INFERENCE IN HIGH-DIMENSIONAL LINEAR MODELS COURSE NOTES

RICHARD LOCKHART

Contents

Version of 29 January 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	Conditional inference given selection; Tibshirani et al (2016)
Feb 13	POSI: conservative inference schemes; Berk et al
Feb 20	Limits ; Leeb and Pötscher
Feb 27	Debiasing: Bühlmann, van de Geer, Meinshausen
Mar 6	Javanmard, Montanari
Mar 13	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.

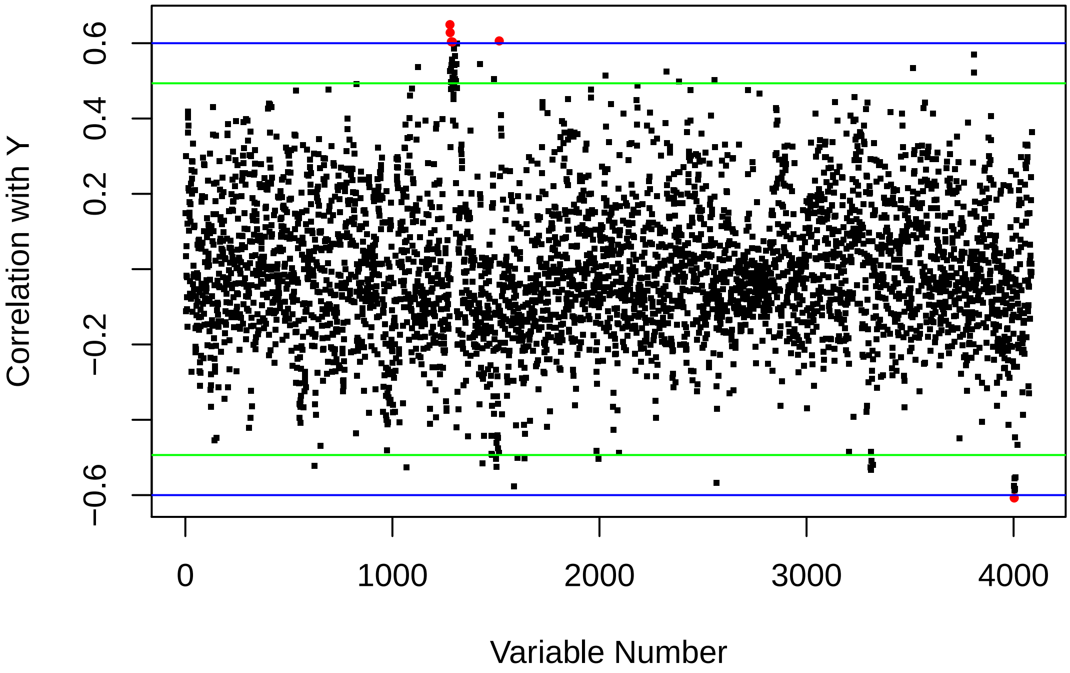
I am going to pretend that we have a sample of n independent and identically distributed vectors $(Y_i, X_{i1}, \ldots, X_{ip})$. I will start with the basic question of whether or not there is any relationship between any of the genes and riboflavin. We will need some notation.

As usual we will stack the 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 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 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 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 rationale 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. Here 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 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. Here are 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 $P(p <\le u) \le u$ for all $u \in [0,1]$ and the inequality is strict for some u. If $p_1, \ldots, p_m$ are any m exact P-values (with any joint law whatsoever) then

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

so

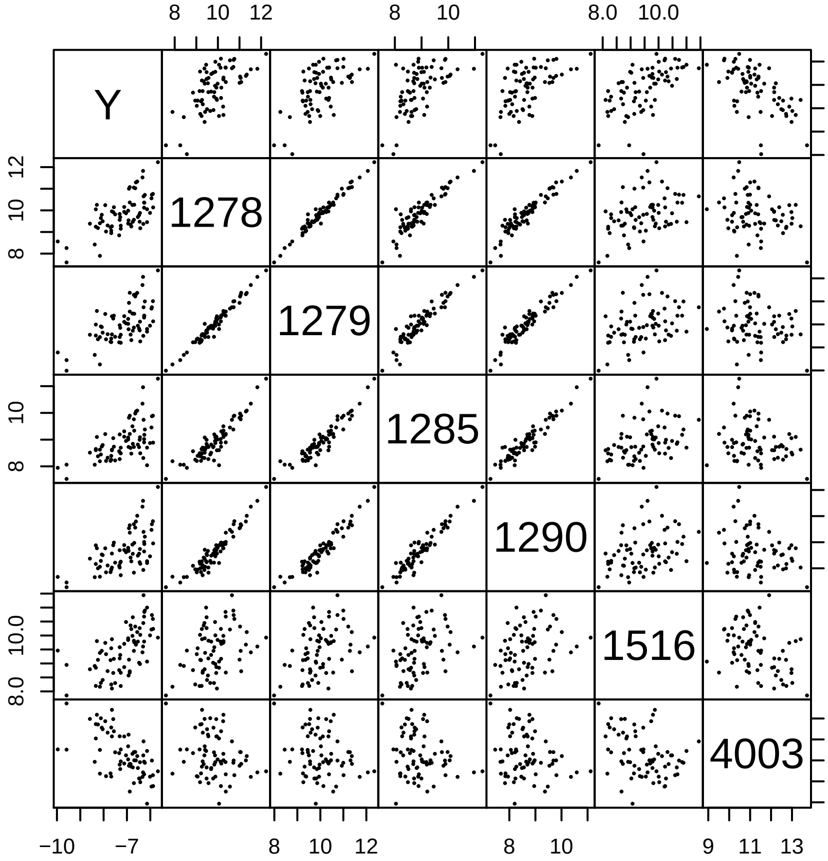
$$p_{\text{Bon}} = m \min_{1\le j\le m}\{p_j\}$$

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

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

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.

- There is no reasonable way the response is independent of the predictors.
- 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

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

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 Y on both $\boldsymbol{U}$ and 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 V as column 1279. Take β to be the slope of Y regressed on variable 1278 (ignoring the selection problems these lectures are actually about) and simulate new vectors Y as described above. The correlation between $\boldsymbol{U}$ and 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.

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

  Bühlmann et al. [2014] uses 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.

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: # 4003.

So taking note of the obvious difficulty I go on: 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 give 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 marginally larger than that of X_{4002} with Y, this ordering is reversed after eliminating 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 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.
- 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 preprocessing of the gene expression data.

More or less the end of what I said in Lecture 1
From here on the notes have not been updated

1.2. **Targets of inference.** 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

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

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.

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 $\boldsymbol{X}^*$ 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 $\boldsymbol{X}\boldsymbol{\theta} = \mathbf{0}$. 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 (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 now 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 $\boldsymbol{U} = \boldsymbol{X}^\top \boldsymbol{Y}$.

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 kilogram 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 per unit of Y.
- In the penalty multiply any β_i by an estimate of scale for the variable X_i.
- Don't shrink the intercept. This is most easily handled by estimating α by $\bar{Y}$, the mean of the responses and then centering $\boldsymbol{Y}$ and each column of $\boldsymbol{X}$ by subtracting means.
- Scale $\boldsymbol{X}$ (after centering) 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.

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 $\sqrt{n}$.

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.)
- 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 $S_i\in\{\pm1\}$; 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 centering 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

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

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.

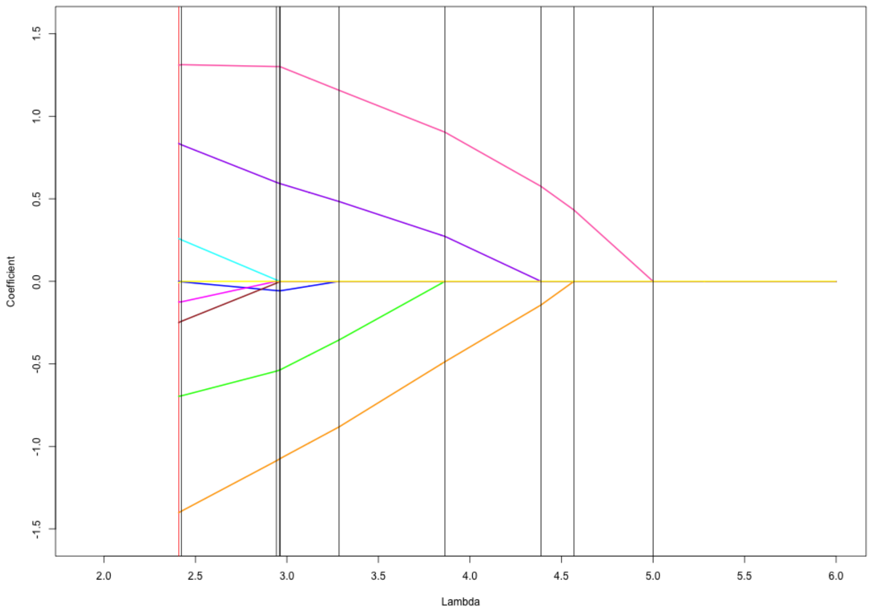
Now I show you in Figure 2.2 a frame from a movie. It shows the values of the 10 entries for $\hat{\beta}_{\lambda j}$ for j as in Table 2.2 plotted against $\lambda \in [\lambda_{10}, 6]$. The movie itself, which simply steps λ down from the right by small increments is available at **??**?. At each 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 $\hat{A}A_\lambda = \emptyset$. 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 **??** 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 $\boldsymbol{X}\hat{\beta}(\lambda)$ 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 **??**. 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 $\hat{\sigma}^2$. 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 β for which some component of the gradient is not defined I will just write down left and right derivatives.

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

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

Here $\beta\pm$ indicate a right $(+)$ or left $(-)$ partial derivative. The right and left derivatives differ, when $\beta_j^* = 0$, by 2λ.

What are these Karush-Kuhn-Tucker conditions? In the following I am going to need $\boldsymbol{X}_A$ to denote the submatrix of $\boldsymbol{X}$ with columns whose indices are in A and $\boldsymbol{X}_j$ to denote the jth column of $\boldsymbol{X}$.

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.

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

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

and

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

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

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

We say that the matrix $\boldsymbol{X}$ has columns in *general position* if that does not happen for any pair $i \neq j$. A design matrix with two identical columns or one column exactly equal to minus the other is a very doubtful design.

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 $\hat{\beta}_\lambda = 0$. Now consider $\lambda = \lambda_1 - \epsilon$ with $\epsilon > 0$ small. I claim that there is a $\epsilon > 0$ for which

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

Proof:

We will check to see that this $\boldsymbol{\beta}^*$ satisfies the conditions. We are saying $A = \{J_1\}$ and solving the equation

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

But remember that $\boldsymbol{X}^\top \boldsymbol{X}$ is the identity and that A has only a single column so that $\boldsymbol{X}_A^\top \boldsymbol{X}_A = 1$ and we must solve $\beta_A - U_{J_1} + S_1\lambda = 0$ which gives the desired formula.

For $j \neq J_1$ the left and right derivatives are

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

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

$$\rho_{jJ_1}(U_{J_1} - \lambda S_1) - U_j - \lambda < 0 < \rho_{jJ_1}(U_{J_1} - \lambda S_1) - U_j + \lambda$$

which becomes

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

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 $|\rho_{jJ_1}| \le 1$ 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 two that $|S_1\rho_{jJ_1}| < 1$ so we will not be dividing by a negative number. We learn that if, for each $j \ne J_1$ 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 $j \ne J_1$. Thus if

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

then

$$\hat\beta_{\lambda j} = \begin{cases} 0 & j \ne J_1 \\ U_{J_1} - \lambda S_1 & j = J_1. \end{cases}$$

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 the fits at λ_1 and λ_2 to get a test of the global null $\beta = 0$. At $\lambda = \lambda_1$ the fitted predictor is 0 and the covariance with Y is 0. At $\lambda = \lambda_2$ the fitted predictor is $X\hat\beta_{\lambda_2}$ the "covariance" is

$$\boldsymbol{Y}^\top\boldsymbol{X}\hat\beta_{\lambda_2}$$

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: 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 $n \to \infty$. 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_1^2 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 [1978b] 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

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

guarantees that

$$n(1 - F(a_n + b_n x)) \to G(x)$$

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 $\psi\prime(x)$. 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 $\psi\prime(x)$. The canonical theoretical choice is $b_n = nf(a_n)$ but there are many asymptotically equivalent choices.

- For a_p and b_p both more or less $\sqrt{2\log p}$ we have
$$a_p(|U_{[1]}| - b_p), a_p(|U_{[2]}| - b_p), \ldots, a_p(|U_{[K]}| - b_p)$$
has joint extreme value limit distribution; Weissman [1978a].
- Weak limit $W_1, \ldots, W_k$ has joint density
$$\exp\left(-w_1 - \cdots - w_k - e^{-w_k}\right) 1(w_k < \cdots < w_1)$$
- In fact we may take
$$a_p = \sqrt{2\log p}$$
and
$$b_p = a_p - \frac{\log\log p + \log\pi}{2a_p}.$$

Consequences

- Implication:
$$a_p(|U_{[1]}| - |U_{[2]}|) \Longrightarrow \text{Exponential}(1).$$
- And $|U_{[1]}|/a_p \to 1$ so
$$|U_{[1]}|(|U_{[1]}| - |U_{[2]}|) \Longrightarrow \text{Exponential}(1).$$

- Indeed under the global null with Gaussian errors
$$U_{[1]}|(|U_{[1]}| - |U_{[2]}|), \ldots, U_{[k]}(U_{[k]} - U_{[k+1]})$$
converges in law to
$$E_1, E_2/2, \ldots, E_k/k$$
where the E_i are iid standard exponential.

How to handle general $\boldsymbol{X}^\top \boldsymbol{X}$

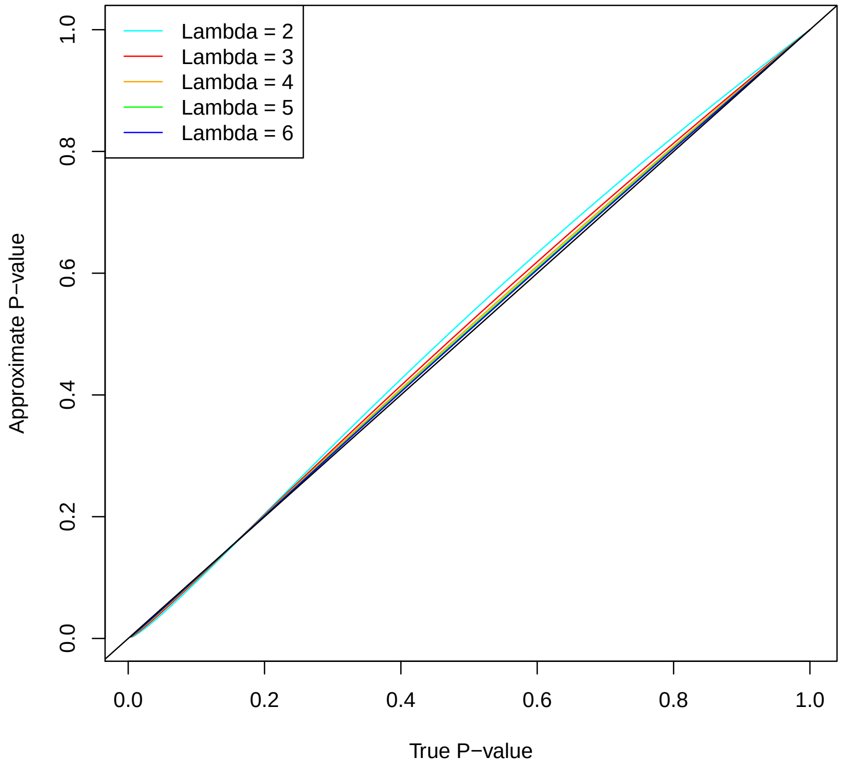
- Notice $U_{[1]}$ is NOT independent of $U_{[2]}$.
- But given $J_1 = j_1$, $U_{[2]}$ computed from the U_j with $j \neq j_1$.
- So conditional law of $U_{[1]}$ given $J_1 = j_1, S_1 = 1$ AND $U_{[2]}$ is Gaussian truncated to range
$$(|U_{[2]}|, \infty).$$
- This part remains true for general designs!
- So what is conditional law of
$$U_{j_1}(U_{j_1} - \lambda_2)$$
given other U_j and $J_1 = j_1$ and $S_1 = 1$?

The tail of the normal distribution is exponential:

- Assume $Z \sim N(0,1)$ and $\mathrm{E}(Z) = 0$ and let $\lambda \to \infty$.
- Then
$$\lim_{\lambda \to \infty} P(Z(Z - \lambda) > x | Z > \lambda) = e^{-x} \text{ for } x > 0.$$
- Much better approx than usual extreme value theory.

There are two natural ways to plot the quality of this approximation. The first, in Figure 2.4 plots the approximate P-value
$$P(Z(Z - \lambda) > v | Z > \lambda) \approx e^{-v}$$
against the exact P-value
$$P(Z(Z - \lambda) > v | Z > \lambda) = P(Z > u(v, \lambda) | Z > \lambda) = \frac{1 - \Phi\{u(v, \lambda)\}}{1 - \Phi(\lambda)}$$
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 2.4 plots the ratio
$$\frac{e^{-v}}{P(Z > u(v, \lambda) | Z > \lambda)}$$
against $P(Z > u(v, \lambda) | Z > \lambda)$ when $\lambda =$???

FIGURE 1. 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 $\lambda \in \{2, 3, 4, 5, 6\}$



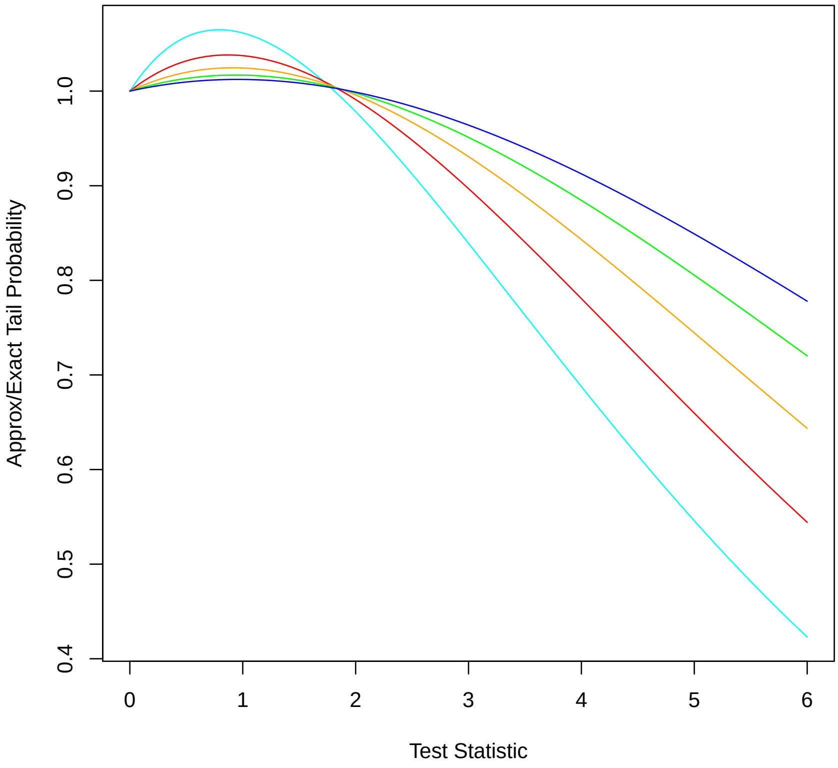
REFERENCES

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.

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

Ishay Weissman. Estimation of parameters and larger quantiles based on the k largest observations. *Journal of the American Statistical Association*, 73

FIGURE 2. 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 ??



(364):812–815, 1978a. ISSN 01621459. URL http://www.jstor.org/stable/2286285.

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