A practical evaluation of recent methods in high-dimensional inference
Talk by Charles Zheng

What I mean by practical evaluation is as follows: I will begin by discussing, on a high-level, the potential utility of inference for high-dimensional linear models, and then eventually show some empirical results for specific methods. What got me interested in the topic is simply being a student at Stanford, I am surrounded by this whirlwind of innovation all being developed for more or less the same central problem. And that is the problem of doing inference in this particular high-dimensional linear model, where some response Y is Gaussian distributed, and its mean is a linear transformation of the covariates X with coefficients β.

And what's motivating this all of this activity? For one, it's a very natural question to ask. After all, classical, low-dimensional linear regression is a central tool in numerous fields; it's used in econometrics to fit theoretical models of the economy to economic data; it's used in clinical studies to control for difference in age, gender, and so on in patients, and so on. So it's an interesting question to see if it would be possible to extend linear regression to the high-dimensional world. However, once we move to the high-dimensional world, we find that the questions we would normally ask for high-dimensional data are quite different from the type of questions which classical linear regression and generalized linear models are used to address. For instance, in biology, we would often like to use high-dimensional data to generate new hypotheses: to identify promising genes for further study. However, the linear model is so general, there may be many other potential applications of high-dimensional inference which have been yet to be uncovered. Therefore, while the domain of application of linear regression, and the connection between the theory of classical linear regression and its main application areas has been well-studied (take, for instance, the causal inference literature), we have to be prepared to start from scratch in thinking about how high-dimensional inference can be applied to a whole new set of problems than what we have seen in classical statistics.

This is a challenging task, because there is a gap between the theoretical settings under which most of these high-dimensional approaches have been derived, and the real-world problems where these inference procedures could be potentially applied. One issue is the fact that high-dimensional procedures are often derived under rather strong assumptions, such as correctness of the linear model, gaussianity of the errors, sparsity of β and so on. Of course, nobody knows of any real dataset where every one of these assumptions is actually met. In this regard it's worth mentioning that classical linear regression can be rather robust in the sense that we understand what happens when the relationship between Y and X is nonlinear, the errors are nongaussian and so on. Thus, observing the behavior of high-dimensional inference when these assumptions are violated is the main subject of this talk. But it's important to note that mismatch between the theoretical model and the real, unknown mechanisms of the data are not the only problem, or even the most important one. Even more crucially, there is a potential mismatch between the *theoretical objectives*, which are to identify non-zero regression coefficients, and the *practical objectives*, which may not even be very clearly defined. At the end of all of our theorizing, the result is an algorithm which takes a couple of inputs: the reponse Y, the covariates X, and some type of error-control parameter, and the output is a subset of those covariates. In the end, the practical usefulness of the procedure depends on how, exactly, the practitioner intends to use this resulting list of variables. One way to put it is that the practitioner wants to arrive at a list of 'interesting' variables–but the meaning of 'interesting' depends on the application. The applications where such inference procedures are useful are the applications where whether or not a variable is 'interesting' is somehow related to the size of its partial regression coefficient. In some particular applications, as we will see in a few slides, one can clearly define what qualities of a variable make it 'interesting' and further, why the size of regression coefficient would tell us something about the 'interestingness' of the variable. But in general the problem of what makes a variable interesting and why the regression coefficients should be of any relevance is a very complicated issue, and one that is beyond the scope of the present talk.

Let's move on to the particular methods I am going to discuss. First is inference for the classical linear model, which was developed by Karl Pearson in the 1930s. As I mentioned before, you do inference for this model in quite general settings, obtaining confidence intervals or confidence sets, and therefore it's easy to control marginal type I error rate or family-wise error rate under classical low-dimensional settings this way. Next, I will present results on three methods for high-dimensional inference.

These particular three high-dimensional inference methods by no means represent an exhaustive or unbiased list: indeed, I started by studying the approaches which I am familiar with, and so it happens that all of these methods are the result of work of Stanford faculty and their collaborators. Another similarity of these methods is that they all involve Lasso regression in some way or another. By varying the lasso penalty from infinity to zero, you get a sequence of models starting with the null model (with zero variables) and increasing in size from there. This gives you a ranking of the variables, where the strength of the variable is measured by the first time it appeared in this sequence of models, with the strongest variables appearing the earliest in this sequence. This ranking of variables is used by the covariance test, which is a procedure resembling forward stepwise regression. The covariance test, by Lockhart, Taylor, and two Tibshiranis is a sequential procedure which uses this ranking: one begins by testing the siginificance of the strongest variable in the ranking; if this variable is accepted, the procedure terminates; otherwise, we continue testing the second-strongest variable, third-strongest and so on until we accept the null hypothesis. This procedure controls for Type I error in *some sense* but it not really clear how the meaning of Type I control offered by this method relates to family-wise error or false discovery rate. Max, a recent alumni, has a paper on how to control FDR using this type of procedure, but for this talk I'm going to apply the covariance test rather naively and see what happens.

The second method I'm going to consider is the debiased lasso (called debiased M-estimator in their paper) approach by Javanmard and Montanari. As we all know, LASSO is a biased estimator. Because of this bias, it's hard to come up with confidence intervals for LASSO, for one, it's not enough to know the variance of the coefficients. Therefore this approach introduces a 'debiased' version of the lasso estimate which is no longer sparse, but now you can make confidence intervals for high-dimensions. This approach is perhaps the most comparable to classical inference, in terms of offering the same kinds of Type I error control. On the other hand, it makes stronger assumptions than the other two approaches I consider here, involving the sparsity of β and also relying on asymptotics.

Of course, high dimensions are precisely where non-classical notions of Type I error control become appealing, and here we have the third method, the knockoff filter, which does multiple hypothesis testing for the linear model controlling for false discovery rate. The current paper on knockoffs by Barber and Candes only considers its application for p smaller than $n/2$, but there is ongoing work on extending knockoffs for high dimensions and also controlling for other Type I error control criteria.

Besides all the other work in high-dimensional regression I'm omitting, there is also by now a separate but related area of selective inference; which is also used in high dimensional settings but it is solving a quite different problem.

And actually, there is one quite important method which I have not yet mentioned for the high-dimensional inference problem, and it's what is most frequently used in practice, marginal screening. Marginal screening is a legitimate statistical technique, but it's solving a totally different hypothesis testing problem than in regression: it is testing correlations between variables and the response. Yet it ends up as a competing method to regression in practical problems, and the reason is because large correlations may be just as interesting to the practitioner as large regression coefficients, depending on the application. It is not unreasonable to use marginal screening even in the case that you are trying to find variables with large regression coefficient, because despite the theoretical counterexamples you could make, chances are, in any real problem the variables with large regression coefficient also have large correlations. Rather, the real shortcoming of marginal screening is in problems where there are *too many* variables correlated with the response, and where many of them are 'redundant.'

Here we see an example of a scenario where this might occur. Take a genome-wide association study, where we are trying to pinpoint the genes which are associated with a particular phenotype, for example an individual's Body Mass Index. The response Y is the subjects' body mass index and the covariates X are their SNPs, as measured by sequencing. One might imagine that the genetic component of the phenotype is contained in only a handful of genes: it is mutations in these genes (or SNPs) that contribute to the individuals' susceptibility to obsesity and therefore their BMI. So here we have a clear criteria for what makes a variable, a particular SNP, interesting: it is that the variable has a direct causal link to the phenotype. But here, the correlation of the variable to the phenotype is not a good match to the criteria for interestingness, since many mutations with indirect links to the phenotype will therefore be correlated to the variable. On the other hand, supposing all the relationships in this diagram are linear, the interesting variables are precisely the variables with nonzero regression coefficients and the uninteresting variables, the variables with no direct link to the phenotype, are precisely the variables with null regression coefficients. So here we have conditions under which regression is the right approach to the problem and marginal screening is the wrong approach. But all of this is conditional. Supposing all the relationships are linear, then there is a direct match between non-null variables and interesting variables; but supposing the links in this diagram are actually nonlinear, then it's no longer true that the non-null variables are necessarily the interesting variables.

This GWAS example is in some ways an ideal example; after all, it is biostatistical applications like this which form much of the original motivation for high-dimensional inference, however even in this specific application we see quite a gap betweeen theory and practice. In the theoretical world, we have a very specific model with some rather strong assumptions. So the only way to tell if these methods are really appropriate for the application is to take a specific problem in the world of genetics which has actually been already solved–to take some phenotype, for example, some Mendelian diseases, which are thoroughly studied and well-understood, for which we can say very confidently, what constitutes the ground truth of which SNPs are interesting for this particular phenotype. Then we would take a number of datasets, or the size which is typical for a new GWAS study, and see if our high-dimensional inference procedures can be used to produce a list of SNPs which are similar to the correct list of SNPs, which we know from all our accumulated knowledge. Notably, Barber and Candes validated their knockoff procedure on virus mutation data, but there the ground truth was by no means very definitive and besides, we would ideally want numerous such validation studies to be performed if we wanted to arrive at a good understanding of the utility of any statistical procedure for the application.

As we have seen, ground truth (in the practical sense) is very tricky to obtain, for one because the very definition of ground truth is dependent on the sceintific details of the problem. It is a little more hopeful, however, to imagine that we could potentially obtain a *statistical ground truth*, in the sense that we know the parameters of the best approximation of our particular model to the truth. What I mean is that as long as we collect enough data–and also as long as we have a definition for what constitutes the population–we could get a good idea of the population regression coefficients β for a particular problem. Whether these coefficients β are meaningful is up to the domain experts, but as statisticians we can use this knowledge to check that we are indeed succeeding at the statistical problem we are trying to solve. This kind of study would be reassuring because our theory does not how well we perform on the hypothesis testing problem in the real world, where our assumptions most likely do not hold.

This kind of practical validation of robustness could be complemented on the theoretical side by an effort to relax the assumptions and develop new, robust methods which hold under these weaker assumptions. Or alternatively, rather than propose new methods, we could simply study how the existing methods perform in specific models in which the assumptions are weakened: taking specific models of nonlinearity, non-Gaussianity, and so on, and observing the precise relation of how the degree of nonlinearity or degree of non-homogeneity affects the properties of the method.

What I'm going to propose lies in none of these categories, but it complements all of these approaches. I'm going to study the performance of these high-dimensional inference problems neither in real data, nor in simulations (where all of the deviations from the assumptions are precisely controlled), but rather in some half-way space between reality and simulation. I'm going to propose validating these procedures using real data augmented with synthetic negative controls.

In some ways, what I'm proposing is a poor man's substitute for validation on real data with statistical ground truth; that is, validation on real data where you know the true value of β. Here, we define β in a model-free way as the coefficients of the best linear prediction of y conditional x, which is given explicitly by this expectiation formula, a population version of the classic $(X'X)^{-1}X'Y$ formula. The whole problem is that in real data we don't know the true β.

But here's an idea. Suppose I have some data consisting of a response Y and low-dimensional (say five-demensional) covariates X. I could easily use ordinary least-squares to get a good estimate of β. But this is a little too easy, so I'm going to turn this into a high-dimensional problem by generating new columns of the design matrix out of pure noise. I'm going to hand this expanded design matrix $\tilde{X}$ and the response Y without telling you which columns were from the original data and which ones were artifically generated. Now if you wanted to do ordinary least squares, you would get very bad results, or you couldn't do it at all since with the added noise variables, the dimension becomes larger than n. However, if you applied high-dimensional inference to do variable selection, you end up selecting mostly columns from the original data. And I could tell if you made any mistakes, since I generated the fake columns. I've created a situation where I know the ground truth, yet in which the problem is not wholly artificial, since I used parts of a real dataset to construct this problem.

Synthetic negative controls goes a little further than this idea of adding pure noise columns. Given a p-dimensional covariate vector x,..