INFERENCE IN HIGH-DIMENSIONAL LINEAR MODELS COURSE NOTES

RICHARD LOCKHART

Contents

Version of 24 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.

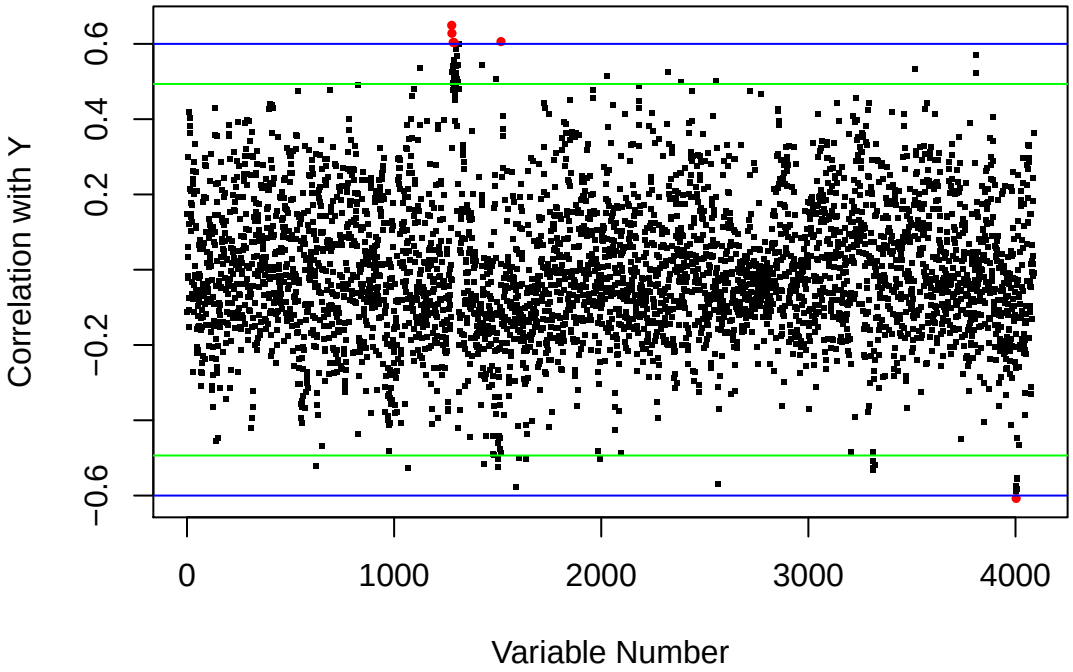
I am going to pretend that we have a sample of n independent and identically distributed vectors $(Y_i, X_{i1}, \ldots, X_{ip})$. I will start with the basic question of whether or not there is any relationship between any of the genes and riboflavin. We will need some notation.

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

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

so

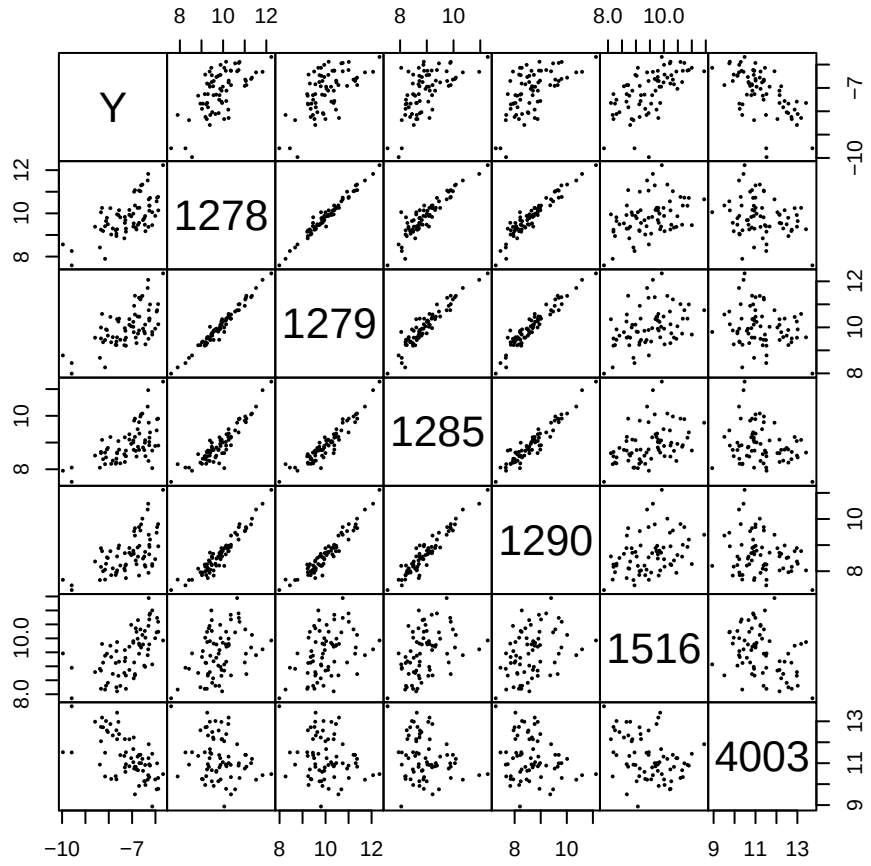
$$p_{\text{Bon}} = m \min_{1 \leq j \leq m} \{p_j\}$$

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

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

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 Y on two columns U, V with $U^TU = V^TV = 1$ and $U^TV = 1 - \epsilon$. 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 $\boldsymbol{U}^T\boldsymbol{V} = \pm 1$.

When we regress Y on U without an intercept we get a fitted slope $\hat{\alpha}_1 = U^T Y$ with mean β and variance 1 while if we regress Y on V without an intercept we get fitted slope $\hat{\alpha}_2 = V^T Y$ with mean $(1-\epsilon)\beta$. The covariance between these two estimates is

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

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

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.

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

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

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.

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

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

by ordinary least squares minimizing

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.