The Annals of Statistics
2014, Vol. 42, No. 2, 501–508
DOI: 10.1214/13-AOS1175E
Main article DOI: 10.1214/13-AOS1175

DISCUSSION: "A SIGNIFICANCE TEST FOR THE LASSO"

BY LARRY WASSERMAN

Carnegie Mellon University

The paper by Lockhart, Taylor, Tibshirani and Tibshirani (LTTT) is an important advancement in our understanding of inference for high-dimensional regression. The paper is a tour de force, bringing together an impressive array of results, culminating in a set of very satisfying convergence results. The fact that the test statistic automatically balances the effect of shrinkage and the effect of adaptive variable selection is remarkable.

The authors make very strong assumptions. This is quite reasonable: to make significant theoretical advances in our understanding of complex procedures, one has to begin with strong assumptions. The following question then arises: what can we do without these assumptions?

1. The assumptions. The assumptions in this paper—and in most theoretical papers on high-dimensional regression—have several components. These include:

(1) The linear model is correct.
(2) The variance is constant.
(3) The errors have a Normal distribution.
(4) The parameter vector is sparse.
(5) The design matrix has very weak collinearity. This is usually stated in the form of incoherence, eigenvalue restrictions or incompatibility assumptions.

To the best of my knowledge, these assumptions are not testable when $p > n$. They are certainly a good starting place for theoretical investigations but they are indeed very strong. The regression function $m(x) = \mathbb{E}(Y|X = x)$ can be any function. There is no reason to think it will be close to linear. Design assumptions are also highly suspect. High collinearity is the rule rather than the exception especially in high-dimensional problems. An exception is signal processing, in particular compressed sensing, where the user gets to construct the design matrix. In this case, if the design matrix is filled with independent random Normals, the design matrix will be incoherent with high probability. But this is a rather special situation.

None of this is meant as a criticism of the paper. Rather, I am trying to motivate interest in the question I asked earlier, namely: what can we do without these assumptions?

Received December 2013.

REMARK 1. It is also worth mentioning that even in low-dimensional models and even if the model is correct, model selection raises troubling issues that we all tend to ignore. In particular, variable selection makes the minimax risk explode [Leeb and Pötscher (2005, 2008)]. This is not some sort of pathological risk explosion, rather, the risk is large in a neighborhood of 0, which is a part of the parameter space we care about.

2. The assumption-free lasso. To begin it is worth pointing out that the lasso has a very nice assumption-free interpretation.

Suppose we observe $(X_1, Y_1), \ldots, (X_n, Y_n) \sim P$ where $Y_i \in \mathbb{R}$ and $X_i \in \mathbb{R}^d$. The regression function $m(x) = \mathbb{E}(Y|X = x)$ is some unknown, arbitrary function. We have no hope to estimate $m(x)$ nor do we have licence to impose assumptions on m.

But there is sound theory to justify the lasso that makes virtually no assumptions. In particular, I refer to Greenshtein and Ritov (2004) and Juditsky and Nemirovski (2000).

Let $\mathcal{L} = \{x^t\beta : \beta \in \mathbb{R}^d\}$ be the set of linear predictors. For a given β, define the predictive risk

$$R(\beta) = E\big(Y - \beta^T X\big)^2,$$

where (X, Y) is a new pair. Let us define the best, sparse, linear predictor $\ell_*(x) = \beta_*^T x$ (in the ℓ_1 sense) where β_* minimizes $R(\beta)$ over the set $B(L) = \{\beta : \|\beta\|_1 \le L\}$. The lasso estimator $\hat{\beta}$ minimizes the empirical risk $\hat{R}(\beta) = \frac{1}{n}\sum_{i=1}^n (Y_i - \beta^T X_i)^2$ over $B(L)$. For simplicity, I will assume that all the variables are bounded by C (but this is not really needed). We make no other assumptions: no linearity, no design assumptions and no models. It is now easily shown that

$$R(\hat{\beta}) \le R(\beta_*) + \sqrt{\frac{8C^2L^4}{n}\log\left(\frac{2p^2}{\delta}\right)}$$

except on a set of probability at most δ.

This shows that the predictive risk of the lasso comes close to the risk of the best sparse linear predictor. In my opinion, this explains why the lasso "works." The lasso gives us a predictor with a desirable property—sparsity—while being computationally tractable and it comes close to the risk of the best sparse linear predictor.

3. Interlude: Weak versus strong modeling. When developing new methodology, I think it is useful to consider three different stages of development:

(1) Constructing the method.
(2) Interpreting the output of the method.
(3) Studying the properties of the method.

I also think it is useful to distinguish two types of modeling. In *strong modeling*, the model is assumed to be true in all three stages. In *weak modeling*, the model is assumed to be true for stage 1 but not for stages 2 and 3. In other words, one can use a model to help construct a method. But one does not have to assume the model is true when it comes to interpretation or when studying the theoretical properties of the method. My discussion is guided by my preference for weak modeling.

4. Assumption-free inference: The HARNESS. Here, I would like to discuss an approach I have been developing with Ryan Tibshirani. We call this: High-dimensional Agnostic Regression Not Employing Structure or Sparsity, or, the HARNESS. The method is a variant of the idea proposed in Wasserman and Roeder (2009).

The idea is to split the data into two halves. $\mathcal{D}_1$ and $\mathcal{D}_2$. For simplicity, assume that n is even so that each half has size $m = n/2$. From the first half $\mathcal{D}_1$, we select a subset of variables S. The method is agnostic about how the variable selection is done. It could be forward stepwise, lasso, elastic net or anything else. The output of the first part of the analysis is the subset of predictors S and an estimator $\hat{\beta} = (\hat{\beta}_j : j \in S)$. The second half of the data $\mathcal{D}_2$ is used to provide distribution-free inferences for the following questions:

(1) What is the predictive risk of $\hat{\beta}$?
(2) How much does each variable in S contribute to the predictive risk?
(3) What is the best linear predictor using the variables in S?

All the inferences from $\mathcal{D}_2$ are interpreted as being conditional on $\mathcal{D}_1$. (A variation is to use $\mathcal{D}_1$ only to produce S and then construct the coefficients of the predictor from $\mathcal{D}_2$. For the purposes of this discussion, we use $\hat{\beta}$ from $\mathcal{D}_1$.)

In more detail, let

$$R = \mathbb{E}|Y - X^T\hat{\beta}|,$$

where the randomness is over the new pair (X, Y); we are conditioning on $\mathcal{D}_1$. Note that in this section I have changed the definition of R to be on the absolute scale which is more interpretable. In the above equation, it is understood that $\hat{\beta}_j = 0$ when $j \notin S$. The first question refers to producing a estimate and confidence interval for R (conditional on $\mathcal{D}_1$). The second question refers to inferring

$$R_j = \mathbb{E}|Y - X^T\hat{\beta}_{(j)}| - \mathbb{E}|Y - X^T\hat{\beta}|$$

for each $j \in S$, where $\beta_{(j)}$ is equal to $\hat{\beta}$ except that $\hat{\beta}_j$ is set to 0. Thus, R_j is the risk inflation by excluding X_j. The third question refers to inferring

$$\beta^* = \arg\min_{\beta \in \mathbb{R}^k} \mathbb{E}(Y - X_S^T\beta)^2$$

the coefficient of the best linear predictor for the chosen model. We call β^* the *projected parameter*. Hence, $x^T\beta^*$ is the best linear approximation to $m(x)$ on the linear space spanned by the selected variables.

A consistent estimate of R is

$$\hat{R} = \frac{1}{m}\sum_{i=1}^{m}\delta_i,$$

where the sum is over $\mathcal{D}_2$, and $\delta_i = |Y_i - X_i^T\hat{\beta}|$. An approximate $1-\alpha$ confidence interval for R is $\hat{R} \pm z_{\alpha/2}s/\sqrt{m}$ where s is the standard deviation of the δ_i's.

The validity of this confidence interval is essentially distribution-free. In fact, if we want to be purely distribution-free and avoid asymptotics, we could instead define R to be the median of the law of $|Y - X^T\hat{\beta}|$. Then the order statistics of the δ_i's can be used in the usual way to get a finite sample, distribution-free confidence interval for R.

Estimates and confidence intervals for R_j can be obtained from $e_1, \ldots, e_m$ where

$$e_i = |Y_i - X^T\hat{\beta}_{(j)}| - |Y_i - X^T\hat{\beta}|.$$

Estimates and confidence intervals for β_* can be obtained by standard least squares procedures based on $\mathcal{D}_2$. The steps are summarized in Figure 1.

The HARNESS bears some similarity to POSI [Berk et al. (2013)] which is another inference method for model selection. They both eschew any assumption that the linear model is correct. But POSI attempts to make inferences that are valid over all possible selected models while the HARNESS restricts attention to the selected model. Also, the HARNESS emphasizes predictive inferential statement.

Here is an example using the wine dataset. (Thanks to the authors for providing the data.) Using the first half of the data, we applied forward stepwise selection and

---

The HARNESS

Input: data $\mathcal{D} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$.

(1) Randomly split the data into two halves $\mathcal{D}_1$ and $\mathcal{D}_2$.

(2) Use $\mathcal{D}_1$ to select a subset of variables S. This can be forward stepwise, the lasso, or any other method.

(3) Let $R = \mathbb{E}((Y - X^T\hat{\beta})^2|\mathcal{D}_1)$ be the predictive risk of the selected model on a future pair (X, Y), conditional on $\mathcal{D}_1$.

(4) Using $\mathcal{D}_2$ construct point estimates and confidence intervals for R, $(R_j : \hat{\beta}_j \neq 0)$ and β_*.

FIG. 1. *The steps in the HARNESS algorithm.*

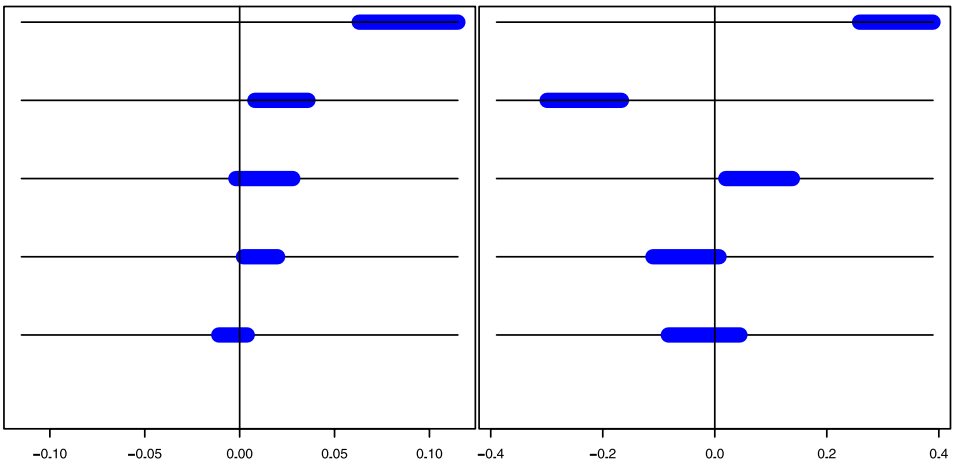

FIG. 2. *Left plot*: *confidence intervals for* R_j. *Right plot*: *confidence intervals for projected parameters*. *From top down the variables are Alcohol*, *Volatile_Acidity*, *Sulphates*, *Total_Sulfur_Dioxide and pH*.

used C_p to select a model. The selected variables are Alcohol, Volatile_Acidity, Sulphates, Total_Sulfur_Dioxide and pH. A 95 percent confidence interval for the predictive risk of the null model is (0.65,0.70). For the selected model, the confidence interval for R is (0.46,0.53). The (Bonferroni-corrected) 95 percent confidence intervals for the R_j's are shown in the first plot of Figure 2. The (Bonferroni-corrected) 95 percent confidence intervals for the parameters of the projected model are shown in the second plot in Figure 2.

5. The value of data-splitting. Some statisticians are uncomfortable with data-splitting. There are two common objections. The first is that the inferences are random: if we repeat the procedure we will get different answers. The second is that it is wasteful.

The first objection can be dealt with by doing many splits and combining the information appropriately. This can be done but is somewhat involved and will be described elsewhere. The second objection is, in my view, incorrect. The value of data splitting is that leads to simple, assumption-free inference. There is nothing wasteful about this. Both halves of the data are being put to use. Admittedly, the splitting leads to a loss of power compared to ordinary methods *if the model were correct*. But this is a false comparison since we are trying to get inferences without assuming the model is correct. It is a bit like saying that nonparametric function estimators have slower rates of convergence than parametric estimators. But that is only because the parametric estimators invoke stronger assumptions.

6. Conformal prediction. Since I am focusing my discussion on regression methods that make weak assumptions, I would also like to briefly mention

Vladimir Vovk's theory of conformal inference. This is a completely distribution-free, finite sample method for predictive regression. The method is described in Vovk, Gammerman and Shafer (2005) and Vovk, Nouretdinov and Gammerman (2009). Unfortunately, most statisticians seem to be unaware of this work which is a shame. The statistical properties (such as minimax properties) of conformal prediction were investigated in Lei, Robins and Wasserman (2014), Lei and Wasserman (2014).

A full explanation of the method is beyond the scope of this discussion but I do want to give the general idea and say why it is related the current paper. Given data $(X_1, Y_1), \ldots, (X_n, Y_n)$, suppose we observe a new X and want to predict Y. Let $y \in \mathbb{R}$ be an arbitrary real number. Think of y as a tentative guess at Y. Form the augmented data set

$$(X_1, Y_1), \ldots, (X_n, Y_n), (X, y).$$

Now we fit a linear model to the augmented data and compute residuals e_i for each of the $n + 1$ observations. Now we test $H_0 : Y = y$. Under H_0, the residuals are invariant under permutations and so

$$p(y) = \frac{1}{n+1} \sum_{i=1}^{n+1} I\big(|e_i| \geq |e_{n+1}|\big)$$

is a distribution-free p-value for H_0.

Next, we invert the test: let $C = \{y : p(y) \geq \alpha\}$. It is easy to show that

$$\mathbb{P}(Y \in C) \geq 1 - \alpha.$$

Thus, C is distribution-free, finite-sample prediction interval for Y. Like the HARNESS, the validity of the method does not depend on the linear model being correct. The set C has the desired coverage probability no matter what the true model is. Both the HARNESS and conformal prediction use the linear model as a device for generating predictions but neither requires the linear model to be true for the inferences to be valid. [In fact, in the conformal approach, any method of generating residuals can be used. It does not have to be a linear model. See Lei and Wasserman (2014).]

One can also look at how the prediction interval C changes as different variables are removed. This gives another assumption-free method to explore the effects of predictors in regression. Minimizing the length of the interval over the lasso path can also be used as a distribution-free method for choosing the regularization parameter of the lasso.

On a related note, we might also be interested in assumption-free methods for the related task of inferring graphical models. For a modest attempt at this, see Wasserman, Kolar and Rinaldo (2013).

7. Causation. LTTT do not discuss causation. But in any discussion of the assumptions underlying regression, causation is lurking just below the surface. Indeed, there is a tendency to conflate causation and inference. To be clear: prediction, inference and causation are three separate ideas.

Even if the linear model is correct, we have to be careful how we interpret the parameters. Many articles and textbooks describe β_j as the change in Y if X_j is changed, holding the other covariates fixed. This is incorrect. In fact, β_j is the *change in our prediction of* Y if X_j is changed. This may seem like nit-picking but this is the very difference between association and causation.

Causation refers to the change in Y as X_j is changed. Association (prediction) refers to the change *in our prediction of* Y as X_j is changed. Put another way, prediction is about $\mathbb{E}(Y|$ observe $X = x)$ while causation is about $\mathbb{E}(Y|$ set $X = x)$. If X is randomly assigned, they are the same. Otherwise, they are different. To make the causal claim, we have to include in the model, every possible confounding variable that could affect both Y and X. This complete causal model has the form

$$Y = g(X, Z) + \varepsilon,$$

where $Z = (Z_1, \ldots, Z_k)$ represents all confounding variables in the world. The relationship between Y and X alone is described as

$$Y = f(X) + \varepsilon'.$$

The causal effect—the change in Y as X_j is changed—is given by $\partial g(x, z)/\partial x_j$. The association (prediction)—the change in our prediction of Y as X_j is changed—is given by $\partial f(x)/\partial x_j$. If there are any omitted confounding variables, then these will be different.

Which brings me back to the paper. Even if the linear model is correct, we still have to exercise great caution in interpreting the coefficients. Most users of our methods are nonstatisticians and are likely to interpret β_j causally no matter how many warnings we give.

8. Conclusion. LTTT have produced a fascinating paper that significantly advances our understanding of high-dimensional regression. I expect there will be a flurry of new research inspired by this paper.

My discussion has focused on the role of assumptions. In low-dimensional models, it is relatively easy to create methods that make few assumptions. In high-dimensional models, low assumption inference is much more challenging.

I hope I have convinced the authors that the low assumption world is worth exploring. In the meantime, I congratulate the authors on an important and stimulating paper.

Acknowledgments. Thanks to Rob Kass, Rob Tibshirani and Ryan Tibshirani for helpful comments.

REFERENCES

BERK, R., BROWN, L., BUJA, A., ZHANG, K. and ZHAO, L. (2013). Valid post-selection inference. *Ann. Statist.* **41** 802–837. MR3099122

GREENSHTEIN, E. and RITOV, Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971–988. MR2108039

JUDITSKY, A. and NEMIROVSKI, A. (2000). Functional aggregation for nonparametric regression. *Ann. Statist.* **28** 681–712. MR1792783

LEEB, H. and PÖTSCHER, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory* **21** 21–59. MR2153856

LEEB, H. and PÖTSCHER, B. M. (2008). Sparse estimators and the oracle property, or the return of Hodges' estimator. *J. Econometrics* **142** 201–211. MR2394290

LEI, J., ROBINS, J. and WASSERMAN, L. (2013). Distribution-free prediction sets. *J. Amer. Statist. Assoc.* **108** 278–287.

LEI, J. and WASSERMAN, L. (2014). Distribution-free prediction bands for non-parametric regression. *J. R. Stat. Soc. Ser. B.* **76** 71–96.

VOVK, V., GAMMERMAN, A. and SHAFER, G. (2005). *Algorithmic Learning in a Random World*. Springer, New York. MR2161220

VOVK, V., NOURETDINOV, I. and GAMMERMAN, A. (2009). On-line predictive linear regression. *Ann. Statist.* **37** 1566–1590. MR2509084

WASSERMAN, L., KOLAR, M. and RINALDO, A. (2013). Estimating undirected graphs under weak assumptions. Preprint. Available at arXiv:1309.6933.

WASSERMAN, L. and ROEDER, K. (2009). High-dimensional variable selection. *Ann. Statist.* **37** 2178–2201. MR2543689

DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
5000 FORBES AVE.
PITTSBURGH, PENNSYLVANIA 15213
USA
E-MAIL: larry@stat.cmu.edu