \boldsymbol{Y} - \rho_j \boldsymbol{v}^\top \boldsymbol{Y} - u_j\right\} = \min_{j:\rho_j=0}\left\{\boldsymbol{\gamma}_j^\top \boldsymbol{Y} - u_j\right\}$$

Then $\boldsymbol{v}^\top \boldsymbol{Y}$ *is independent of* $\left(V^{\rm lo}(\boldsymbol{Y}), V^{\rm hi}(\boldsymbol{Y}), V^0(\boldsymbol{Y})\right)$ *and the event* $\boldsymbol{\Gamma Y} > \boldsymbol{u}$ *is the event*

$$V^{\rm lo}(\boldsymbol{Y}) < \boldsymbol{v}^\top \boldsymbol{Y} < V^{\rm hi}(\boldsymbol{Y}), \quad V^0(\boldsymbol{Y}) > 0$$

Proof: The independence is just a matter of computing covariances. The terms $\boldsymbol{\gamma}_j^\top \boldsymbol{Y} - \rho_j \boldsymbol{v}^\top \boldsymbol{Y}$ are the residuals when $\boldsymbol{v}^\top \boldsymbol{Y}$ is regressed on $\boldsymbol{\gamma}^T \boldsymbol{Y}$ so the covariances are 0. The condition $\rho_j = 0$ just amounts to saying $\boldsymbol{\gamma}_j^\top \boldsymbol{Y}$ has covariance 0 with $\boldsymbol{v}^T \boldsymbol{Y}$. The inequalities are elementary algebra. You should check back to the arguments surrounding the covariance test for the first variable in.

It remains to establish the equivalence of the events. We have

$$\begin{aligned} u_j \le \gamma_j^\top \boldsymbol{Y} &\Leftrightarrow u_j - \rho_j \nu^\top \boldsymbol{Y} \le \gamma_j^\top \boldsymbol{Y} - \rho_j \boldsymbol{v}^\top \boldsymbol{Y} \\ &\Leftrightarrow \rho_j \nu^\top \boldsymbol{Y} \ge u_j - \left(\gamma_j^\top \boldsymbol{Y} - \rho_j \boldsymbol{v}^\top \boldsymbol{Y}\right) \end{aligned}$$

For any j for which $\rho_j > 0$ we divide by ρ_j to see that the indicated inequality is equivalent to

$$\nu^\top \boldsymbol{Y} \ge \frac{u_j - \left(\gamma_j^\top \boldsymbol{Y} - \rho_j \boldsymbol{v}^\top \boldsymbol{Y}\right)}{\rho_j}.$$

For a j for which $\rho_j < 0$ we also divide through by ρ_j but now the direction of the inequality is reversed; the indicated inequality is equivalent to

$$\nu^\top \boldsymbol{Y} \le \frac{u_j - \left(\gamma_j^\top \boldsymbol{Y} - \rho_j \boldsymbol{v}^\top \boldsymbol{Y}\right)}{\rho_j}$$

when $\rho_j < 0$. When $\rho_j = 0$ the original inequality is unchanged. All the inequalities for j such that $\rho_j > 0$ hold if and only if $\boldsymbol{v}^\top \boldsymbol{Y} \ge V^{\rm lo}$ and the corresponding inequalities for j such that $\rho_j < 0$ all hold if and only if $\boldsymbol{v}^\top \boldsymbol{Y} \le V^{\rm hi}$. •

The polyhedral lemma tells you that the conditional distribution of $\boldsymbol{v}^T \boldsymbol{Y}$ given the event that $\boldsymbol{Y}$ lands in the polyhedron given and the value of the residual when $\boldsymbol{v}^\top \boldsymbol{Y}$ is regressed on $\boldsymbol{\Gamma Y}$ is normal between two limits. It shows you how

to compute those limits. The independence means that the conditioning did not change the variance of $\boldsymbol{v}^\top \boldsymbol{Y}$ which is

$$\boldsymbol{v}^\top \boldsymbol{\Sigma} \boldsymbol{v}.$$

3.4. **Extra ideas.** Tibshirani et al. [2016] contains a number of other ideas.

- Whenever the polyhedral lemma applies you can get tests and confidence intervals for $\boldsymbol{v}^\top \theta$ using the Gaussian distribution with mean $\boldsymbol{v}^\top \theta$ conditioned to lie in the interval $[V^-, V^+]$.
- Sometimes the matrix Γ can be made smaller at the price of replacing the deterministic vector $\boldsymbol{u}$ (which is often just full of 0s) by a random vector $\boldsymbol{U}$; this vector should not be confused with the notation $\boldsymbol{U} = \boldsymbol{X}^\top \boldsymbol{Y}$ already in use. The simplification is acceptable is $\boldsymbol{U}$ is also uncorrelated with $\boldsymbol{v}^\top \boldsymbol{Y}$; typically that will happen when it is computed from the same vector of residuals.
- The paper details construction of the matrix Γ for LASSO and for LARS (like LASSO but no deletions are allowed — the LASSO and LARS paths coincide up to the first variable deletion for LASSO).
- The paper gives various approximate versions, for higher k, of the spacings test.
- You could also compute the conditional distribution of λ_1 given only $J_1 = j_1$ *without* condition on S_1. This conditions on $\boldsymbol{Y}$ belonging to the union of two polytopes and the computations are harder.
- Sequential testing procedures are described but not studied in detail.
- There is a discussion by Larry Brown and Kory Johnson which expresses doubt that the methods have any value.
- The methods are implemented in `selectedInference` in `R`.
- Estimation of σ is again swept under the rug. The `R` package implements two methods. One uses an over estimate of σ which is just the sample standard deviation of $\boldsymbol{Y}$. This can be very conservative, of course. The other selects λ by cross-validation in the LASSO framework and then uses a suggestion for that context. In the `RIboflavin` example that estimate of σ is based on a fit to some 30 predictors. I feel, but cannot prove, that the estimate is too liberal – that is, too small – when I am thinking about whether or not to put in a single predictor or a second predictor. Of course

the ordinary least squares estimate using the selected variable will be too liberal as well, typically.

The conditional inference method here accepts the model selection method as a good one and then tries simply to give confidence intervals for parameters of interest. Part of the difficulty lies in the fact that a good model selection procedure must use lots of the information in the data to select the model and have little left over for the fitting part. One might try to think about writing the log-likelihood in the form

$$\log\left\{P(\hat{M}=M;\boldsymbol{\beta})\right\}+\log\left\{f_{\boldsymbol{Y}}(y|\hat{M}=M;\boldsymbol{\beta})\right\}.$$

The Hessian of this is the total information about $\boldsymbol{\beta}$ and a strongly peaked first term would seem to rule out a strongly peaked second term.

Here is a short list of some of the ways in which this work has been expanded:

- Lee et al. [2016] is concerned with the same theoretical structure – inference conditional on a polyhedron but applied to LASSO at a single value of λ for instance rather than as a tool along the LASSO / LARS path.
- Fithian et al. [2014] is concerned with the trade-off between conditioning on more information and the power of post-selective inference. In Tibshirani et al. [2016] we condition on the model selected and on the bounds V^+ and V^-. If we could condition solely on the model selected we would expect to be able to get more testing power. This paper introduces the term *data carving*, arguing using classic optimal hypothesis testing theory that data splitting is inadmissible. Typically data splitting partitions data D at randm into D_1 and D_2, uses D_1 to perform model selection and then does inference on the model parameters using D_2 only. Fithian et al. [2014] notes that this amounts to conditioning on D_1 and argues for conditioning only on the model selected using D_1, leaving an analyst free to re-used the rest of the information in D_1 as well as all the information in D_2 for inference within the selected model.
- Fithian et al. [2015] suggests an alternate, and more powerful, approach to using the P-values of Tibshirani et al. [2016] and provides an explicit method for controlling the False Discovery Rate.
- Tian and Taylor [2015] suggests adding noise to the model selection step in the process in order to save more information for post-selection inference.

4. De-biasing and de-sparsifying

In this section I am going to look at van de Geer et al. [2014], Javanmard and Montanari [2013], Javanmard and Montanari [2014b], Javanmard and Montanari [2014a], Javanmard and Montanari [2015], and Zhang and Zhang [2014]. They too study inference after using LASSO to select a model. Having found the LASSO estimate $\hat{\boldsymbol{\beta}}_\lambda$ for some λ they adjust the estimate to try to remove its bias and then give confidence intervals for the coefficients. The three papers overlap a great deal but got published in widely separated venues at around the same time.

A central inferential issue is always what you want to estimate. The papers given here assume that there is a vector $\boldsymbol{\beta}_0$, depending on n and p (which depends itself on n) such that

$$\boldsymbol{Y} = \boldsymbol{X}\beta_0 + \boldsymbol{\epsilon}$$

with iid entries in the error vector. The goal is to give confidence intervals for all the entries in $\boldsymbol{\beta}_0$ or at least for any specific entry in $\boldsymbol{\beta}_0$. There is a sense in which one is abandoning the model selection goal of the LASSO but there is also the realistic understanding that when one selects a model with a given estimated active set $\hat{A}$ there may well be $j \notin \hat{A}$ for which $\beta_{0j} \neq 0$.

These three papers make very similar suggestions. They produce confidence intervals for components β_j by finding estimates whose estimation error $(\hat{\beta}_j - \beta_j)$ has the form

$$\boldsymbol{a}_j^\top \boldsymbol{Y} + o_P\left(\sqrt{\boldsymbol{a}_j^\top \boldsymbol{a}_j}\right)$$

with

$$\boldsymbol{a}_j^\top \boldsymbol{a}_j = O(1/n).$$

They achieve this by correcting the bias in some estimator. Zhang and Zhang [2014] proceed by finding a linear estimator which has approximately the properties of ordinary least squares, then using the LASSO to remove the bias of that estimate. Javanmard and Montanari [2014b] and van de Geer et al. [2014] start from the LASSO estimates and adjust them to remove bias and get an error expansion as above.

4.1. Zhang and Zhang [2014].

I am going to start with Zhang and Zhang [2014] because in this paper the formulas are developed one entry at a time and the presentation seems simplest to me. In any least squares problem where $\boldsymbol{X}^\top \boldsymbol{X}$ is

not singular the jth entry in the ordinary least squares estimate of $\boldsymbol{\beta}$ is of the form

$$\hat{\beta}_{j\text{OLS}} = \frac{\boldsymbol{x}_j^{\perp\top}\boldsymbol{Y}}{||\boldsymbol{x}_j^{\perp}||^2} = \frac{\boldsymbol{x}_j^{\perp\top}\boldsymbol{Y}}{\boldsymbol{x}_j^{\perp\top}\boldsymbol{x}_j^{\perp}} = \frac{\boldsymbol{x}_j^{\perp\top}\boldsymbol{Y}}{\boldsymbol{x}_j^{\perp\top}\boldsymbol{x}_j}$$

for

$$\boldsymbol{x}_j^{\perp} = \left\{\boldsymbol{I} - \boldsymbol{X}_{-j}\left(\boldsymbol{X}_{-j}^{\top}\boldsymbol{X}_{-j}\right)^{-1}\boldsymbol{X}_{ij}^{\top}\right\}\boldsymbol{x}_j.$$

This vector is the projection of $\boldsymbol{x}_j$ onto the orthogonal complement of the column space of $\boldsymbol{X}_{-j}$ which is the $\boldsymbol{X}$ matrix with column j removed. Notice in particular in the last equality that the inner product of $\boldsymbol{x}_j^{\perp}$ with itself is the same as its inner product with $\boldsymbol{x}_j$ because $\boldsymbol{x}_j^{\perp}$ is $\boldsymbol{x}_j$ minus something in the column space of $\boldsymbol{X}_{-j}$.

Zhang and Zhang [2014] make one key observation about $\boldsymbol{x}_j^{\perp}$. If a vector $\boldsymbol{z}$ is in the column space of $\boldsymbol{X}$, perpendicular to the column space of $\boldsymbol{X}_{-j}$ and has $\boldsymbol{z}^{\top}\boldsymbol{x}_j = \boldsymbol{z}^{\top}\boldsymbol{z}$ then $\boldsymbol{z} = \boldsymbol{x}_j^{\perp}$. (The orthogonal complement of the column space of $\boldsymbol{X}_{-j}$ within the column space of $\boldsymbol{X}$ is the set of all vectors of the form $a\boldsymbol{x}_j^{\perp}$ for a scalar a. Thus $\boldsymbol{z} = a\boldsymbol{x}_j^{\perp}$. But then $\boldsymbol{z}^{\top}\boldsymbol{z} = a^2\|\boldsymbol{x}_j^{\perp}\|^2$ while $\boldsymbol{z}^{\top}\boldsymbol{x}_j = \boldsymbol{z}^{\top}\boldsymbol{x}_j^{\perp} = a\|\boldsymbol{x}_j^{\perp}\|^2$. The only non-zero solution has $a = 1$.)

More or less the end of what I said in Lecture 5

If $\boldsymbol{X}_{-j}$ has rank less than $p-1$ then the matrix inverse won't exist and will need to be replaced by a generalized inverse (or by selecting a full rank submatrix of $\boldsymbol{X}_{-j}$). But when $p > n$ it will be the case that $\boldsymbol{x}_j$ is in the column space of $\boldsymbol{X}_{-j}$ and the projection being discussed is 0. Thus $\hat{\beta}_{j\text{OLS}}$ is undefined. In these circumstances Zhang and Zhang [2014] suggests replacing $\boldsymbol{x}_j^{\perp}$ with some other vector $\boldsymbol{z}_j$.

Imagine we used

$$\hat{\beta}_{j,\text{alt}} = \frac{\boldsymbol{z}_j^{\top}\boldsymbol{Y}}{\boldsymbol{z}_j^{\top}\boldsymbol{x}_j}.$$

This estimate has mean

$$\frac{\boldsymbol{z}_j^{\top}\boldsymbol{x}\boldsymbol{\beta}}{\boldsymbol{z}_j^{\top}\boldsymbol{x}_j} = \beta_j + \frac{\sum_{k\neq j}\boldsymbol{z}_j^{\top}\boldsymbol{x}_k\beta_k}{\boldsymbol{z}_j^{\top}\boldsymbol{x}_j}$$

so it has bias

$$\frac{\sum_{k\neq j}\boldsymbol{z}_j^{\top}\boldsymbol{x}_k\beta_k}{\boldsymbol{z}_j^{\top}\boldsymbol{x}_j}.$$

It has variance

$$\tau_j^2 = \frac{\boldsymbol{z}_j^\top \boldsymbol{z}_j}{\left(\boldsymbol{z}_j^\top \boldsymbol{x}_j\right)^2}.$$

The idea is to select, for each j, a vector $\boldsymbol{z}_j$ by regression $\boldsymbol{x}_j$ on $\boldsymbol{X}_{-j}$ using LASSO to regularize the fit. If we can estimate the bias and remove it by subtraction in such a way that the residual bias is small compared to σ_j and without inflating the variance then we get confidence intervals using

$$\hat{\beta}_{j,\text{alt}} \pm 2\hat{\sigma}_j$$

where I am now imagining that we have in hand a consistent estimate of σ rather than pretending it is known.

Zhang and Zhang [2014] make several particular suggestions. To remove the bias they suggest using the Scaled LASSO to derive an initial estimate, $\hat{\boldsymbol{\beta}}^{\text{i}}$ for $\boldsymbol{\beta}$ and an estimate $\hat{\sigma}^2$ of σ^2. In the Scaled LASSO we minimize

$$J_\lambda(\boldsymbol{\beta}, \sigma) = \frac{\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2}{2n\sigma} + \frac{\sigma}{2} + \lambda|\boldsymbol{\beta}|_1$$

If we knew the value of $\boldsymbol{\beta}$ then we could carry out the minimization over σ easily and get the estimate

$$\hat{\sigma}^2(\boldsymbol{\beta}, \lambda) = \frac{\|\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda\|^2}{n}.$$

Profiling out this estimate we must minimize

$$J_{\lambda,\text{scaled}}(\boldsymbol{\beta}) = \frac{\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2}{n} + \lambda\|\boldsymbol{\beta}\|_1.$$

This last is called the *square-root* LASSO (because we have replace the Error Sum of Squares in the LASSO penalty by its square root); the square-root LASSO is studied by Belloni et al. [2011]. Remember this suggestion makes λ unitless if the covariates have been scaled to be unitless. As a consequence there is theory suggesting a (fairly) specific value of λ, namely,

$$\lambda_{\text{univ}} = \sqrt{\frac{2\log p}{n}}.$$

(In fact the theory developed in Zhang and Zhang [2014] requires this to be modified by multiplying by some constant larger than 1 and replacing $\log(p)$ by $\log(p)+c$ for some $c > 0$; the paper's simulation studies use the universal value unmodified.)

Next we need to select the $\boldsymbol{z}_j$. To do so we return to the difference between the true and the estimated bias:

$$\left|\frac{\sum_{k\neq j}\boldsymbol{z}_j^\top\boldsymbol{x}_k\left(\hat\beta_k^{\mathrm{i}}-\beta_k\right)}{\boldsymbol{z}_j^\top\boldsymbol{x}_j}\right| \le \tau_j \max_{k\neq j}\left\{\left|\frac{|\boldsymbol{z}_j^\top\boldsymbol{x}_k|}{\boldsymbol{z}_j^\top\boldsymbol{x}_j}\right|\right\}\|\hat{\boldsymbol{\beta}}^i-\boldsymbol{\beta}\|_1.$$

Zhang and Zhang [2014] use the symbol η_j for the middle term.

The most important point here is this: you can do whatever you want with the matrix $\boldsymbol{X}$ as long as you don't do anything that depends on $\boldsymbol{Y}$. So the suggestion is to run square root lasso regressing each $\boldsymbol{x}_j$ on $\boldsymbol{X}_{-j}$. Each such LASSO (called 'node-wise') can be given its own penalty λ_j and then the penalties can be altered so that all the quantities η_j and τ_j are adjusted. Zhang and Zhang [2014] propose a specific algorithm which searches over values of the λ_j to find a solution which makes η_j small without letting τ_j get too large.

Here is a proposition which summarizes the basic strategy.

Proposition 1. *Consider the usual linear regression model with homoscedastic normal errors and fix some integer j and a vector $\boldsymbol{z}_j$, not perpendicular to $\boldsymbol{x}_j$. Suppose $\hat{\boldsymbol{\beta}}^{\mathrm{in}}$ is some initial estimate. Define*

$$\tau_j = \frac{\|\boldsymbol{z}_j\|}{|\boldsymbol{z}_j^\top\boldsymbol{x}_j|}$$

and

$$\eta_j = \frac{\max_{i\neq j}\{|\boldsymbol{z}_j^\top\boldsymbol{x}_i|\}}{\|\boldsymbol{z}_j\|}.$$

Define

$$\hat\beta_j = \frac{\boldsymbol{z}_j^\top\boldsymbol{Y}}{\boldsymbol{z}_j^\top\boldsymbol{x}_j} - \frac{\sum_{i\neq j}\boldsymbol{z}_j^\top\boldsymbol{x}_i\hat\beta_i^{\mathrm{in}}}{\boldsymbol{z}_j^\top\boldsymbol{x}_j}.$$

Then

$$\eta_j\|\hat{\boldsymbol{\beta}}^{\mathrm{in}}-\boldsymbol{\beta}\|_1 = O_P(1)$$

implies that

$$\frac{\hat\beta_j-\beta_j}{\sigma\tau_j} \overset{d}{\rightsquigarrow} N(0,1)$$

If $\hat\sigma$ is consistent in the sense $\hat\sigma/\sigma\to 1$ in probability then

$$\frac{\hat\beta_j-\beta_j}{\hat\sigma\tau_j} \overset{d}{\rightsquigarrow} N(0,1).$$

Proof: Write

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\tau_j} = \frac{\boldsymbol{z}_j^\top \boldsymbol{\epsilon}}{\|\boldsymbol{z}_j\|} - \frac{\sum_{i \neq j} \boldsymbol{z}_j^T \boldsymbol{x}_i (\hat{\beta}_i^{\text{in}} - \beta_i)}{\boldsymbol{z}_j^\top \boldsymbol{x}_j}.$$

The first term on the right hand side has a standard normal distribution. The absolute value of the second term is bounded by

$$\frac{\max_{i \neq j} \left\{ |\boldsymbol{z}_j^T \boldsymbol{x}_i| \right\}}{|\boldsymbol{z}_j^\top \boldsymbol{x}_j|} \times \|\hat{\boldsymbol{\beta}}^{\text{in}} - \boldsymbol{\beta}\|_1 = \eta_j \|\hat{\boldsymbol{\beta}}^{\text{in}} - \boldsymbol{\beta}\|_1.$$

The results follow. •

The crucial inequality at the end can be rewritten in a different way which highlights the overlap between this paper and the others I have mentioned. Define the vector $\boldsymbol{\kappa_j}$ with entries

$$\kappa_{j,i} = \frac{\boldsymbol{z}_j^T \boldsymbol{x}_i}{\boldsymbol{z}_j^\top \boldsymbol{x}_j}$$

and the jth standard basis vector $\boldsymbol{e}_j$ whose entries are all 0 except for entry j which is 1. Notice that $\kappa_{j,j} = 1$ so that

$$\eta_j = \max_i \{|\kappa_{j,i} - e_{j,i}|\} = \|\boldsymbol{\kappa}_j - \boldsymbol{e}_j\|_\infty.$$

The inequality is therefore

$$\left| \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}\tau_j} - \frac{\boldsymbol{z}_j^\top \boldsymbol{\epsilon}}{\|\boldsymbol{z}_j\|} \right| \leq \|\boldsymbol{\kappa}_j - \boldsymbol{e}_j\|_\infty \|\hat{\boldsymbol{\beta}}^{\text{in}} - \boldsymbol{\beta}\|_1.$$

Zhang and Zhang [2014] described their method as starting with a linear estimator and correcting its bias using the LASSO to get and inital estimate. But we can write

$$\begin{aligned}
\hat{\beta}_j &= \frac{\boldsymbol{z}_j^\top \boldsymbol{Y}}{\boldsymbol{z}_j^\top \boldsymbol{x}_j} - \frac{\sum_{i \neq j} \boldsymbol{z}_j^T \boldsymbol{x}_i \hat{\beta}_i^{\text{in}}}{\boldsymbol{z}_j^\top \boldsymbol{x}_j} \\
&= \frac{\boldsymbol{z}_j^\top \boldsymbol{Y}}{\boldsymbol{z}_j^\top \boldsymbol{x}_j} - \frac{\sum_{i} \boldsymbol{z}_j^T \boldsymbol{x}_i \hat{\beta}_i^{\text{in}}}{\boldsymbol{z}_j^\top \boldsymbol{x}_j} + \hat{\beta}_j^{\text{in}} \\
&= \hat{\beta}_j^{\text{in}} + \frac{\boldsymbol{z}_j^\top \left(\boldsymbol{Y} - \boldsymbol{X} \hat{\boldsymbol{\beta}}^{\text{in}} \right)}{\boldsymbol{z}_j^\top \boldsymbol{x}_j}
\end{aligned}$$

This formulation looks like the initial estimator corrected by the extent to which the residual is not perpendicular to $\boldsymbol{z}_j$. It is this form which Javanmard and Montanari [2014b] and van de Geer et al. [2014] use.

4.2. **van de Geer et al. [2014] and Javanmard and Montanari [2014b].** I am going to write out the KKT conditions in a way which is convenient for the theoretical analysis of the suggestions made by van de Geer et al. [2014] and Javanmard and Montanari [2014b]. The vector $\hat{\boldsymbol{\beta}}_\lambda$ solves the equation

$$\boldsymbol{X}^\top \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{U} + \lambda\boldsymbol{\kappa} = \lambda\boldsymbol{\kappa} - \boldsymbol{X}^\top\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda\right) = 0$$

for a vector $\boldsymbol{\kappa}$ with

$$\|\boldsymbol{\kappa}\|_\infty \le 1$$

and for any j with $\beta_j \neq 0$

$$\kappa_j = \text{sign}\beta_j.$$

If we could multiply through by $\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}$ we would see that $\hat{\boldsymbol{\beta}}_\lambda$ is the ordinary least squares estimate minus $\lambda\left(\boldsymbol{X}^\top\boldsymbol{X}\right)^{-1}\boldsymbol{\kappa}$. We could then correct the bias of the LASSO estimate by adding back in this last term; that is just another way of saying we could use OLS! Both van de Geer et al. [2014] and Javanmard and Montanari [2014b] suggest that when you can't find an exact inverse of $\boldsymbol{X}^\top\boldsymbol{X}$ you find a matrix $\boldsymbol{M}$ which is the best possible substitute. Then you use the estimate

$$\hat{\boldsymbol{b}} \equiv \hat{\boldsymbol{\beta}}_\lambda + \lambda\boldsymbol{M}\boldsymbol{\kappa} = \hat{\boldsymbol{\beta}}_\lambda + \boldsymbol{M}\boldsymbol{X}^\top\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda\right).$$

This has the same structure as Zhang and Zhang's estimator.

Both papers then observe (using the notation $\boldsymbol{R} = \boldsymbol{X}^\top\boldsymbol{X}$)

$$\begin{aligned}\hat{\boldsymbol{b}} &= \hat{\boldsymbol{\beta}}_\lambda + \boldsymbol{M}\boldsymbol{X}^\top\left(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda\right)\\ &= \boldsymbol{M}\boldsymbol{X}^\top(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \boldsymbol{M}\boldsymbol{R}\boldsymbol{\beta} + (\boldsymbol{I} - \boldsymbol{M}\boldsymbol{R})\hat{\boldsymbol{\beta}}_\lambda\\ &= \boldsymbol{\beta} + \boldsymbol{M}\boldsymbol{X}^\top\boldsymbol{\epsilon} + (\boldsymbol{I} - \boldsymbol{M}\boldsymbol{R})\left(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\right)\end{aligned}$$

Notice the re-appearance of the sum of the parameter being estimated, a linear function of the errors, and a term which depends on the error in the LASSO estimate. As did Zhang and Zhang [2014] we observe

$$\left|(\boldsymbol{I} - \boldsymbol{M}\boldsymbol{R})\left(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\right)\right|_\infty \le \|\boldsymbol{I} - \boldsymbol{M}\boldsymbol{R}\|_\infty\|\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\|_1.$$

The only difference here is that Zhang and Zhang describe the inequality one component at a time. Both papers now seek to control the two terms on the right hand side; the two terms are the same L_1 error that Zhang and Zhang [2014] study and the L_∞ norm which is actually just $\max_j \eta_j$.

Javanmard and Montanari [2014b] consider directly the selection of a matrix $\boldsymbol{M}$. They define the coherence parameter $\mu_*(\boldsymbol{X}, \boldsymbol{M})$ by

$$\mu_*(\boldsymbol{X}, \boldsymbol{M}) = \|\boldsymbol{I} - \boldsymbol{M}\boldsymbol{R}\|_\infty = \max_i \|\boldsymbol{e}_i - \boldsymbol{R}_i\|_\infty$$

where ${}_i$ is column i of $\boldsymbol{M}$. The variance of $\boldsymbol{M}\boldsymbol{X}^\top\boldsymbol{\epsilon}$ is $\boldsymbol{M}\boldsymbol{R}\boldsymbol{M}$; the ith element on the diagonal of this matrix is just ${}_i^\top\boldsymbol{R}_i$. So for each i Javanmard and Montanari [2014b] describe an algorithm to minimize ${}_i^\top\boldsymbol{R}_i$ subject to

$$\|\boldsymbol{e}_i - \boldsymbol{R}_i\|_\infty \leq \mu$$

for some preselected μ. If the condition is achievable for each i then you get

$$\|\boldsymbol{I} - \boldsymbol{M}\boldsymbol{R}\|_\infty \leq \mu.$$

This is precisely the bound needed on the η_j for Zhang and Zhang [2014].

4.3. **Theoretical Results.** Each of these papers contains a theorem which has the following flavour. I will state the theorem without the technical details being properly spelled out to give the flavour.

Theorem 6. *Suppose the design matrix $\boldsymbol{X}$ satisfies a compatibility condition with compatibility constant ϕ^2. Then there is a constant C which depends on the detailed properties of $\boldsymbol{X}$, the sparsity $s_0 = |\{j : \beta_j \neq 0\}|$, and the penalty parameter λ such that*

$$P\left(\left\|(\boldsymbol{I} - \boldsymbol{M}\boldsymbol{R})\left(\hat{\boldsymbol{\beta}}_\lambda - \boldsymbol{\beta}\right)\right\|_\infty > C\right) \leq \epsilon$$

for some specific small ϵ.

The compatibility condition is actually rather like a minimum eigenvalue condition in flavour but much weaker. The smallest eigenvalue of a matrix $\boldsymbol{Q}$ is

$$\lambda_{\min}(\boldsymbol{Q}) = \inf\left\{\frac{\boldsymbol{x}^\top\boldsymbol{Q}\boldsymbol{x}}{\boldsymbol{x}^\top\boldsymbol{x}} : \boldsymbol{x} \neq \boldsymbol{0}\right\}.$$

The compatibility constant replaces the L_2 norm in the denominator with an L_1 norm squared computed using only the entries in S_0 and then takes the infimum over a much smaller set of $\boldsymbol{\beta}$:

$$\phi_0^2 = \inf_{\boldsymbol{\beta}}\left\{\frac{\boldsymbol{\beta}^\top\boldsymbol{R}\boldsymbol{\beta}}{\|\boldsymbol{\beta}_{S_0}\|_1^2/|S_0|} : \|\boldsymbol{\beta}_{S_0^c}\|_1 \leq 3\|\boldsymbol{\beta}_{S_0}\|_1\right\}.$$

I remark that

$$\|\boldsymbol{\beta}_{S_0}\|_2^2 \geq \|\boldsymbol{\beta}_{S_0}\|_1^2/|S_0|$$

because the difference is $|S_0|$ times the variance of the entries in β_{S_0}. The ratio of these two quantities can be as big as $|S_0|$. Since the compatibility constant appears in the denominator of the constant C the use of the L_1 version in the denominator can make C smaller by a factor of $\sqrt{|S_0|}$.

In each paper there is then a section of results for random designs. These are the models in which we have an iid sample $(Y_i, X_{i1}, \ldots, X_{ip})$. Typically there is a lower bound imposed on the smallest eigenvalue of the variance covariance matrix of the Xs. van de Geer et al. [2014] require sparsity in the inverse of this variance covariance matrix; they require that the largest number of non-zero entries in any row of the inverse be $o(n/\log(p))$. They also require (as do Javanmard and Montanari [2014b]) $|S_0| = o(\sqrt{n}/\log(p))$. When $n = 71$ and $p = 4088$ we have

$$\sqrt{n}/\log(p) = 2.33$$

Now we see one of the difficulties with this theory. The condition in the theory says small compared to 2.33. But that would seem to mean we should not expect much without the number of non-zero entries in $\boldsymbol{\beta}$ being 1 or 2 at the largest. The number of non-zero entries in a row in the inverse variance covariance matrix is supposed to be small compared to $71/\log(4088) \approx 19$. So most entries in this inverse are 0, meaning that most pairs of variables are conditionally independent given all the other variables. That feels like a strong assumption in the context of our example.

5. Post Selection Inference (PoSI)

Lockhart et al. [2014] and Tibshirani et al. [2016], have different inferential targets than do Zhang and Zhang [2014], van de Geer et al. [2014], and Javanmard and Montanari [2014b]. The latter 3 are interested in all p of the coefficients in a model in which Y is regressed on all p of the X variables. They remove the sparseness when they adjust the LASSO estimate. They try to give conditions under which you might trust all p of the resulting confidence intervals simultaneously. But assumptions on the coefficients are necessary. Regardless of which sparse model your LASSO fit selects you get, in principal, confidence limits for all

p parameters and no sparseness. Of course, it is not possible to prove that things are zero – only to assess evidence against the assertion that they are zero.

Lockhart et al. [2014] and Tibshirani et al. [2016] take the view that the parameters you will be interested in depend on the model you select. We have seen that in designs with highly correlated covariates LASSO (and other variable selection methods) have a substantial probability of getting the model wrong. For this reason these two papers and related work focus on the best linear approximation to the mean of $\boldsymbol{Y}$ in terms of a linear combination of the columns of $\boldsymbol{X}$. Then they analyze estimates conditional on which model is selected. Their analysis focuses on some specific model selection strategies for which the event that a specific model is selected has a specific structure — a union of polyhedra.

5.1. **PoSI: Berk et al. [2013].** Berk et al. [2013] takes an unconditional approach. They too imagine that you will run some model selection scheme and then be interested in the regression of $\boldsymbol{Y}$ on the columns of the design corresponding to the selected model. They too imagine that you might not have the model right and suggest evaluating how well you do at making inferences about the best possible approximation to the mean vector of $\boldsymbol{Y}$ by a linear combination of the selected columns of $\boldsymbol{X}$. The really nice thing about this work is that they handle all model selection schemes simultaneously – even ones designed to pick out the most significant variables.

The key idea is an extension of common strategies in the multiple comparisons literature. So I will start by reviewing Tukey and Scheffé confidence intervals.

Consider first data of the form Y_{ij} where i from 1 to p labels p populations and j running from 1 to n_i labels the members of a sample from population i. The classic one way ANOVA model is that these are p independent samples from normally distributed populations with common standard deviation σ and means μ_i. Of course this is a regression model. (Notationally this doesn't match the rest of these notes; the μ_i are the regression parameters and I ought to use β_i for their names. But that would be ahistoric.)

Analyses used to proceed something like this. First run an F test for the null hypothesis that all the means are equal. If that hypothesis is rejected then give confidence intervals for parameters of interest. In the following discussion I am going to ignore that pre-test step and pretend you would do the subsequent analyses in any case.

Typically interest focuses not on the absolute levels of the groups (that is, not on the μ_i) but on comparisons between groups. Tukey's method supposes that you are only interested in all possible pairwise comparisons: $\theta_{ij} = \mu_i - \mu_j$; a total of $p(p-1)/2$ parameters built out of the p values of μ_i. Let $\hat{\sigma}$ be the usual estimate of σ – the root mean square error. The vector T with entries $T_{ik}, 1 \le i < k \le p$ given by

$$T_{ik} = \frac{(\bar{Y}_i - \bar{Y}_k) - (\mu_i - \mu_k)}{\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_k}}}$$

is a pivot. Its joint distribution, for homoscedastic normal populations, does not depend on the parameters. So there is a constant $K_{\mathrm{T},\alpha}$ such that

$$P_{\boldsymbol{\mu},\sigma}\left(\forall i, k |T_{ik}| \le K_{\mathrm{T},\alpha}\right) = 1 - \alpha.$$

If we use this constant then the family of intervals

$$\bar{Y}_i - \bar{Y}_k \pm K_{\mathrm{T},\alpha}\hat{\sigma}\sqrt{\frac{1}{n_i} + \frac{1}{n_k}}$$

has simultaneous coverage probability $1 - \alpha$ for all parameter values.

If we knew in advance which populations i and k we would like to compare we would use the same confidence interval but with the constant $K_{\mathrm{T},\alpha}$ replace by the smaller value from Student's t distribution with the appropriate degrees of freedom, $\nu = \sum_i (n_i - 1)$. Using the larger multiplier gives us the luxury of being able to look at the means and pick out the differences which most interest us and still give valid coverage probability statements for the 'interesting' comparison.

Scheffé intervals permit more elaborate data snooping of this kind. In the same situation where Tukey's intervals are proposed Scheffé imagines you might make other comparisons. You might be interested in say, $(\mu_2 + \mu_3)/2 - \mu_1$, a different contrast of the entries in the mean vector. Suppose $\boldsymbol{a}$ is a vector of length p with entries which sum to 0; then $\boldsymbol{a}$ is a *contrast* vector. The family of all statistics $T_{\boldsymbol{a}}$ indexed by $\boldsymbol{a} \neq \mathbf{0}$ given by

$$T_{\boldsymbol{a}} = \frac{\boldsymbol{a}^\top(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})}{\hat{\sigma}\sqrt{\sum_i \frac{a_i^2}{n_i}}}$$

is again pivotal. Define

$$T_{\mathrm{S}} = \sup_{\boldsymbol{a}:\boldsymbol{a}\neq 0} \{|T_{\boldsymbol{a}}|\}.$$

Again T_{S} is pivotal. Moreover we can actually compute the supremum since $T_{\boldsymbol{a}}^2$ is a ratio two quadratic forms in $\boldsymbol{a}$. we find

$$T_{\mathrm{S}} = \frac{(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^\top (\boldsymbol{X}^\top \boldsymbol{X})^{-1} (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})}{\hat{\sigma}^2} = (p-1)F$$

where $\boldsymbol{X}$ is the usual design matrix for one way ANOVA so that $\boldsymbol{X}^\top \boldsymbol{X}$ is diagonal with $n_1, \ldots, n_p$ down the diagonal and F is the usual central F pivot which has degrees of freedom $p-1$ and $\nu = \sum_i (n_1 - 1)$. Thus we can use the intervals

$$\boldsymbol{a}^\top (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}) \pm \sqrt{p-1} F_{p-1,\nu,\alpha} \hat{\sigma} \sqrt{\boldsymbol{a}^\top (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{a}}$$

as simultaneous confidence intervals and pick out any contrast we find interesting without letting the error rate go above α.

The idea in Berk et al. [2013] is to use multipliers $K_{\mathrm{PoSI},\alpha}$ which are between the Tukey and Scheffé multipliers and give you control over all intervals for individual slopes in the regression of $\boldsymbol{Y}$ on any subset of the predictors in $\boldsymbol{X}$. Since they give simultaneous coverage over all models they give coverage guarantees after selection; this guarantee leads to the name *Post Selection Inference* or PoSI.

Here is some generic notation. We are regressing $\boldsymbol{Y}$ on $\boldsymbol{X}$ as usual and assuming homoscedastic normal errors. We suppose that $\boldsymbol{Y}$ has mean vector μ and error vector $\boldsymbol{\epsilon}$ with iid $N(0, \sigma^2)$ components. We treat $\boldsymbol{X}$ as fixed; if $\boldsymbol{X}$ is actually random we are conditioning on $\boldsymbol{X}$ and assuming

$$\mu = \mathrm{E}(\boldsymbol{Y} | \boldsymbol{X})$$

and that the conditional distribution, given $\boldsymbol{X}$, of $\boldsymbol{\epsilon}$ defined by

$$\boldsymbol{\epsilon} = \boldsymbol{y} - \mathrm{E}(\boldsymbol{Y} | \boldsymbol{X})$$

is iid $N(0, \sigma^2)$.

For a given model $A \subset \{1, \ldots, p\}$ we define the slope $\beta_{j \cdot A}$ for $j \in A$ to be the jth element in

$$\left(\boldsymbol{X}_A^\top \boldsymbol{X}_A\right)^{-1} \boldsymbol{X}_A^\top \boldsymbol{\mu}.$$

This is exactly analogous to the suggestion of Tibshirani et al. [2016] although the latter contemplated vectors ν other than those which are rows of $\left(\boldsymbol{X}_A^\top \boldsymbol{X}_A\right)^{-1} \boldsymbol{X}_A^\top$. I strongly encourage the reading of the introductory material in Berk et al. [2013] where a pretty persuasive case is made in support of the appropriateness of these parameters.

Let $\boldsymbol{a}_{j\cdot A}$ be the jth row of $\left(\boldsymbol{X}_A^\top \boldsymbol{X}_A\right)^{-1} \boldsymbol{X}_A^\top$. Consider the natural estimate of

$$\beta_{j\cdot A} = \boldsymbol{a}_{j\cdot A}\boldsymbol{\mu},$$

namely,

$$\hat{\beta}_{j\cdot A} = \boldsymbol{a}_{j\cdot A}\boldsymbol{Y}.$$

Then the family of statistics $T_{j\cdot A}$ (for all $A \subset \{1,\ldots,p\}$ and all $j \in A$ with A not empty) given by

$$T_{j\cdot A} = \frac{\boldsymbol{a}_{j\cdot A}(\boldsymbol{Y}-\boldsymbol{\mu})}{\hat{\sigma}\sqrt{\boldsymbol{a}_{j\cdot A}^\top(\boldsymbol{X}_A^T\boldsymbol{X}_A)^{-1}\boldsymbol{a}_{j\cdot A}}}$$

is pivotal. Define $K_{\mathrm{PoSI},\alpha}$ by

$$P\left(\max_{A,j\in A}|T_{j\cdot A}| \le K_{\mathrm{PoSI},\alpha}\right) = 1-\alpha.$$

Then the family of intervals

$$\boldsymbol{a}_{j\cdot A}\boldsymbol{Y} \pm K_{\mathrm{PoSI},\alpha}\hat{\sigma}\sqrt{\boldsymbol{a}_{j\cdot A}^\top(\boldsymbol{X}_A^T\boldsymbol{X}_A)^{-1}\boldsymbol{a}_{j\cdot A}}$$

has simultaneous coverage probability $1-\alpha$.

The PoSI suggestion is simply to run whatever model selection scheme you like to get $\hat{A}$, a selected model. Then use, for $j \in \hat{A}$, the intervals

$$\mathrm{CI}_{j\cdot\hat{A}} \equiv \boldsymbol{a}_{j\cdot\hat{A}}^\top\boldsymbol{Y} \pm K_{\mathrm{PoSI},\alpha}\hat{\sigma}\sqrt{\boldsymbol{a}_{j\cdot\hat{A}}^\top(\boldsymbol{X}_{\hat{A}}^T\boldsymbol{X}_{\hat{A}})^{-1}\boldsymbol{a}_{j\cdot\hat{A}}}$$

Since each of these intervals is one of the intervals used above and since there is only chance α that any of the larger set of intervals misses its target we see

$$P\left(\forall j \in \hat{A} : \boldsymbol{a}_{j\cdot A}^\top\boldsymbol{\mu} \in \mathrm{CI}_{j\cdot\hat{A}}\right) \ge 1-\alpha.$$

Here are some comments on these ideas:

- The constant $K_{\mathrm{PoSI},\alpha}$ depends on the design matrix $\boldsymbol{X}$. It is, of course, less than Scheffé's constant. An important question is: how much less? Berk et al. [2013] show that for an orthogonal design the constant is on the order of $\sqrt{\log p}$. On the other hand there are sequences of designs where the constant grows at the rate $\sqrt{p}$; this is the same rate as the growth of Scheffé's constant. I want to emphasize that Scheffé's ideas are rarely used. The confidence intervals are larger than people like – mathematically

best possible for the kind of guarantee given, but a case where the truth is unappealing.

- The constants can be computed for p up to about 20. So this is not really high dimensional inference. The event whose probability is to be calculated is a polyhedron in the space of values of $\boldsymbol{X}^\top \boldsymbol{Y}$. Good algorithms for larger p would be welcome I believe.
- You need $n > p$ for all these estimates to exist. This is another sign that PoSI is not yet high dimensional.
- You can also study model selection procedures which do not select from all submodels, A, but from a restricted list of sub-models. For instance, in polynomial regression you might insist that each A have the form $\{0, \ldots, d\}$ corresponding to a polynomial of degree d and then limit d to some maximal degree. This kind of restriction can make the constant smaller.
- Since $n > p$ one potential estimate of σ is the RMSE of the full model. But this requires the full model to be right: $\boldsymbol{\mu}$ must be in the column space of $\boldsymbol{X}$. You might look for exact replicates and use the Pure Error Sum of Squares. The paper suggests you might get $\hat{\sigma}$ from a previous experiment; I find this one a bit hard to believe. You might deliberately overestimate σ; in the extreme the sample standard deviation of the Y's does this. But now the coverage probabilities are not exact at all; you need an estimate $\hat{\sigma}$ for which $\nu\hat{\sigma}^2/\sigma^2$ has a χ^2_ν distribution and is independent of all the estimates in all the models.

6. Limits to inference, Leeb and Pötscher [2005]

Leeb and Pötscher [2005] makes some important comments on the problem of inference in high dimensions. All statistical inference, even "model free" methods, relies on model assumptions. The assumption of iid sampling, for instance, can be a strong one.

I think the simple example given in this paper really captures the difficulties pretty well. Consider the model

$$Y_i = \alpha U_i + \beta V_i + \epsilon_i$$

with the usual iid $N(0,1)$ errors; I do the case σ known. We will be interested in doing model selection between two models: $A = \{1\}$, where only U is active

(and $\beta = 0$), and $A = \{1, 2\}$, the full model. Suppose that we have a sequence of designs in which $\boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{V}^\top \boldsymbol{V} =$ and $\boldsymbol{U}^\top \boldsymbol{V} = \rho$ with $|\rho| < 1$. If you knew that β were 0 you would like to estimate α using the restricted least squares estimate

$$\hat{\alpha}_R = \frac{\boldsymbol{U}^\top \boldsymbol{Y}}{\boldsymbol{U}^\top \boldsymbol{U}}.$$

If you knew $\beta \neq 0$ you would prefer the full least squares estimate

$$\hat{\alpha}_F = \frac{(\boldsymbol{U} - \rho \boldsymbol{V})^\top \boldsymbol{Y}}{1 - \rho^2}.$$

While $\hat{\alpha}_F$ is unbiased the mean of $\hat{\alpha}_R$ is

$$\mathrm{E}(\hat{\alpha}_R) = \alpha + \rho\beta.$$

On the other hand

$$\mathrm{Var}(\hat{\alpha}_R) = 1 < \mathrm{Var}(\hat{\alpha}_F) = \frac{1}{1 - \rho^2}.$$

Superficially we can have the best of both worlds. Here is a procedure with the oracle property. Let $\hat{\beta}_F$ be the ordinary least squares estimate of β for the full model given by

$$\hat{\beta}_F = \frac{(\boldsymbol{V} - \rho \boldsymbol{U})^\top \boldsymbol{Y}}{1 - \rho^2}.$$

This estimate has variance $1/(1 - \rho^2)$. Now define the *pre-test* estimator of α by

$$\hat{\alpha} = \begin{cases} \hat{\alpha}_F & \left| \frac{\hat{\beta}^{\mathrm{OLS}}}{\hat{\sigma}_{\hat{\beta}^{\mathrm{OLS}}}} \right| > c_n \\ \\ \hat{\alpha}_R & \left| \frac{\hat{\beta}^{\mathrm{OLS}}}{\hat{\sigma}_{\hat{\beta}^{\mathrm{OLS}}}} \right| \leq c_n. \end{cases}$$

Thus we are considering the model selection procedure

$$\hat{A} = \begin{cases} \{1, 2\} & \left| \frac{\hat{\beta}^{\mathrm{OLS}}}{\hat{\sigma}_{\hat{\beta}^{\mathrm{OLS}}}} \right| > c_n \\ \\ \{1\} & \left| \frac{\hat{\beta}^{\mathrm{OLS}}}{\hat{\sigma}_{\hat{\beta}^{\mathrm{OLS}}}} \right| \leq c_n. \end{cases}$$

Then for every α and β we have

$$\lim_{n \to \infty} P_{\alpha\beta}(\hat{A} = A_0) = 1$$

where A_0 is the true model ($\{1\}$ if $\beta = 0$ and $\{1, 2\}$ otherwise). In this procedure the constants c_n are increasing to ∞ slowly enough that $c_n/\sqrt{n} \to 0$.

This oracle property permits people to argue that they can get procedures which work just as well as if the true model were known. Consider the following confidence interval procedure for α. If $\hat{A} = \{1\}$ use the interval

$$CI_1 = \boldsymbol{U}^\top \boldsymbol{Y} \pm z_\alpha \equiv \hat{\alpha}_{(0)}^\top \boldsymbol{Y} \pm z_\alpha.$$

If $\hat{A} = \{1, 2\}$ use the ordinary least squares estimator

$$\hat{\alpha}^{\text{OLS}} = \frac{(\boldsymbol{U} - \rho \boldsymbol{V})^\top \boldsymbol{Y}}{1 - \rho^2}.$$

and compute

$$CI_{1,2} = \hat{\alpha}^{\text{OLS}} \pm z_\alpha \frac{1}{\sqrt{1 - \rho^2}}$$

The latter interval is wider, of course. The idea is that you get the benefit, when $\beta = 0$, of narrower confidence intervals and you have the apparent confidence guarantee

$$\lim_{n \to \infty} P_{\alpha\beta}(\alpha \in \text{CI}_n) = 1 - \alpha$$

where CI_n means use the interval corresponding to $\hat{A}$.

But there is a catch and the point of this paper is that there is always a catch. The assertions above are, in formal notation

$$\forall \alpha, \beta \, \forall \epsilon > 0 \, \exists N \, \forall n : n \geq N \Rightarrow |P_{\alpha\beta}(\alpha \in CI_n) - (1 - \alpha)| \leq \epsilon.$$

The historic definition of the confidence level of an interval is the minimum, over the parameter space, of the coverage probability. So the theorem you want is

$$\forall \epsilon > 0 \, \exists N \, \forall \alpha, \beta, n : n \geq N \Rightarrow |P_{\alpha\beta}(\alpha \in CI_n) - (1 - \alpha)| \leq \epsilon.$$

This one is not true. And the fact that it is not true does not depend on the particular nature of the oracle procedure at hand.

Here is what I see as the guts of the problem. Define

$$H_n(\alpha, \beta) = P_{\alpha\beta}(\hat{A} = \{1\}).$$

The oracle conclusion arises from the fact that

$$\lim_{n \to \infty} H_n(\alpha, \beta) = H_\infty(\alpha, \beta) = 1(\beta = 0).$$

But now H_n is an analytic function of its arguments converging pointwise to its discontinuous limit H_∞. That convergence cannot be uniform because the continuous image of a connected set is connected. For all large n there must be points α_n, β_n where $H_n(\alpha_n, \beta_n) = 1/2$; that is, there are parameter values where you have a 50% chance of picking the wrong model. (In fact you can take $\alpha_n = 0$ because the intervals are equivariant in α.) For these points α_n, β_n you use CI_1 half the time and $CI_{1,2}$ about half the time. Now I want to prove that you must get bad coverage probabilities for some value of α, β. Consider now a sequence

7. Some simulations

The design matrix in the riboflavin data set is problematic for the theory we have seen. I tried the following simple little trials. I generated N(0,1) errors and took $Y_i = \beta x_{1278,i} + \epsilon_i$. I scaled the design matrix to keep $\boldsymbol{X}^\top \boldsymbol{X}$ a correlation matrix. I tried β running from $\sqrt{2\log(p)}$ (which is about 4) to 9 with $\sigma = 1$. Using the multi-split method in `hdi` with all the default settings and taking the family wise error rate to be $\alpha = 0.05$ I get, for 20 Monte Carlo data sets, the selected variables shown in Table 7.

TABLE 2. Variables selected when the `multi-split` method is run with the default settings using the centred and standardize design matrix to generate data with only coefficient #1278 non-zero. That coefficient, β, runs from 5 to 9. For each of 20 Monte Carlo runs the table shows for that β the variables selected at the 5% level.

Run	$\beta = 5$	$\beta = 6$	$\beta = 7$	$\beta = 8$	$\beta = 9$
1	None	None	1278 1423	1278 1423	1278
2	None	None	1278	1278	1278
3	None	None	None	1278	1278
4	None	None	1278 1279 1312	1278	1278
5	None	1423	1423	1423	1423
6	None	None	None	None	1278 1423
7	None	None	None	1278	1278 1279
8	None	1278	1278	1278	1278
9	None	None	None	None	1279
10	None	None	1423	1278 1423	1278
11	1278 1279	1278 1279	1278 1279	1278 1279	1278 1279
12	None	None	1279 1290	1279 1290	1279 1290
13	None	None	1278	1278	1278 1279
14	None	1278	1278	1278	1278
15	1278	1278	1278	1278	1278
16	1278	1278	1278	1278	1278
17	1278 1279	1278 1279	1278 1279	1278 1279	1278 1279
18	None	None	None	1278 1279	1278 1279
19	1310	1278 1310	1278	1278	1278
20	None	1278	1278	1278	1278

REFERENCES

Anestis Antoniadis. Comments on: ℓ_1-penalization for mixture regression models. *TEST*, 19(2):257–258, 2010. ISSN 1863-8260. doi: 10.1007/s11749-010-0198-y. URL http://dx.doi.org/10.1007/s11749-010-0198-y.

A. Belloni, V. Chernozhukov, and L. Wang. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, 98(4):791, 2011. doi: 10.1093/biomet/asr043. URL +http://dx.doi.org/10.1093/biomet/asr043.

Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao. Valid post-selection inference. *Ann. Statist.*, 41(2):802–837, 04 2013. doi: 10.1214/12-AOS1077. URL http://dx.doi.org/10.1214/12-AOS1077.

Peter Bühlmann, Markus Kalisch, and Lukas Meier. High-dimensional statistics with a view toward applications in biology. *Annual Review of Statistics and Its Application*, 1(1):255–278, 2014. doi: 10.1146/annurev-statistics-022513-115545. URL /brokenurl#http://dx.doi.org/10.1146/annurev-statistics-022513-115545.

W. Fithian, D. Sun, and J. Taylor. Optimal Inference After Model Selection. *ArXiv e-prints*, October 2014.

W. Fithian, J. Taylor, R. Tibshirani, and R. Tibshirani. Selective Sequential Model Selection. *ArXiv e-prints*, December 2015.

A. Javanmard and A. Montanari. Hypothesis Testing in High-Dimensional Regression under the Gaussian Random Design Model: Asymptotic Theory. *ArXiv e-prints*, January 2013.

A. Javanmard and A. Montanari. Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. *IEEE Transactions on Information Theory*, 60(10):6522–6554, Oct 2014a. ISSN 0018-9448. doi: 10.1109/TIT.2014.2343629.

A. Javanmard and A. Montanari. De-biasing the Lasso: Optimal Sample Size for Gaussian Designs. *ArXiv e-prints*, August 2015.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15: 2869–2909, 2014b. URL http://jmlr.org/papers/v15/javanmard14a.html.

Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor. Exact post-selection inference, with application to the lasso. *Ann. Statist.*, 44(3):907–927, 06 2016. doi: 10.1214/15-AOS1371. URL http://dx.doi.org/10.1214/15-AOS1371.

Hannes Leeb and Benedikt M. Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21(1):21–59, 002 2005. doi: 10.1017/S0266466605050036. URL https://www.cambridge.org/core/article/div-class-title-model-selection-and-inference-facts-and-fiction-div/EF3C7D79D5AFC4C6325345A3C8E26296.

Hannes Leeb and Benedikt M. Ptscher. Can one estimate the conditional distribution of post-model-selection estimators? *Ann. Statist.*, 34(5):2554–2591, 10 2006. doi: 10.1214/009053606000000821. URL http://dx.doi.org/10.1214/009053606000000821.

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.

Tingni Sun and Cun-Hui Zhang. Comments on: ? 1-penalization for mixture regression models. *TEST*, 19(2):270–275, 2010. ISSN 1863-8260. doi: 10.1007/s11749-010-0201-7. URL http://dx.doi.org/10.1007/s11749-010-0201-7.

Tingni Sun and Cun-Hui Zhang. Scaled sparse linear regression. *Biometrika*, 99 (4):879, 2012. doi: 10.1093/biomet/ass043. URL +http://dx.doi.org/10.1093/biomet/ass043.

X. Tian and J. E. Taylor. Selective inference with a randomized response. *ArXiv e-prints*, July 2015.

Ryan J. Tibshirani. The lasso problem and uniqueness. *Electron. J. Statist.*, 7: 1456–1490, 2013. doi: 10.1214/13-EJS815. URL http://dx.doi.org/10.1214/13-EJS815.

Ryan J. Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620, 2016. doi: 10.1080/01621459.2015.1108848. URL http://dx.doi.org/10.1080/01621459.2015.1108848.

Sara van de Geer, Peter Bhlmann, Yaacov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, 42(3):1166–1202, 06 2014. doi: 10.1214/14-AOS1221. URL http://dx.doi.org/10.1214/14-AOS1221.

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.

Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014. ISSN 1467-9868. doi: 10.1111/rssb.12026. URL http://dx.doi.org/10.1111/rssb.12026.