1 To Do

Link back to earlier work in finding that CVSS and AIC were not pointing for us to pick the same model.

2 Motivation

Problem is motivated by selecting features for a regression using features derived from the singular value decomposition of the document/term matrix of text documents. The response is a property associated with a document (in our case, the price of a house described by a listing) and the features are eigenwords from the text of a collection of listings. Rather than try to find the best subset, the monotone nature of the predictive value of these features suggests picking in order, much like one does with an autoregression (albeit on a larger scale). Should we use the first 10, 30, or 50 singular vectors (*i.e.*, eigenwords) as features? The problem is complicated by the presence of outliers produced by the decomposition. Given the Ziphian distribution associated with text, outliers are to be expected.

More generally, leveraged observations produce surprising effects on regression models, particularly by producing collinearity among the explanatory features. Leveraged observations in our setting are simply rows of the design matrix X that have large variance compared to typical rows.

3 Stylized Problem

The following example illustrates what can happen. For this example, we simulated $n = 1000$ observations $X_i = (X_{i1}, \ldots, X_{ip})'$ of $p = 100$ independent Gaussian variables. The first n_o

obsevations have variance σ_o^2, and the remaining observations have variance 1. otherwise.

$$X_{ij} \sim \begin{cases} N(0,\sigma_o^2), & i \le n_o \\ N(0,1), & n_o < i \le n \ . \end{cases} \tag{1}$$

These first 50 cases are the leveraged observations; the observations are independent, but not identically distributed. The model for the response is the usual Gaussian regression with *constant* error variance,

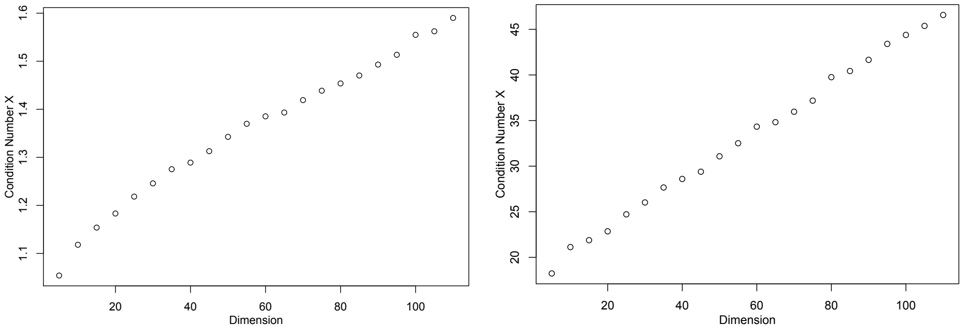
$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \sigma\epsilon_i, \quad \epsilon_i \sim N(0,1) \ . \tag{2}$$

Notice that the model is *not* heteroscedastic; rather, the changing variances occur among the values of the X_{ij}.

Now suppose that we know that only the first k elements of β are non-zero. How should we pick the best fitting regression, that minimizes the squared error of prediction? As will become clear, this question is not well posed. A popular choice for this context is to use *AIC*, which we illustrate in a small simulation. Suppose $n = 2000$ with $p = 150$, with $n_o = 50$ leverage points with $\sigma_o = 100$. Before continuing, we note that the presence of these leverage points produces a surprising amount of collinearity given that the design points are independent. Figure 1 graphs the condition number of the leading columns of the design matrix, varying the number of columns included. (The condition number is the ratio of the largest to smallest singular value of a non-square matrix.) The left frame shows the condition number if $\sigma_o = 1$ (*i.e.*, without leverage points); the condition number grows roughly linearly in the number of columns (slope ≈ 0.0046). The right frame of Figure 1 shows the condition number in the context of our simulated data with $n_o = 50$ and $\sigma_o = 100$. Again, the condition number increases linearly, but with a much steeper slope near 0.26. But for the scale of the y-axis, the figures are almost identical. Notice that the linear growth persists for matrices with more than n_o columns; larger matrices do not quickly "outgrow" the influence of the leverage points. The amount of collinearity grows at a steady fixed rate (for these dimensions) as the number of columns increases.

Figure 2 shows the impact of this colllinearity on the problem of picking a model. For this illustration, $k = 75$ features are predictive. We set $\beta_j = c$ for $j = 1, \ldots, k$ with c chosen

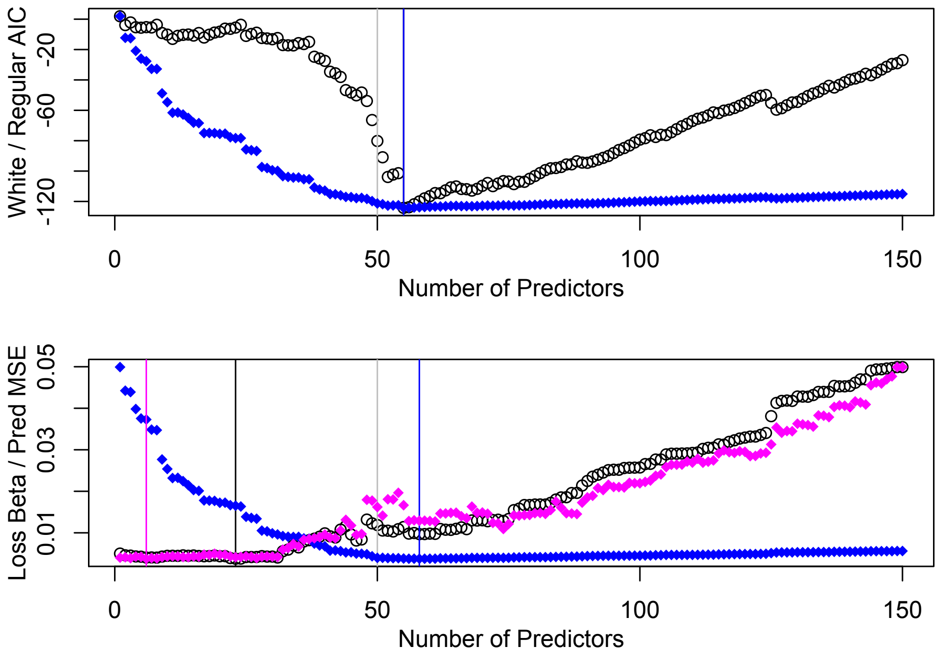
Figure 1: Condition numbers for matrices with increasing numbers of columns, with with no leveraged outliers (left) or with $n_o = 50$ leveraged outliers with standard deviation $\sigma_o = 100$ (right).



so that these coefficients lie, on average about 5 standard errors from 0. These are very significant effects, and $R^2 \approx 0.53$ for the fully specified model with all 75 features. The top frame of Figure 2 plots of two forms of AIC for models of increasing size. The White version of AIC uses a heteroscedastic consistent variance estimate rather than the fitted estimate. (We also use the estimate of σ^2 from the prior model when computing AIC.) The version of AIC corrected for heteroscedasticity shows a dramatic valley and relatively steep increase past its minimum. Though showing different trends, the two versions of AIC both pick a model with about 50 features, substantially fewer than $k = 75$ and matching the number of leverage points. In both panels of Figure 2, we have normalized the statistics to lie on a common range for visual comparisons. The vertical gray line in this and other figures highlights the number of leveraged outliers in the data (in this case, $n_o = 50$). Other colored colored segments denote the position of the minimum value.

The lower panel shows how well these models perform. Whereas both versions of AIC are computable; the lower panel shows unobservable errors that are computable only because we have simulated these data from known populations. The black points in the figure show the model loss. Let $X_{0:j}$ refer to a matrix comprised of a column of 1's followed by the first j columns of X, and similarly let $\beta_{0:j}$ and $\hat{\beta}_{0:j}$ refer to the associated true coefficients and least squares estimates (including the intercept). The model loss (black points) is then the

Figure 2: Simulation of AIC (normal in blue and heteroscedastic consistent, top) and model loss and mean squared prediction errors (bottom), in the presence of leverage points. The loss is shown in black, with the “in-sample” error $\|X_{1:j}(\hat{\beta}_{1:j} - \beta_{1:j}\|^2$ and the corresponding “out-of-sample” error when the estimates are applied to new observations.



sequence

$$\|\beta_{0:j} - \hat{\beta}_{0:j}\|^2 \ , \quad j = 1, \ldots, 150 \ .$$

The sequence of blue points in the figure are the "in-sample" mean squared errors, computed as

$$\text{In-Sample MSE: } \|X_{0:j}(\hat{\beta}_{0:j} - \beta_{0:j})\|^2$$

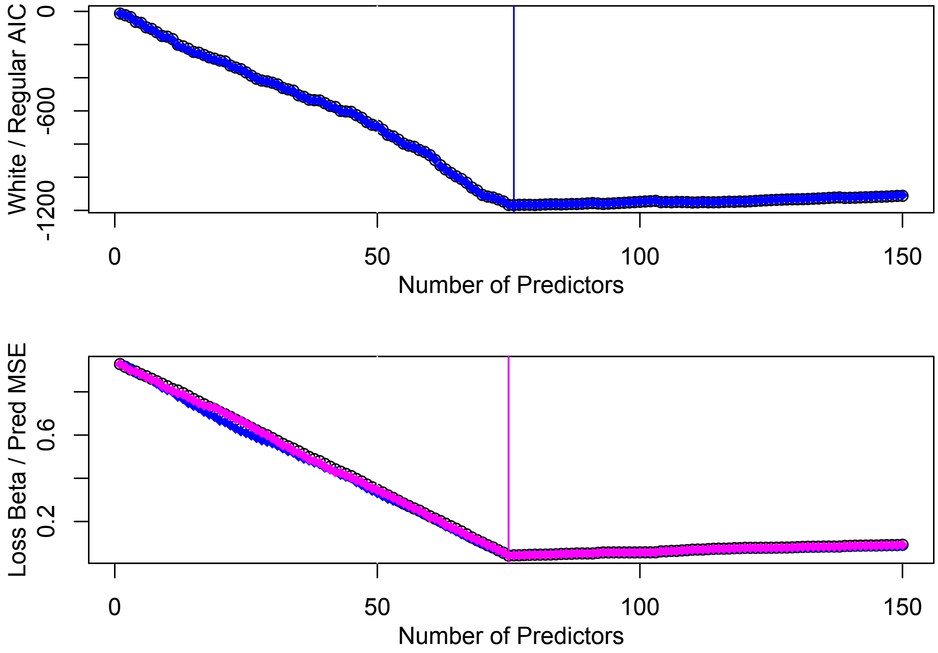
This sequence is highly correlated with the trend of the usual *AIC* statistic in the upper panel. The "out-of-sample" version of the mean squared error is computed by drawing a second realization of the design matrix from the same process (and the same choices for simulating leverage points), independently, and computing the prediction error for these features. With this independent realization of X labeled $\widetilde{X}$, the magenta points in the lower panel of Figure 2 show

$$\text{Out-of-Sample MSE: } \|\widetilde{X}_{0:j}(\hat{\beta}_{0:j} - \beta_{0:j})\|^2$$

This version of the mean squared error closely mimics the trend in the loss, which after all is its expectation (over the distribution of X).

In the presence of leverage points, *AIC* thus tries to pick a model that predicts *the observed* cases well. After all, that's the only data visible. What we more often experience in practice, however, is the behavior seen of the out-of-sample MSE. Thus, *AIC* is solving one problem (evidently pretty well), but it may not be the problem we care about. Without the leverage points (and the collinearity they induce), the issues raised here vanish. Figure 3 shows the same statistics and random variables as Figure 2, but wihout the leverage points. Both versions of *AIC* align closely and pick the correct model with $k = 75$ features, and all of the error measures agree.

Figure 3: Simulation of AIC (normal in blue and heteroscedastic consistent, top) and model loss and mean squared prediction errors (bottom), without leverage points.



4 xx

5 xx

6 Derivations