When 'false' models predict better than 'true' ones: Paradoxes of the bias-variance tradeoff

Momin M. Malik[*]
Institute for Software Research
Department of Computer Science
Carnegie Mellon University
momin.malik@cs.cmu.edu

Hemank Lamba
Institute for Software Research
Department of Computer Science
Carnegie Mellon University
hlamba@cs.cmu.edu

December 2017
v1.6

1 Introduction

The distinction between a model whose purpose is *prediction* and a model that tries to capture associations (and possibly capturing a causal or data-generating process) for the purpose of *explanation* or *information* (what Mullainathan and Spiess [2017] contrast respectively as $\widehat{\beta}$ problems and $\widehat{Y}$ problems) is little-appreciated outside of statistics and machine learning theory (Hofman et al., 2017; Shmueli, 2010; Breiman, 2001), but it is enormously consequential. It is deeply unintuitive that the best-fitting model from a given class could somehow be 'less true' than a worse-fitting model, although we know from theory (and can demonstrate via simulation) that this is a definite possibility.

The consequences are far-ranging. In the social sciences, policymakers who take advice that model selection procedures and ensemble methods that reduce variance are more 'true' than theory-driven model building (Hindman, 2015) might end up using these methods for planning interventions, failing to recognize that such use cases, while present, are limited and must be specifically identified as such (Athey, 2017; Kleinberg et al., 2015); and that treatment effects cannot be reliable measured with model selection (Rolling and Yang, 2014). In bioinformatics, Yang and Yang (2016) worry that the limitations of model selection techniques mean that their use will not necessarily yield trustworthy, reproducible scientific results. On the other hand, not appreciating the role of flexible models may lead to their lack of use for exploratory purposes in science or for engineering goals (Lin, 2015), including policy applications in which prediction is sufficient (Kleinberg et al., 2015). Shmueli (2010) charges that the limited awareness of this distinction outside of statistics is a major failure of statistical communication; in exploring aspects that separate prediction from explanation, we hope to contribute to understandings of this distinction that can be further communicated.

In this paper, we propose explaining the gap between predictive performance and model 'truth' in terms of the bias-variance tradeoff. That is, we will show how the bias-variance tradeoff of the prediction error, $\text{EPE} = \sigma^2 + \text{bias}^2(\widehat{f}) + \text{Var}(\widehat{f})$, relates to model truth, where some estimate $\widehat{f}$ has the same terms, functional form, and (potential) features as f.

We operationalize 'true' and 'false' in terms *model specification*: a 'true' model is one that is correctly specified, i.e. that has the same terms and the same functional form as the data-generating process (DGP). A 'false' model is one that is incorrectly specified (misspecified), i.e. that has different terms and/or a different functional form from the true DGP. Generally we would assume that, given the correct specification, consistent estimation and correct inferences will follow from application of standard statistical procedures; we note that post-selection inference (Berk et al., 2013) raises the possibility of true inferences of given terms under potential misspecification, in which case the 'truth' is of confidence intervals around fitted weights for given terms (in this case, inferences would be correct if the confidence interval around the estimated coefficient of a feature not present in the true model contains zero).

[*]Corresponding author.

We do not consider the obvious cases of fitted models being false because the true DGP is not from a tractable class (or, more philosophically, that the DGP cannot be meaningfully described by any mathematical approximation), or that we do not know that we have not measured (or we cannot measure) all causal features. We recognize that these are the core reasons why "all models are wrong" (Box, 1979), but they also do not permit formal analysis. Instead, we look at three topics: first, an explicit example of the conditions under which an underspecified linear model will predict better than a true model. Second, extending this approach, we review what would happen if we apply the lasso in such a situation, looking at the conditions under which the lasso might recover the best-predicting terms (the terms of the 'oracle' predictor). In these two topics, the bias-variance tradeoff explains why biased models can perform better than unbiased ones.

For our third topic we consider a complication to this view: we review the ways in which the bias-variance decomposition is specific to L_2 loss, and that other loss functions do not necessarily have a similar tradeoff. We might wonder if, under other loss functions, there is no bias-variance tradeoff such that decreasing variance at the cost of increasing bias might not actually lead to better performance; as it turns out, not only is this *not* the case, but the situation is even stranger: we review a generalization of the bias and variance that shows how, in 0-1 loss, increasing the variance without changing the bias can actually lower the error rate. This points towards the difference between prediction and explanation being even more subtle and strange than what can be understood through the bias-variance tradeoff. Still, developments such as post-selection inference point the way to reconciling these two goals by making inferential statements that take into account the uncertainty introduced by models striving to maximize prediction.

2 Misspecification and prediction

A true model will be, by definition, unbiased (i.e., when its terms and form are used for estimation). So, any model that achieves lower test error than the true model by lowering estimator variance at the cost of estimator bias will, by definition, be trading off 'truth' for predictive performance. Of course, we try to use procedures that are both asymptotically unbiased and consistent, such that both bias and variance go to zero asymptotically; but in finite samples, the tradeoff matters. Kunst (2008), in a simulation study that also shows a true model performing worse than a model built for prediction, notes that there is a "conflict between the aims of optimizing finite-sample prediction and of finding the true data-generating model class."

We look to the literature to ask the question, will the DGP specification produce an oracle predictor, or can the oracle predictor itself be 'false'? After all, the oracle predictor is defined in terms of being the true minimizer of the MSE, and not in terms of representing the true DGP. Shmueli (2010) gives an example, drawn from Wu et al. (2007), of the conditions under which an underspecified linear model has a lower expected MSE (i.e., a lower expected prediction error) than a correctly specified linear model. We illustrate the proof specified by Wu et al. (2007). Consider that the true model is linear, given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where $\mathbf{X} \in \mathbb{R}^{n \times p+q}$.

The optimal estimate for the coefficients $\boldsymbol{\beta}$ is given by OLS estimates, which is

$$\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{H}\mathbf{y}$$

This is under the following assumptions:

- $\mathbf{X}$ has full column rank and is deterministic.
- The stochastic component $\boldsymbol{\varepsilon} = \varepsilon_1, ..., \varepsilon_n$ is an uncorrelated random sequence with mean zero and variance σ^2.

Given these assumptions, the OLS estimate is unbiased: $\mathbb{E}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}) = \boldsymbol{\beta}$ and $\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}) = \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}$.

The variance σ^2 can be estimated using the following expression:

$$s^2_{OLS} = \frac{S(\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})}{n-m} := \frac{(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})^T(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}})}{n-m}$$

where $m = p + q$ is the total number of parameters in the true model.

Now, consider an *underspecified* model as an example of misspecified model. The underspecified model in this case will be one which fails to include a subset of predictors.

Let the underspecified model that leaves out q features (subscript $-q$) be as follows:

$$\mathbf{y} = \mathbf{X}_{-q}\boldsymbol{\beta}_{-q} + \mathbf{e}$$

where $\mathbf{e} = \mathbf{X}_q\boldsymbol{\beta}_q + \boldsymbol{\varepsilon}$.

The parameters for this model can be learned by minimizing the objective function,

$$(\mathbf{y} - \mathbf{X}_{-q}\boldsymbol{\beta}_{-q})^T(\mathbf{y} - \mathbf{X}_{-q}\boldsymbol{\beta}_{-q})$$

which gives underspecified OLS estimates $\widehat{\boldsymbol{\beta}}_{-q}$:

$$\widehat{\boldsymbol{\beta}}_{-q} = \mathbf{H}_{-q}\mathbf{y}$$

The bias of this estimate (recall, assuming deterministic $\mathbf{X}$) is given by:

$$\begin{aligned}
\mathbb{E}(\widehat{\boldsymbol{\beta}}_{-q}) &= \mathbf{H}_{-q}\mathbb{E}(\mathbf{y}) \\
&= (\mathbf{X}_{-q}^T\mathbf{X}_{-q})^{-1}(\mathbf{X}_{-q}^T\mathbf{X}_{-q})\boldsymbol{\beta}_{-q} + (\mathbf{X}_{-q}^T\mathbf{X}_{-q})^{-1}(\mathbf{X}_{-q}^T\mathbf{X}_q)\boldsymbol{\beta}_q \\
&= \boldsymbol{\beta}_{-q} + \mathbf{A}_{-q}\boldsymbol{\beta}_q
\end{aligned}$$

where $\mathbf{A}_{-q} = (\mathbf{X}_{-q}^T\mathbf{X}_{-q})^{-1}(\mathbf{X}_{-q}^T\mathbf{X}_q)$ is the projection of $\mathbf{X}_{-q}$ on $\mathbf{X}_q$.The covariance of the estimate is given by

$$\mathrm{Cov}(\widehat{\boldsymbol{\beta}}_{-q}) = \sigma^2(\mathbf{X}_{-q}^T\mathbf{X}_{-q})^{-1}$$

These parameter estimates are generally biased, unless $\mathbf{A}_{-q}\boldsymbol{\beta}_q = 0$, which occurs only if $\boldsymbol{\beta}_q = 0$ or when $\mathbf{X}_q$ and $\mathbf{X}_{-q}$ are orthogonal.

The variance estimated from the under-specified model is:

$$s_{-q}^2 = \frac{S(\widehat{\boldsymbol{\beta}}_{-q})}{n-p} = \frac{(\mathbf{y} - \mathbf{X}_{-q}\widehat{\boldsymbol{\beta}}_{-q})^T(\mathbf{y} - \mathbf{X}_{-q}\widehat{\boldsymbol{\beta}}_{-q})}{n-p}$$

Note that under the true model, if we split $\widehat{\boldsymbol{\beta}}_{\mathrm{OLS}}$ into the OLS weights of the set of predictors left out by the underspecified model $\widehat{\boldsymbol{\beta}}_{-q,\mathrm{OLS}}$, and the OLS weights of the set of predictors of the underspecified model $\widehat{\boldsymbol{\beta}}_{q,\mathrm{OLS}}$, we can specify our estimates in the following sense:

$$\begin{bmatrix} \widehat{\boldsymbol{\beta}}_{-q,\mathrm{OLS}} \\ \widehat{\boldsymbol{\beta}}_{q,\mathrm{OLS}} \end{bmatrix} \sim \left(\begin{bmatrix} \boldsymbol{\beta}_{-q} \\ \boldsymbol{\beta}_q \end{bmatrix}, \sigma^2 \begin{bmatrix} \Gamma & \Psi \\ \Psi^T & \Omega \end{bmatrix} \right)$$

where

$$\begin{bmatrix} \Gamma & \Psi \\ \Psi^T & \Omega \end{bmatrix} = \begin{bmatrix} \mathbf{X}_{-q}^T\mathbf{X}_{-q} & \mathbf{X}_{-q}^T\mathbf{X}_q \\ \mathbf{X}_q^T\mathbf{X}_q & \mathbf{X}_q^T\mathbf{X}_q \end{bmatrix}^{-1}$$

Using matrix algebra, we can derive the following results:

$$\begin{aligned}
\Gamma &= (\mathbf{X}_{-q}^T(\mathbf{I}_n - \mathbf{H}_q)\mathbf{X}_{-q})^{-1} \\
&= (\mathbf{X}_{-q}^T\mathbf{X}_{-q})^{-1} + \mathbf{A}_{-q}(\mathbf{X}_q^T(\mathbf{I}_n - \mathbf{H}_{-q})\mathbf{X}_q)^{-1}\mathbf{A}_{-q}^T \\
\Omega &= (\mathbf{X}_{-q}^T(\mathbf{I}_n - \mathbf{H}_{-q})\mathbf{X}_q)^{-1} \\
&= (\mathbf{X}_q^T\mathbf{X}_q)^{-1} + \mathbf{A}_q(\mathbf{X}_{-q}^T(\mathbf{I}_n - \mathbf{H}_q)\mathbf{X}_{-q})^{-1}\mathbf{A}_q^T \\
\Psi &= -\mathbf{A}_{-q}\Omega
\end{aligned}$$

where

$$\mathbf{H}_q = \mathbf{X}_q(\mathbf{X}_q^T\mathbf{X}_q)^{-1}\mathbf{X}_q^T$$
$$\mathbf{H}_{-q} = \mathbf{X}_{-q}(\mathbf{X}_{-q}^T\mathbf{X}_{-q})^{-1}\mathbf{X}_{-q}^T$$
$$\mathbf{A}_q = (\mathbf{X}_q^T\mathbf{X}_q)^{-1})(\mathbf{X}_q^T\mathbf{X}_{-q})$$

Consider the mean squared error (MSE) as an empirical measure of loss. The MSE in the matrix can be specified by the 'mean squared error matrix' (MSEM). This is given by:

$$\text{MSEM}(\widehat{\boldsymbol{\beta}}) = E(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})E(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T = \text{Cov}(\widehat{\boldsymbol{\beta}}) + \Delta\Delta^T$$

where Δ is the bias of the estimator.

The MSEM difference between the coefficient for the true model and the under-specified model is given by

$$\text{MSEM}(\widehat{\boldsymbol{\beta}}_{q,\text{OLS}}) - \text{MSEM}(\widehat{\boldsymbol{\beta}}_{-q}) = \mathbf{A}_{-q}(\sigma^2(\mathbf{X}_q^T - (\mathbf{I}_n - \mathbf{H}_{-q})\mathbf{X}_q)^{-1} - \boldsymbol{\beta}_q\boldsymbol{\beta}_q^T)\mathbf{A}_{-q}^T$$

This is positive semi-definite if

$$\mathbf{X}_q^T - (\mathbf{I}_n - \mathbf{H}_{-q})\mathbf{X}_q \succeq 0$$

which is only possible when

$$\boldsymbol{\beta}_q^T\mathbf{X}_q^T(\mathbf{I}_n - \mathbf{H}_{-q})\mathbf{X}_q\boldsymbol{\beta}_q \leq q\sigma^2$$

Introduce R_c, the critical ratio, defined as:

$$R_c := \boldsymbol{\beta}_q^T\mathbf{X}_q^T(\mathbf{I}_n - \mathbf{H}_{-q})\mathbf{X}_q\boldsymbol{\beta}_q$$

From the above inequality, we obtain that the expected prediction error (expected MSE) for the underspecified model is lower than that of the correctly specified (true) model when:

$$R_c < q\sigma^2$$

R_c also provides an upper bound for the bias of the noise variance estimate under the underspecified model. Recall that $E(s^2_{\text{OLS}}) = \sigma^2$ and

$$E(s^2_{-q}) = \sigma^2 + \frac{\boldsymbol{\beta}_q^T\mathbf{X}_q^T(\mathbf{I}_n - \mathbf{H}_{-q})\mathbf{X}_q\boldsymbol{\beta}_q}{n-p}$$

Under what conditions is $R_c < q\sigma^2$? Some examples given by Shmueli (2010) are:

1. When σ^2 is large;
2. When the true absolute value of parameters of the left-out predictors, $\|\boldsymbol{\beta}_{-q}\|_1$, is small;
3. When there are high correlations among input variable settings so that the trace of $\mathbf{X}_q^T(\mathbf{I}_n - \mathbf{H}_{-q})\mathbf{X}_q$ is small; and
4. When there are a limit range on input conditions so that the trace of $\mathbf{X}_q^T(\mathbf{I}_n - \mathbf{H}_{-q})\mathbf{X}_q$ is small.

2.1 Simulation example

It helps to illustrate this concretely. Take $p = 13$, $q = 3$, $n = 100$, use 5-fold cross validation (80 observations for training, 20 for testing) and generate the data $\mathbf{X} = (\mathbf{X}_{-q}, \mathbf{X}_q)$ such that $\mathbf{X}_q$ is correlated with q features in $\mathbf{X}_{-q}$. We draw from a multivariate normal with mean zero and a covariance matrix that has a 1 along the diagonal, and a correlation of .99 between each pair of X_8 and X_{11}, X_9 and X_{12}, and X_{10} and X_{13}. The correlation matrix for generated data that captures this covariance structure is shown in figure (1).

After fixing the design matrix $\mathbf{X}$, we take $\boldsymbol{\beta}_{-q} = 10 \cdot \mathbf{1}_{10}$, $\boldsymbol{\beta}_q = \mathbf{1}_3$, and $\sigma = 10 \cdot \sqrt{R_c/q}$. Note that for $n = 100$ this gives $\sigma = 4.789$ (we calculate the R_c and the resulting σ from all of $\mathbf{X}$, not just the training instances), which is larger than the magnitude of the coefficients of the left-out predictors.

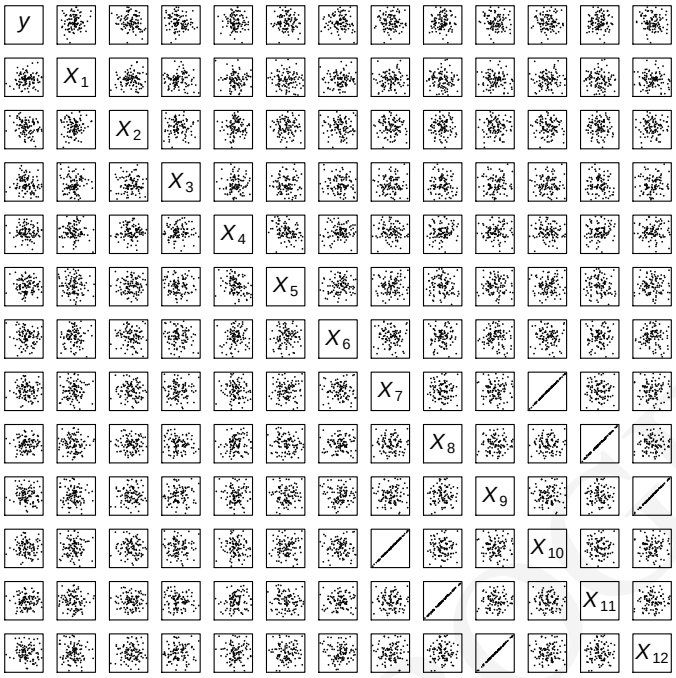

Figure 1: The correlations between the fixed X's, along with one generated y. We split $\mathbf{X}$ into two groups: the first group, $\mathbf{X}_{-q}$, is X_1 through X_{10}. The second group, $\mathbf{X}_q$, is X_{11} through X_{13}.

We compare the test performance of a fit on all the features $X_1, ..., X_{13}$, versus a fit on only features $X_1, ..., X_{10}$. Looking ahead to section 3, we also include the performance of the lasso and ridge regression (with regularization parameters λ chosen as the value that minimizes MSE on an additionally generated validation set of 20 observations), the performance of all-subset feature selection (again with the subset of features chosen as the subset that minimizes MSE on the same validation set), and a random forest of regression trees. As we can see from the figure (2), the true model has worse test performance than the underspecified model. But, again looking ahead to section 3, the random forest and regularizations perform even better than the underspecified model or all-subset regression! Also looking ahead to section 4, we can also ask if the relative model performances are different under mean absolute error (MAE), i.e. L_1 loss instead of L_2 loss, to ensure that the difference is not just an artifact of our choice of loss function; figure (3) shows that the result is the same (in section 4, we will see how the bias-variance tradeoff only breaks down for asymmetric loss functions), although the difference between the performance of the true model and the various 'false' models is less stark.

Wu et al. (2011) note that the MSE asymptotically has an F-distribution; but regardless of the choice of a t-test to test the means, a Wilcoxon rank-sum test, or an F-test, the difference between the MSE of the correctly specified model and underspecified model is significant, as is the difference between the underspecified model and the lasso or all-subset feature selection (although not between the lasso, ridge regression, or a random forest). As we increase n, the ordering remains the same (i.e., there indeed is a difference in the average test MSE), although the difference becomes no long significant.

We can also examine how the 'false' models are achieving their lower MSE; figure (4) shows how the underspecified model manages to regularize the estimates, underestimating the parameters but improving performance. We can also see the clear connection to Stein's paradox (Efron and Morris, 1977); the same (still counterintuitive) process is at work here as there, where individual coefficients regularized to be below their true value manage to, together, decrease the loss. Meanwhile, the OLS estimates are scattered. That the interval of all OLS estimates does not contain the true value of

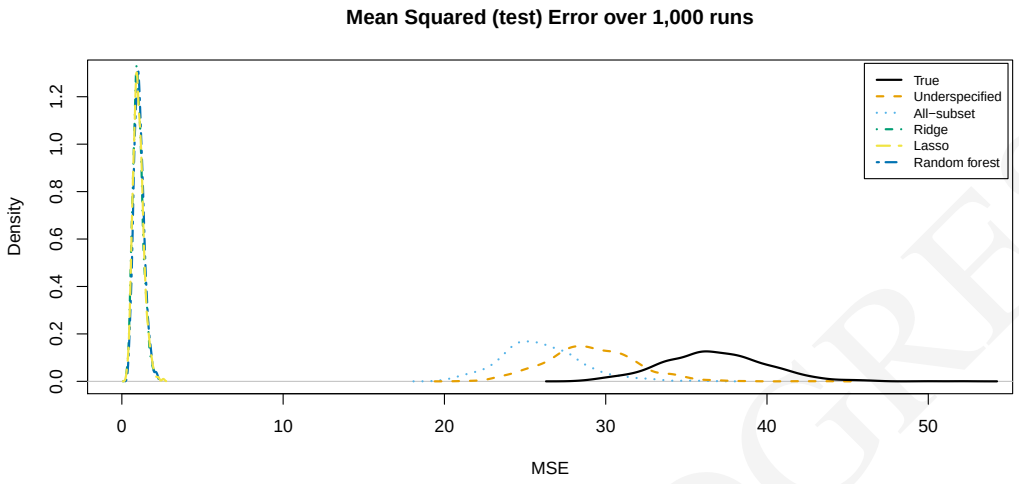

Figure 2: Comparison of average 5-fold test MSE over 1,000 runs for a correctly specified fit (true model), and the fits of various selection techniques.

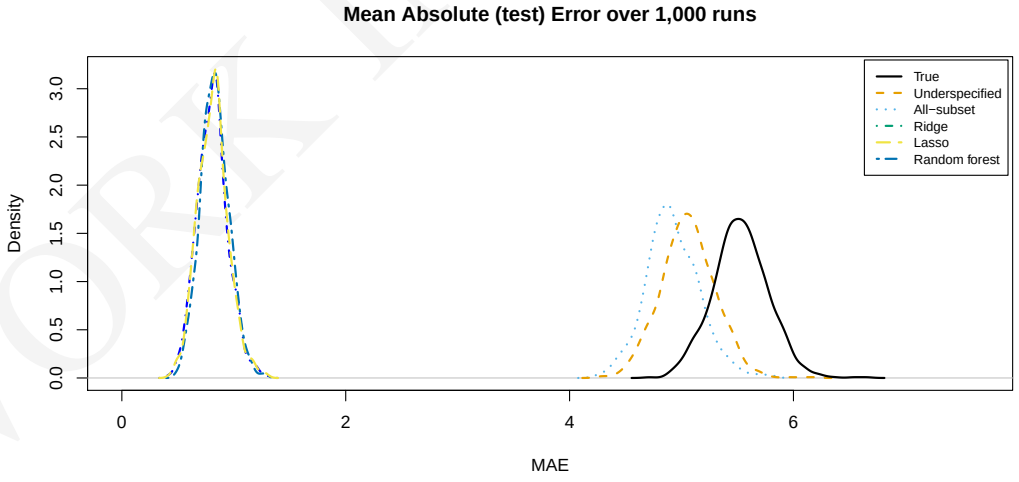

Figure 3: Comparison of average 5-fold test MAE over 1,000 runs for a correctly specified fit (true model), and the fits various selection techniques.

some coefficients is not to say that OLS is failing to be consistent (which would rather suggest that something is amiss in the simulation), only that $n_{\text{train}} = 80$ is too small for the asymptotic guarantees to kick in. The greatly inflated estimate of X_9 in both OLS and all-subset regression, and the estimate of X_{12} in both OLS and all-subset regression that is of large magnitude and with wrong sign (!), is likely due to the specific draw of $\mathbf{X}$ (since, within each group $\mathbf{X}_q$ and $\mathbf{X}_{-q}$, the variables are interchangeable).

However, how the lasso and ridge performances relate to their estimates is still mysterious; for both, only X_1 has a higher estimate than all the other coefficients (although still less than its true value). The random forest measures feature importance with GINI coefficients rather than in the same scale as the original coefficients, so we cannot compare the magnitudes, but we can look at the relative importance: unlike the lasso and ridge regression, the three pairs of correlated features have a slightly higher average importance than the other features, but otherwise there is no symmetry-breaking estimated coefficient like X_1 for the lasso and ridge regression.

The reader might also be interested to know, in terms of the traditional (although now fallen out of favor) goodness-of-fit of metrics of R^2 and adjusted R^2, it was .99 across the runs but coefficients were seldom significant. Note that the boxes in fig. 4 are not the same as confidence intervals of particular trials, but the two are similar, and the boxes that overlap with zero (the thin gray line) correspond to variables that were nonsignificant across most or all trials.

Of course, R^2 can be high but the features not significant when the irreducible noise of the response is extremely high (one reason its use has fallen out of favor among statisticians), as we have here.

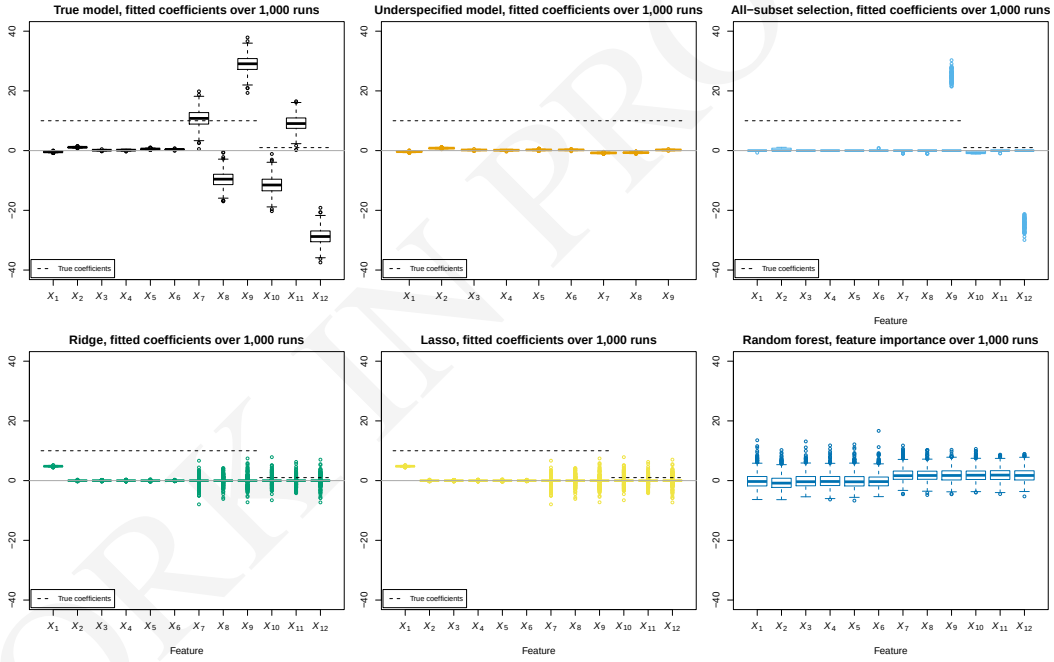

Figure 4: The size of the fitted coefficients over 1,000 runs for the various models. The dashed black line is the true coefficient size for the two groups of features. Colors correspond to those in figs. (2) and (3). We take selected-out features to have a coefficient of zero for plotting the boxplots of the lasso and all-subset regression.

3 Selection approaches

3.1 Irrepresentability

A prominent technique for selecting optimal variables is the lasso. However, the lasso's objective function is aimed at minimizing the MSE, and not actually finding the true variables. Zhao and Yu (2006) studied the lasso for its model selection consistency, and presented a condition they called

the **irrepresentable condition**, which is necessary and sufficient for consistency of lasso to pick out the true variables. Previously, Knight and Fu (2000) have shown lasso estimates are consistent for fixed $\boldsymbol{\beta}$ and p as $n \to \infty$. And Meinshausen and Bühlmann (2004) have shown that under certain conditions, the lasso is consistent in estimating the dependency between Gaussian variables even when number of variables grow faster than dimensions. Zhao and Yu (2006) argue that there exists two types of problems: first, whether there exists an optimal regularization λ which gives consistent selection, and second, for each random realization, whether there exists a regularization constant that selects the true model. The authors show that for both the problems, the irrepresentable condition is necessary and sufficient.

To define the irrepresentable condition, we need to first define the model and the settings.

1. Data is generated by linear model: $\mathbf{y}_n = \mathbf{X}_n\boldsymbol{\beta}^n + \boldsymbol{\varepsilon}_n$
2. The Lasso estimates are given by $\widehat{\boldsymbol{\beta}^n} = (\widehat{\beta}_1^n, \widehat{\beta}_2^n, \ldots, \widehat{\beta}_p^n)$ is defined by $\widehat{\boldsymbol{\beta}^n}(\lambda) = \text{argmin}_{\boldsymbol{\beta}} \, ||\mathbf{y}_n - \mathbf{X}_n\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1$.

Irrepresentable Conditions: Assume $\boldsymbol{\beta}^n = (\beta_1^n, \ldots, \beta_q^n, \beta_{q+1}^n, \ldots, \beta_p^n)$ where $\beta_j^n \neq 0$ for $j = 1, \ldots, q$ and $\beta_j^n = 0$ for $j = q+1, \ldots, p$. Let $\boldsymbol{\beta}_{(1)}^n = (\beta_1^n, \ldots, \beta_q^n)^T$ and $\boldsymbol{\beta}_{(2)}^n = (\beta_{q+1}^n, \ldots, \beta_p^n)^T$. Now, $\mathbf{X}_n(1)$ is the first q columns of $\mathbf{X}$ and $\mathbf{X}_n(2)$ is the last $p-q$ columns of $\mathbf{X}$. Let $C^n = \frac{1}{n}\mathbf{X}_n^T\mathbf{X}_n$. Similarly, $C_{11}^n = \frac{1}{n}\mathbf{X}_n(1)^T\mathbf{X}_n(1)$, $C_{22}^n = \frac{1}{n}\mathbf{X}_n(2)^T\mathbf{X}_n(2)$, $C_{12}^n = \frac{1}{n}\mathbf{X}_n(1)^T\mathbf{X}_n(2)$, and $C_{21}^n = \frac{1}{n}\mathbf{X}_n(2)^T\mathbf{X}_n(1)$.

C^n can therefore be expressed as:

$$C^n = \begin{pmatrix} C_{11}^n, C_{12}^n \\ C_{21}^n, C_{22}^n \end{pmatrix}$$

The **strong irrepresentable** condition is that there exists a positive constant vector η such that $|C_{21}^n(C_{11}^n)^{-1}\text{sign}(\boldsymbol{\beta}_{(1)}^n)| \leq \mathbf{1} - \eta$ where $\mathbf{1}$ is a $p-q$ vector of 1's.

The **weak irrepresentable** condition is that $|C_{21}^n(C_{11}^n)^{-1}\text{sign}(\boldsymbol{\beta}_{(1)}^n)| \leq \mathbf{1}$.

Under these conditions, Zhao and Yu (2006) shown that model selection is consistent. Additionally, they show that, if a non-important feature is highly correlated with any of the important features, the lasso will be unable to distinguish it from the true features, even with large amounts of data. Given that we deliberately designed the simulation above to have such correlations between important (large coefficient) and unimportant (small coefficient) features, it is unsurprising that the lasso frequently selected out some features (even though all of them were 'true' features that generated the observed data). However, the simulation also had σ that was very large compared to the coefficients (in particular, the coefficients of the unimportant features).

3.2 Post-selection inference

Currently, we have only talked about the distinction between two worlds: models whose purpose is *prediction* and models whose purpose is *explanation*. Model selection gives superior predictive performance, but as we see above, should not be used for information about the 'true' features of the DGP. But a promising new area, *post-selection inference* (Berk et al., 2013; Lee et al., 2016), provides us with a way to both do inference and prediction.

The classical way of doing statistical inference (using the standard errors of the model fitted with a given set of predictors) is invalid after model selection since the data selection process is itself stochastic, which would be ignored by naive application of usual inference methods. Post-selection inference is a way to get confidence intervals around parameter estimates that take into account the uncertainty introduced by model selection. The result is generally wider confidence intervals, but this is what we want; a feature may be selected into the model, but it may have a confidence interval wide enough to not be able to reject a null hypothesis that the true population parameter (or parameter of the underlying DGP) is different from zero. Here, the state of the art here is still emerging (with many innovations happening only in the past year), so we decided to focus on other topics. Still, post-selection inference deserves special mention as a place that is bringing together prediction and inference. Because of how fraught the process of inference is given the perniciousness

and insidiousness of omitted variable bias, and how predictions are far more reliable but generally unsatisfying for the scientific goals of understanding (and intervention, which requires knowledge of the DGP), post-selection inference may be a way to use predictions to study underlying phenomenon in more than an exploratory way.

4 Generalizing the bias-variance tradeoff

If we lack all the features/variables of the DGP, then we suffer from omitted variable bias; this can cause all our estimates to be biased anyway. Thus, we should not mind trading off being 'unbiased' for having lower variance, since here the 'unbiased' estimates of the given specification would be biased estimates of the true parameters of the DGP. However, in cases where we *do* have all the variables of the DGP, we have demonstrated above that the true model (for which an OLS fit will asymptotically give unbiased parameter estimates—although, as we saw above, there is sizable finite-sample bias) does not always have the lowest expected prediction error; this is obviously due to the bias-variance tradeoff. Certain underspecified, biased models (whether underspecified through a hand-picked feature subset, by the lasso, or by all-subset selection) manage a lower MSE by decreasing the variance.

However, how generally can we apply the understanding of truth and falsity in terms of the bias-variance tradeoff? The notion of the tradeoff comes from the decomposition of L_2 loss for an estimator for $Y = f(X) + \varepsilon$, with $\mathbb{E}[\varepsilon] = 0$ and $\mathrm{Var}[\varepsilon] = \sigma^2$,

$$\begin{aligned}
\mathrm{EPE}(x) &= \mathbb{E}\big[\big(Y - \widehat{f}(x)\big)^2 | X = x\big] \\
&= \mathrm{Var}(Y) + \mathbb{E}\big[\big(\widehat{f}(x) - f(x)\big)^2 | X = x\big] + \mathbb{E}\big[\big(\widehat{f}(x) - \mathbb{E}[\widehat{f}(x)]\big)^2 | X = x\big] \\
&= \sigma^2 + \mathrm{bias}^2\big(\widehat{f}(x)\big) + \mathrm{Var}\big(\widehat{f}(x)\big)
\end{aligned}$$

which consists of irreducible error (or, the variance of the response), the squared bias of the estimator, and the variance of the estimator. But this algebra is specific to the case of L_2 loss; if we use other loss functions, there might be a different decomposition (or not possible decomposition). Is our notion of the tradeoff a consequence of our choice of loss function?

The answer is, surprisingly, partially yes: for example, Friedman (1997) shows that increasing the variance of an estimator can actually decrease the error rate for 0-1 loss. However, speaking of bias and variance outside of L_2 loss requires generalizing the notions. James (2003) distinguishes *generalized variance* from a *variance effect* (VE), and *generalized bias* from a *systematic effect* (SE). He defines a 'systematic' operator S,

$$S\widehat{f} = \mathrm{argmin}_{\mu}\, \mathbb{E}_{\widehat{f}}\big(\widehat{f} - \mu\big)$$

which acts on the distribution of the estimator $\widehat{f}$ to extract the 'systematic' component of it. For L_2 loss, $S\widehat{f} = \mathbb{E}[\widehat{f}]$; for L_1 loss, this turns out to be the median. This operator leads to three possible generalizations of variance, which are equivalent under L_2 loss but not necessarily for other loss functions $L(\cdot, \cdot)$:

$$\mathrm{Var}(\widehat{f}) = \mathbb{E}_{\widehat{f}}[L(\widehat{f}, S\widehat{f})] \tag{1}$$

$$\mathrm{Var}(\widehat{f}) = \mathbb{E}_{\widehat{f}}[L(SY, \widehat{f})] - L(SY, S\widehat{f}) \tag{2}$$

$$\mathrm{Var}(\widehat{f}) = \mathbb{E}_{Y,\widehat{f}}[L(Y, \widehat{f})] - \mathbb{E}_{Y}[L(Y, S\widehat{f})] \tag{3}$$

James (2003) argues that a generalization of the variance of an estimator $\widehat{f}$ should, among other properties, depend only on the estimator, and not the response variable Y. The only potential generalization that fulfills this is (1). Similarly, a generalization of bias squared should measure the difference between the systematic parts of the estimator, $S\widehat{f}$, and the systematic parts of the response, SY. Again, from several possibilities, only $\mathrm{bias}^2(Y, \widehat{f}) = L(SY, S\widehat{f})$ fulfills this.

Separately, James (2003) defines the effect of variance on the prediction error as the *variance effect*, and similarly the effect of the bias on the prediction error as the *systematic effect*, noting that "For example, in general, it is possible to have an estimator with high variance but for this variance to

have little impact on the prediction error... It is even possible for increased variance to decrease the prediction error". For L_2 loss (and, James proves, for strictly convex loss functions), generalized variance is equal to the variance effect and generalized bias is equal to the systematic effect, but not so for 0-1 and other (non-strictly convex) loss functions. James (2003) defines the VE and SE respectively as

$$\begin{aligned}\mathrm{VE}(Y,\widehat{f}) &= \mathbb{E}_{\widehat{f},Y}[L(Y,\widehat{f}) - L(Y,S\widehat{f})] \\ \mathrm{SE}(Y,\widehat{f}) &= \mathbb{E}_{Y}[L(Y,S\widehat{f}) - L(Y,SY)]\end{aligned}$$

VE expresses the difference in loss between a zero-variance, i.e. constant, estimator, and SE expresses the difference in loss between using the systematic component of the predictor and the systematic component of the estimator to make predictions (e.g., for L_2 loss, comparing the predictions from using the mean of Y to predictions made using the mean of $\widehat{f}$).

Consider a classification problem with 0-1 loss, with class probabilities $\pi_1, ..., \pi_k$ and predicted class probabilities $\widehat{\pi}_1, ..., \widehat{\pi}_k$ from a classifier $\widehat{f}$. SY is the majority class, $S\widehat{f}$ is the class predicted to be the majority, and

$$\begin{aligned}\mathrm{Var}(Y) &= \mathbb{E}_Y[L(Y,SY)] = \mathbb{E}_Y[I(Y = SY)] = \mathbb{P}(Y \neq SY) = 1 - \max_i \ \pi_i \\ \mathrm{Var}(\widehat{f}) &= \mathbb{E}_{\widehat{f}}[L(\widehat{f},S\widehat{f})] = \mathbb{E}_{\widehat{f}}[I(\widehat{f} = S\widehat{f})] = \mathbb{P}(\widehat{f} \neq S\widehat{f}) = 1 - \max_i \ \widehat{\pi}_i \\ \mathrm{bias}(Y,\widehat{f}) &= I(SY \neq S\widehat{f})\end{aligned}$$

That is, the variance of a multinomial response is 1 minus the probability of being the majority class, the variance of the classifier (the estimator) is 1 minus the probability of being the class predicted to be most likely, and the bias is 0 if the majority class of the response and the classifier agree and 1 otherwise.

$$\begin{aligned}\mathrm{VE}(Y,\widehat{f}) &= \mathbb{E}_{\widehat{f},Y}[L(Y,\widehat{f}) - L(Y,S\widehat{f})] = \mathbb{E}_{\widehat{f},Y}[I(Y \neq \widehat{f}) - I(Y \neq S\widehat{f})] \\ &= \mathbb{P}(Y \neq \widehat{f}) - \mathbb{P}(Y \neq S\widehat{f}) = \left(1 - \textstyle\sum_{i=1}^k \pi_i\widehat{\pi}_i\right) - \left(1 - \pi_{\mathrm{argmax}_i\,\widehat{\pi}_i}\right) \\ &= \pi_{\mathrm{argmax}_i\,\widehat{\pi}_i} - \textstyle\sum_{i=1}^k \pi_i\widehat{\pi}_i \\ \mathrm{SE}(Y,\widehat{f}) &= \mathbb{E}_Y[L(Y,S\widehat{f}) - L(Y,SY)] = \mathbb{E}_Y[I(Y \neq S\widehat{f}) - I(Y \neq SY)] \\ &= \mathbb{P}(Y \neq S\widehat{f}) - \mathbb{P}(Y \neq SY) = \left(1 - \pi_{\mathrm{argmax}_i\,\widehat{\pi}_i}\right) - \left(1 - \max_i \ \pi_i\right) \\ &= \max_i \ \pi_i - \pi_{\mathrm{argmax}_i\,\widehat{\pi}_i}\end{aligned}$$

The VE and SE are here clearly not respectively equal to the variance or the bias squared. James (2003) then gives the following numerical example: at fixed input $X = x$, let $\mathbb{P}(Y|X = x) = (0.5, 0.4, 0.1)$, and consider classifiers $\widehat{f}_1$ and $\widehat{f}_2$ with respective distributions over training samples $\mathbb{P}(\widehat{f}_1(x)|X = x) = (0.4, 0.5, 0.1)$ and $\mathbb{P}(\widehat{f}_2(x)|X = x) = (0.1, 0.5, 0.4)$.[2] They both have the same distribution of probabilities, so they have the same variance, which comes out to 0.5. Both have a 'bias' of 1 by the definition of generalized bias, since both give a plurality of probability mass to the nonmajority class. Both also have the same systematic error, 0.1. However,

$$\begin{aligned}\mathrm{VE}(\widehat{f}_1,Y) &= 0.4 - (0.5 \cdot 0.4 + 0.4 \cdot 0.5 + 0.1 \cdot 0.1) = .40 - 0.41 = -0.01 \\ \mathrm{VE}(\widehat{f}_2,Y) &= 0.4 - (0.5 \cdot 0.1 + 0.4 \cdot 0.5 + 0.1 \cdot 0.4) = .40 - 0.29 = 0.11\end{aligned}$$

[2]Note that many simple classifiers, at a fixed input $X = x$, would deterministically give the same output: something like $\mathbb{P}(\widehat{f}(x)|X = x) = (1, 0, 0)$, which would be the Bayes classifier at this level of $X = x$. The classifiers considered here have some randomness for a fixed level of the features. The reason why a classifier would assign incorrect probabilities (whether deterministically assigning 1 to the wrong class, or being random but with incorrect class probabilities) at a fixed input is that the relationship between the features and the class labels might be very spiky and the estimator too smooth to be able to capture these spikes. In this case, the estimator could do poorly at certain levels of the features for the sake of performing better overall.

Thus, James (2003) concludes, even though $\widehat{f}_1$ and $\widehat{f}_2$ have the same *amount* of variance, the variance of $\widehat{f}_1$ has caused the error rate to decrease while the variance of $\widehat{f}_2$ has caused the error rate to increase![3] The reason is that the variance of $\widehat{f}_1$ results in more classifications being made to the true majority class.

While a classifier that predicted the majority class at a fixed input $X = x$ (or a classifier that captured the true probabilities over fixed $X = x$) would have even lower prediction error than $\widehat{f}_1$ or $\widehat{f}_2$ and hence this is not an example of a false model predicting better than a real one, this toy example does demonstrate how actions that increase the variance of an estimator without affecting the 'bias' may decrease the prediction error. This means that, outside of strictly convex loss functions, the ways in which the bias-variance tradeoff explains the difference between 'true' models and ones that predict well are not straightforward.

This puts a damper on the hope that we might find parsimonious and predictive models (Forster and Sober, 1994) outside of the curve-fitting or L_2 case. But this provides a way in which we might understand how classifiers can improve predictive performance by complex, rather than simple, relationships to 'true' underlying trends: in the example above, having true most frequent class be the second-most-frequently predicted class resulted in the same *increase* in variance as another random shuffling of predictions, but this variance decreased rather than increased the prediction error.

5 Lessons and future work

By taking a case of an underspecified linear model, we were able to create a narrative through the paradoxical disconnect between prediction and explanation. We saw, and concretely demonstrated in a simulation, how large noise and correlated variables lead to true models predicting better than false models; and furthermore, that model selection procedures similarly output false models but ones that predict even better than the specific form of underspecification that we theoretically explored. Second, we explored further into a condition that helps explain the performance of the lasso in this case, the irrepresentable condition, which both explains how the true features of the DGP will not necessarily be the ones that lead to the lowest error but that this is a separate issue from achieving the oracle prediction rate.

At the heart of the paradox as presented is the bias-variance tradeoff; false models can achieve better performance by lowering the variance. However, the tradeoff cannot fully explain how false models relate to true ones. Generalizations of the bias-variance tradeoff shows that for classification specifically and non-strictly convex loss functions in general, the prediction error can be reduced by actually increasing variance, showing the complexity of the connection between prediction error and model truth.

Fortunately, post-selection inference gives a promising start to statistical investigations into ways to combine prediction and explanation. Especially as we deal with larger and more complex phenomena, far beyond where it feasible to have theory-driven model specifications, understanding and resolving the difference between prediction and explanation will be a central challenge for future scientific progress.

The simulation we introduce here, providing an explicit case where 'false' models fit better than the 'true' one, also gives us opportunities to apply other techniques. Mainly, this is ideal for also testing and comparing post-inference selection, as well as methods of estimating undirected graphical models and of causal discovery. These are relatively brief and straightforward extensions; further extensions would be to vary some of the simulation conditions. If we broaden the scope slightly and consider problems such as omitted variable bias and dependencies that also contribute to models being wrong, we can examine how techniques built to achieve the best model performance interact with 'truth' in such cases that appear frequently.

[3] Again, consider a classifier that is deterministic at fixed inputs; no matter which class it predicts to be the majority class, $\mathrm{VE}(\widehat{f}_2, Y) = \pi_{\mathrm{argmax}_i\, \widehat{\pi}_i} - (1 \cdot \pi_{\mathrm{argmax}_i\, \widehat{\pi}_i} + 0 + 0) = 0$ and there is no variance to contribute to the error rate; so obviously, in the case where the incorrect class is chosen at $X = x$, adding some randomization to the outputs can improve the performance of the classifier. James (2003) gives the example of boosting as a case of this.

Another extension would be to pursue an example of Mullainathan and Spiess (2017), who partition the American Housing Survey into subsets and apply the lasso separately to each of them. They find that the predictive performance is similar across the partitions, but the variables that are selected into the model vary wildly, which is an excellent real-world illustration of why we should not rely on the lasso for interpreting which variables are 'important' for the process in question. Extending this work to trying out other feature selection techniques on the American Housing Survey or a similar large social data set would help give intuition of how techniques other than the lasso are acting. And, for being able to compare fitted model outputs to known true coefficients or feature importance, we could repeat the exercise on data simulated from fitted models.

The distinction between prediction and explanation, and the way in which the bias-variance tradeoff operates, is far from intuitive or obvious. There remains much more work to be done, and a need for further simulation demonstrations to complement precise but abstract theoretical results, to better develop understandings of how predictive modeling works and how statistical education can best educate practitioners and consumers of statistical knowledge about how to interpret models.

References

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483–485.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837.

Box, G. E. P. (1979). Robustness in the strategy of scientific model building. Technical Report #1954, University of Madison-Wisconsin, Mathematics Research Center.

Breiman, L. (2001). Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231.

Efron, B. and Morris, C. (1977). Stein's paradox in statistics. *Scientific American*, 236(5):119–127.

Forster, M. and Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science*, 45(1):1–35.

Friedman, J. H. (1997). On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 1(1):55–77.

Hindman, M. (2015). Building better models: Prediction, replication, and machine learning in the social sciences. *The ANNALS of the American Academy of Political and Social Science*, 659(1):48–62.

Hofman, J. M., Sharma, A., and Watts, D. J. (2017). Prediction and explanation in social systems. *Science*, 355(6324):486–488.

James, G. M. (2003). Variance and bias for general loss functions. *Machine Learning*, 51(2):115–135.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Obermeyer, Z. (2015). Prediction policy problems. *American Economic Review*, 105(5):491–95.

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics*, 28(5):1356–1378.

Kunst, R. M. (2008). Cross validation of prediction models for seasonal time series by parametric bootstrapping. *Austrian Journal of Statistics*, 37(3&4):271–284.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927.

Lin, J. (2015). On building better mousetraps and understanding the human condition: Reflections on big data in the social sciences. *The ANNALS of the American Academy of Political and Social Science*, 659(1):33–47.

Meinshausen, N. and Bühlmann, P. (2004). Consistent neighbourhood selection for sparse high-dimensional graphs with the lasso. Technical report, Seminar für Statistik, Eidgenössische Technische Hochschule (ETH) Zürich.

Mullainathan, S. and Spiess, J. (2017). Machine learning: An applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106.

Rolling, C. A. and Yang, Y. (2014). Model selection for estimating treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(4):749–769.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3):289–310.

Wu, S., Harris, T. J., and McAuley, K. B. (2007). The use of simplified or misspecified models: Linear case. *The Canadian Journal of Chemical Engineering*, 85(4):386–398.

Wu, S., McAuley, K. B., and Harris, T. J. (2011). Selection of simplified models: II. Development of a model selection criterion based on mean squared error. *The Canadian Journal of Chemical Engineering*, 89(2):325–336.

Yang, W. and Yang, Y. (2016). Toward an objective and reproducible model choice via variable selection deviation. *Biometrics*, 73(1):20–30.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563.