1 To Do

- Basically a regression with 50 X's that only uses the other data when collinearity becomes very severe.
- Breiman citations
- Link back to earlier work in which CVSS and AIC were pointing to different models. The results here might be a good explanation of that phenomenon.
- Figure out what it means to have 'comparable' situations among IID, leveraged cases, and autoregressive models: plan is to (a) set condition number (b) vary coefficients to obtain specified/target R2.
- Decide how to handle the intercept. It mucks up the comparison analysis later since $\mathbb{E}\, X'X$ has that non-stochastic leading n.

2 Motivation

This problem is motivated by selecting features for a regression in text analysis. The features for this regression are derived from the singular value decomposition of the document/term matrix of a large collection of text documents. The value of the response is a characteristic of the subject of the document. In our application, documents are real-estate listings that verbally describe properties. The response is the listed price of the home. The regression features are eigenwords from the text of a collection of listings, singular vectors from the document/term matrix. Rather than try to find the best subset, the monotone predictive value of these features suggests picking them 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 (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 a form of 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. Let $(X_{i1}, \ldots, X_{ip})'$ for $i = 1, \ldots, n$ denote a vector of p independent Gaussian variables. These vectors form the rows of the design matrix X in a linear regression. To produce the leverage effects observed in text regressions, we assume the first n_o observations have variance σ_o^2, and the remaining observations have variance 1:

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

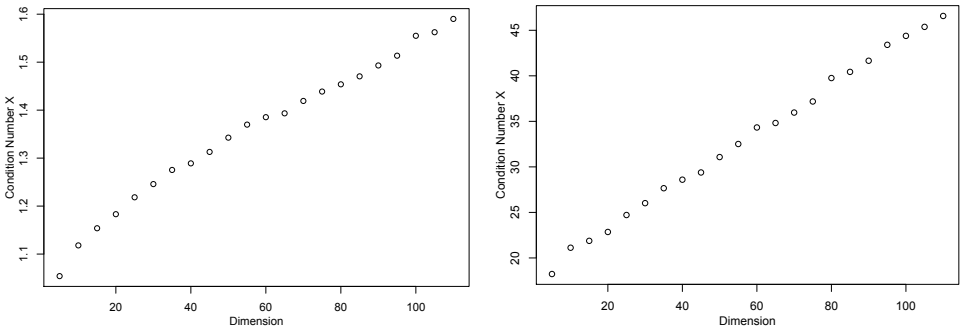
The observations are independent, but not identically distributed. The model for the response is the usual Gaussian regression with *constant* error variance. If we let X_j with a single subscript denote the jth column of the $n \times p$ matrix X with elementsX_{ij}, then the regression model is

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \sigma\epsilon, \quad \epsilon_i \sim N(0,1) \ . \tag{2}$$

Notice that the model is *not* heteroscedastic; rather, the changing variances occur among the rows of X.

Now suppose that we know that only the first k elements of β are non-zero. How should we pick the best fitting regression, one that minimizes the squared error of prediction? As will become clear, this question is not well-posed. A popular choice for this context is pick for k the model that minimizes AIC, which we illustrate with a simulated example. For this example, we observe $n = 2000$ cases of $p = 150$ available features, with $n_o = 50$ leverage points with standard deviation $\sigma_o = 100$. Before continuing to the response, we note that the presence of these leverage points produces a surprising amount of collinearity given that the

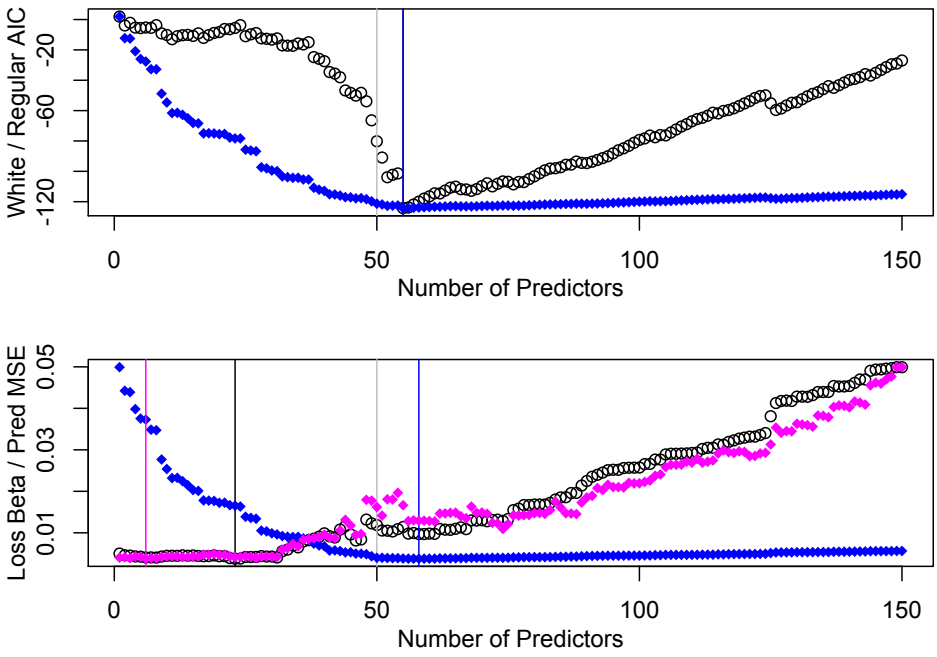
Figure 1: Condition numbers for matrices with increasing numbers of columns, with with no leveraged outliers (left) or with $n_o = 50$ leveraged outliers with standard deviation $\sigma_o = 100$ (right).



design points are independent. Figure 1 graphs the condition number of the leading columns of X, 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. The linear growth in the condition number 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 these leverage points on the problem of picking a model. For this illustration, $k = 75$ features are predictive (*i.e.*, have non-zero regression coefficient). We set the error variance to $\sigma = 1$ and $\beta_j = c$ for $j = 1, \ldots, k$, with c chosen so that these coefficients lie, on average, about 5 standard errors from 0 when fitting the model with all 75 features included. As such, these features are very significant effects, and $R^2 \approx 0.53$ for the fully specified model that uses $X_1, \ldots, X_{75}$. The top frame of Figure 2 plots two versions of *AIC* for models of increasing size. The White version of *AIC* uses a heteroscedastic consistent estimate of the variance of $\hat{\beta}$ rather than the usual expression. (We also use the estimate

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 fixed-design mean squared error (blue, equation 4) and the corresponding random-design error (magenta, equation 5).



of σ^2 from the prior model with $j-1$ features when computing AIC for the model with j features.) Because these do not lie in the same range, we rescaled both AIC statistics to lie on a common range for visual comparisons. 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 pick a model with about 50 features, substantially fewer than $k = 75$ and instead matching the number of leverage points. 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 vertical segments denote the position of the minimum value of each series, identifying the "best" model as judged by each metric.

The lower panel of Figure 2 shows how well each model in this sequence of fits performs. Whereas both versions of AIC are computable; the lower panel graphs unobservable random variables that are computable only because we simulated these data from known models.

The black points in the figure show the loss when estimating β. 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 $\hat{\beta}_{0:j}$ denote the corresponding true coefficients and least squares estimates (including the intercept). The model loss (black points) is then the sequence

$$\|\beta_{0:j} - \hat{\beta}_{0:j}\|^2 \,, \quad j = 1, \ldots, 150 \,. \tag{3}$$

The sequence of blue points in the figure are the mean squared errors under a fixed-X design, computed as

$$\text{Fixed-Design MSE: } \|X_{0:j}(\hat{\beta}_{0:j} - \beta_{0:j})\|^2 \tag{4}$$

Again, for the sake of comparison, the lower panel of Figure 2 graphs the series using a common y-axis. The fixed-design MSE presumes that were we to predict new data, the new observations would have the same values of the observed model features. The sequence of fixed-design MSEs is highly correlated with the trend of the usual AIC statistic shown in the upper panel of Figure 2. In contrast, the random-design mean squared error shown in magenta in Figure 2 measures the accuracy in the more realistic context in which we predict new data from the same population, but not at the same values of the explanatory variables. This MSE is computed by independently simulating a second realization of the design matrix from the same process, including the same choices for simulating leverage points, and computing the prediction error obtained at these values of the model features. If we denote an independent realization of X by $\widetilde{X}$, then the magenta points in the lower panel of Figure 2 show

$$\text{Random Design MSE: } \|\widetilde{X}_{0:j}(\hat{\beta}_{0:j} - \beta_{0:j})\|^2 \tag{5}$$

The random design MSE closely mimics the trend in the loss, which is proportional to its expectation (over the distribution of X):

$$\begin{aligned}
\mathbb{E}\,\|\widetilde{X}_{0:j}(\hat{\beta}_{0:j} - \beta_{0:j})\|^2 &= \mathbb{E}\,(\hat{\beta}_{0:j} - \beta_{0:j})'\widetilde{X}'_{0:j}\widetilde{X}_{0:j}(\hat{\beta}_{0:j} - \beta_{0:j}) \\
&= \mathbb{E}\,\mathrm{tr}\widetilde{X}'_{0:j}\widetilde{X}_{0:j}(\hat{\beta}_{0:j} - \beta_{0:j})(\hat{\beta}_{0:j} - \beta_{0:j})' \\
&= \mathrm{tr}\mathbb{E}\,\widetilde{X}'_{0:j}\widetilde{X}_{0:j}(\hat{\beta}_{0:j} - \beta_{0:j})(\hat{\beta}_{0:j} - \beta_{0:j})' \\
&= \mathrm{tr}\mathbb{E}\,(\widetilde{X}'_{0:j}\widetilde{X}_{0:j})\mathbb{E}\,((\hat{\beta}_{0:j} - \beta_{0:j})(\hat{\beta}_{0:j} - \beta_{0:j})') \\
&= \mathrm{tr}D_j(n,c)\mathbb{E}\,((\hat{\beta}_{0:j} - \beta_{0:j})(\hat{\beta}_{0:j} - \beta_{0:j})')
\end{aligned}$$

$$\approx \quad c\,\mathbb{E}\,(\hat{\beta}_{0:j} - \beta_{0:j})'(\hat{\beta}_{0:j} - \beta_{0:j})\;, \tag{6}$$

where $D_j(n, c)$ is a diagonal matrix with leading diagonal element n followed by j copies of the constant $c = n_o\sigma_o^2 + (n - n_o)$. The approximation results from replacing n by c in this leading diagonal matrix.

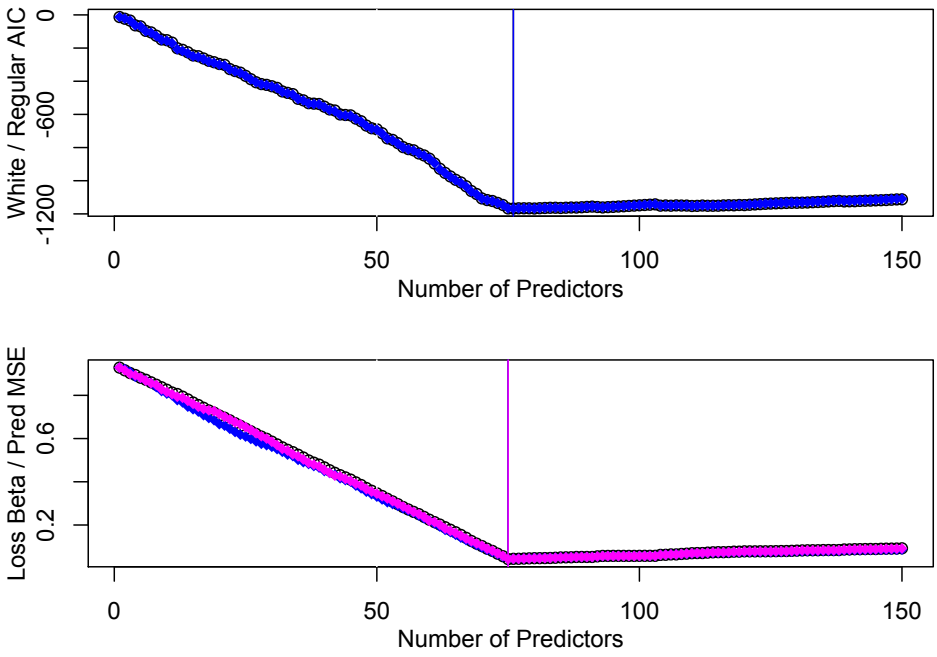
In the presence of leverage points, *AIC* thus picks the model that predicts well at the *observed* design points. After all, that's the only data visible. What we more often experience in practice, however, is the behavior of the random-design MSE. Out-of-sample leverage points possess a very different configuration than those observed in-sample, and the model cannot predict this new configuration so well as that which was observed. In this sense, *AIC* answers one question (evidently pretty well), but it may not be the question we need to answer.

These problems vanish without the leverage points, even in common models used to introduce collinearity among the features. Figure 3 shows the same statistics and random variables as Figure 2, but for "ideal" data that lack leverage points and the resulting collinearities. Again, $b = 2000$ with $p = 150$ possible features, of which the first 75 have non-zero coefficients. Both versions of *AIC* align closely and pick the correct model with 75 features, and all of the error measures in the second panel agree.

{**ras:** *The following results are dodgy: depends a lot on how you make the data.* }

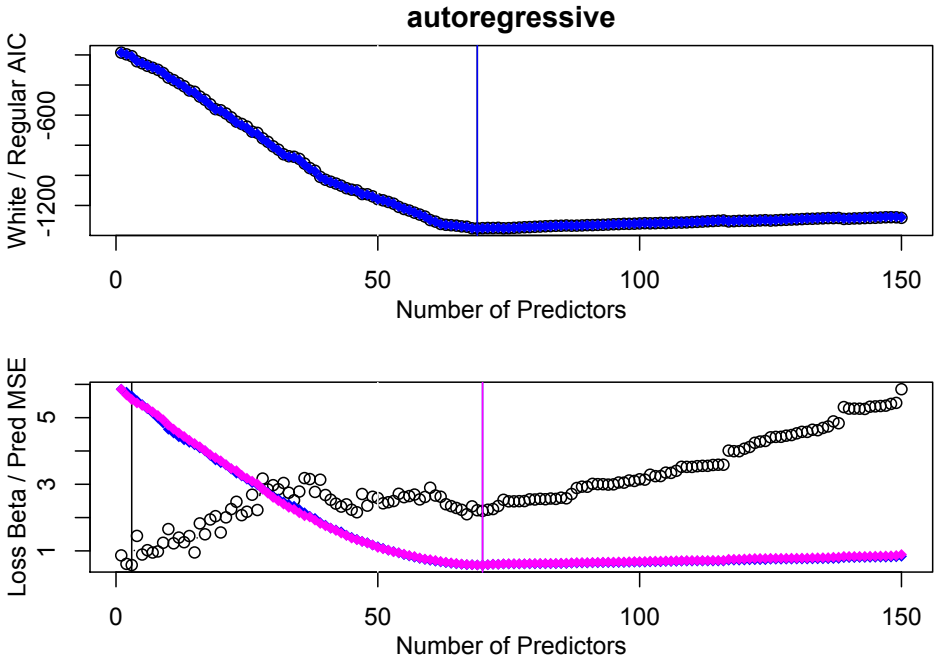
An obvious question to pose at this point is to ask: How do these plots look under other forms of collinearity? For example, a common model used to generate correlated predictors is the first-order autoregression, in which $\mathrm{Corr}(X_j, X_k) = \phi^{|j-k|}$ for $|\phi| < 1$. Figure 4 shows the two versions of *AIC* and mean squared errors as in Figures 2 and 3, but with autoregressive dependence among the columns of X. Specifically, we set $\phi = 0.935$ so that the condition number of X approximates that produced by the leveraged outliers in the first example. Although the condition numbers roughly match, the two collinear design matrices produce different distributions of singular values, as seen in Figure 5. The $n_o = 50$ leverage points produce 50 large singular values, whereas autoregressive collinearity produces

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



a geometrically decaying collection. (Notice also that adding more columns to the design increases the collinearity produced by leverage, but not for autoregressive dependence. In the AR model with $s < t$, X_s is independent of X_t given X_{s+1}. As as result, leverage produced much higher VIFs for the full set of $p = 150$ features in the range of 150 to 200 in this example, compared to 25-30 for the autoregressive model.) In this setting, the performance of the two versions of AIC shown in Figure 4 closely agree, and both pick the correct model with $k = 75$ features. The sequences of fixed- and random-design mean squared errors are also very similar with minimum values near $k = 75$. The squared loss estimating β is not proportional to the fixed-design MSE in this case because $\mathbb{E}\, X'X \neq c\, I$ – the columns of X are not independent in expectation. As a result, $\mathbb{E}\,\|\hat{\beta} - \beta\|^2$ is rather different than $\mathbb{E}\,\|X(\hat{\beta} - beta)\|^2$. The minimum squared loss occurs for a model near those that minimize the two forms of the MSE, the trend in the loss is rather different and more volatile.

Figure 4: Simulation of AIC (normal in blue and heteroscedastic consistent, top) and model loss and mean squared prediction errors (bottom), when fitting a model with autoregressive dependence among the explanatory features.



4 xx

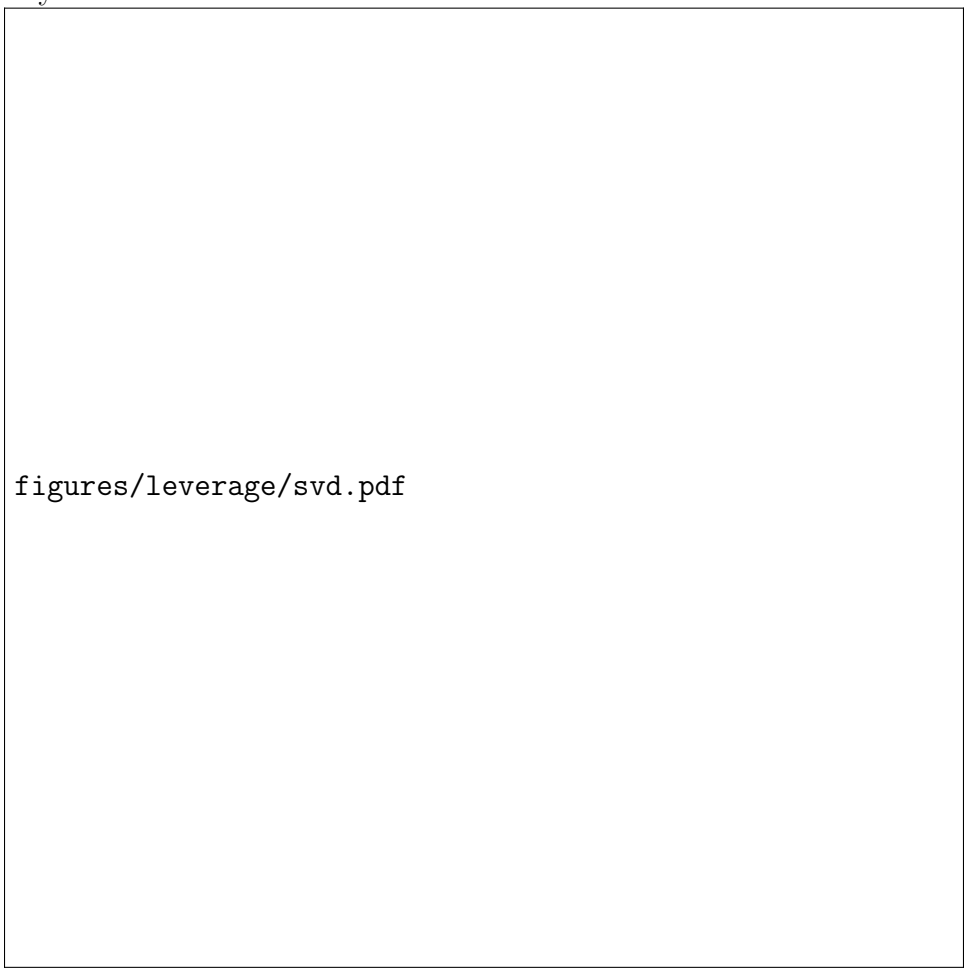
5 Derivations

6 Literature

? discuss variable selection in the random-design model and point out possible "startling" differences from what occurs in the fixed-design model. They recommend cross validation and bootstrap resampling, and note that smaller models are generally preferred with random designs. For some analysis, they write the fixed-design MSE (which they call "model error") as (for the correctly specified "full" model to avoid issues of specification error)

$$(\hat{\beta} - \beta)G(\hat{\beta} - \beta) \quad \text{for } G = X'X \ .$$

Figure 5: Singular values produced by leverage points (black) are larger, with linear decay, compared to those produced by autoregressive collinearity (red). Both matrices have approximately the same condition number.



They write random-design MSE as (after taking the expectation over the distribution of the new $\widetilde{X}$ for which $\mathbb{E}\, G = \Gamma$)

$$(\hat{\beta} - \beta)\Gamma(\hat{\beta} - \beta)(\hat{\beta} - \beta)(\Gamma G^{-1})G(\hat{\beta} - \beta),$$

in order to capture the difference from the fixed-design in the product $\Gamma\, G^{-1}$. The larger $\Gamma\, G^{-1} - I$, the larger the difference between the fixed and random-design errors.

The rest of paper is simulation, with 40 features and $n = 60$ or 160 (done on a Cray, no less). Their emphasis is on variable selection via stepwise, using CVSS, bootstrap or C_p-like methods to select the best subset of features. Simulation uses AR collinear structure, either lognormal or normal. The lognormal case produces rancom number of **leveraged observations**. Nonzero coefficients were in clusters of "adjacent" features, scaled so that $R^2 = 0.75$. Nice conclusion is preference for 10-fold CV over leave-one-out.

With **leveraged observations**, they point out the importance of stratifying on the leverage points when forming the folds. (They sorted by leverages in the full design, then sampled from these strata.) Find that such adjustments did not improve the estimates.

? Gives a correction when using 10-fold cross validation to estimate the MSE of a regression (cited in Breiman).

? notes that cross validation estimates the random-design MSE and should not be used to estimate the MSE for fixed designs (unless in asymtopia).