In any event, though, the effects of overfitting are clear.

4.3 Bias vs. Variance

Let's take a closer look, in an RS context. Say we believe (3.14) is a good model for the setting described in that section, i.e. men becoming more liberal raters as they age but women becoming more conservative. If we omit the interaction term, than we will underpredict older men and overpredict older women. This biases our ratings.

On the other hand, we need to worry about sampling variance. Consider the case of opinion polls during an election campaign, in which the goal is to estimate p, the proportion of voters who will vote for Candidate Jones. If we use too small a sample size, say 50, our results will probably be inaccurate. This is due to sampling instability: Two pollsters, each randomly sampling 50 people, will sample different sets of people, thus each having different values of $\widehat{p}$, their sample estimates of p. For a sample of size 50, it is likely that their two values of $\widehat{p}$ will be substantially different from each other, whereas if the sample size were 5000, the two estimates would likely be close to each other. In other words, the variance of $\widehat{p}$ is too high if the sample size is just 50.[2]

In a parametric regression setting, increasing the number of terms roughly means that the sampling variance of the $\widehat{\beta}_i$ will increase.

So we have the famous *bias/variance tradeoff*: As we use more and more terms in our regression model (predictors, polynomials, interaction terms), the bias decreases but the variance increases. This "tug of war" between these decreasing and increasing quantities typically yields a U-shaped curve: As we increase the number of terms from 1, mean absolute prediction error will at first decrease but eventually will increase. Once we get to the point at which it increases, we are *overfitting*.

This is particularly a problem when one has many dummy variables. For instance, there are more than 42,000 ZIP (postal) codes in the US; to have a dummy for each would almost certainly be overfitting. If we have only, say, 100,000 rows in our data, on average each ZIP code would have only about 2 rows, hardly enough for a good estimate of the effect of that code.

4.4 Mathematical Analysis of the Bias vs. Variance Tradeoff

Let's take a more precise look, employing a simple mathematical model.

[2]The repeatable experiment here is randomly choosing 50 people. Each time we perform this experiment, we get a different set of 50 people, thus a different value of $\widehat{p}$. The latter is a random variable, and thus has a variance.

4.4.1 The Setting

Suppose we have the samples of men's and women's heights, $X_1, ..., X_n$ and $Y_1, ..., Y_n$. Assume for simplicity that the population variance of height is the same for each gender, σ^2. The means of the two populations are designated by μ_1 and μ_2.

Say we wish to guess the height of a new person who we know to be a man but for whom we know nothing else. We do not see him, etc.

Suppose for just a moment that we actually know the distribution of X, i.e. the *population* distribution of male heights. What would be the best constant g to use as our guess for a person about whom we know nothing other than gender?

It is easily shown that the mean squared error MSE

$$E[(g - X)^2] \tag{4.1}$$

is minimized by setting $g = \mu_1$. Our best guess for this unseen man's height is the mean height of all men in the population. (Note that "mean" above averaged over all possible men in the population.)

Of course, we don't know μ_1, but we can do the next-best thing, i.e. use an estimate of it from our sample. The natural choice for that estimator would be

$$T_1 = \overline{X}, \tag{4.2}$$

the mean height of men in our sample.

4.4.2 Context of Interest: Very Small Sample

But what if our sample size n is really small, say $n = 5$? That's awfully small. We may wish to consider pooling the women's heights into our estimate, in order to get a larger sample. Then we would estimate μ_1 by incorporating the sample mean of women's heights, $\overline{Y}$:

$$T_2 = \frac{\overline{X} + \overline{Y}}{2}, \tag{4.3}$$

It may at first seem obvious that T_1 is the better estimator. Women tend to be shorter, after all, so pooling the data from the two genders would induce a bias, defined as

$$\text{bias} = \text{mean of the estimator - true population value} \tag{4.4}$$

Here "mean" refers to the average of the estimator over all possible samples from this population. It can be shown that for a sample mean M, drawn from a population with mean ν,

$$E(M) = \nu \tag{4.5}$$

In other words, M has 0 bias. Thus our T_1 here has 0 bias. But that is not the case for T_2:

$$E(T_2) = 0.5E(\overline{X}) + 0.5E(\overline{Y}) = (\mu_1 + \mu_2)/2 < \mu_1 \tag{4.6}$$

In other words. T_2 would have a negative bias.

For an estimator of T of some population quantity θ, its *mean square error* is defined to be

$$MSE = E[(T - \theta)^2] \tag{4.7}$$

One can derive that

$$\text{MSE} = \text{variance of the estimator} + \text{bias of the estimator}^2 \tag{4.8}$$

In other words, *some amount of bias may be tolerable*, if it will buy us a subtantial reduction in variance. After all, women are not that much shorter than men, so the bias might not be too bad. Meanwhile, the pooled estimate should have lower variance, as it is based on $2n$ data points, rather than n.

Before continuing, note first that T_2 is based on a simpler model than is T_1, as T_2 ignores gender. We thus refer to T_1 as being based on the more complex model.

So, the question becomes, which has the smaller MSE, T_1 or T_2? In other words:

> Which is smaller, $E[(T_1 - \mu_1)^2]$ or $E[(T_2 - \mu_1)^2]$?

4.4.3 Drawing Conclusions from This Example

After some elementary math stat operations, one can show that T_1 is a better predictor than T_2 if

$$\left(\frac{\mu_2 - \mu_1}{2}\right)^2 > \frac{\sigma^2}{2n} \tag{4.9}$$

Granted, we don't know the values of the μ_1 and σ^2, so in a real situation, we won't really know whether to use T_1 or T_2. But the above analysis makes the point that under some circumstances, it really is better to pool the data in spite of bias.

So you can see that T_1 is better only if either

- n is large enough, or
- the difference in population mean heights between men and women is large enough, or
- there is not much variation within each population, e.g. most men have very similar heights

In other words:

> A more complex model is more accurate than a simpler one only if either
>
> (a) we have enough data to support it, or
> (b) the complex model is sufficiently different from the simpler one

A very rough, intuitive way to view (a) is that our data is being "shared" by all the parameters to be estimated. In our example above, the simple model had one parameter, μ while the complex one had two, μ_1 and μ_2. Due to this "sharing," each parameter in the complex version has "a smaller piece of the pie."

In Section 4.2, we ran an **lm()** model with 2624 parameters, definitely a complex model. Was $n = 100000$ large enough to satisfy (a) above? We don't know, but again, it raises the issue of possible overfitting.

4.5 Can Anything Be Done about It?

So, where is the "happy medium," the model that is rich enough to capture most of the dynamics of the variables at hand, but simple enough to avoid variance issues? Unfortunately, **there is no good answer to this question.**

4.5.1 Rough Rule of Thumb

One quick rule, backed up by mathematical theory, is that one should have $p < \sqrt{n}$, where p is the number of predictors, including polynomial and interaction terms (not to be confused with the quantity of the same name in our polling example above), and n is the number of cases in our sample. But this is certainly not a firm rule by any means, and I find it tends to be overly conservative.

the blurriness.

- Applying NMF or PCA/SVD to a whole collection of images, e.g. MNIST, further heightens this approximate nature of the process.

- But we need to do something to avoid overfitting, i.e. some kind of dimension reduction, and finding a low-rank approximation does that.

5.8 The Bias vs. Variance Tradeoff

The blurriness in that second picture is really an issue of bias, as follows. Consider a given pixel, say in the 3rd row and 52nd column. That pixel's intensity in the second picture will be a weighted average of various pixels in the first picture. Some of the latter may be in locations within the picture that are somewhat far away from the 3rd row and 52nd column. This biases the pixel in the second picture.

On the other hand, there definitely is a variance issue. Let's review what this entails.

Recall from Chapter 4 that an intuitive way to view the variance issue in overfitting is that our data are being "shared" by the various things we're estimating, so that in a rough sense, each of these things has less data to itself. Less data means more sample-to-sample variability, i.e. higher variance. In linear regression with p features, we are estimating $p + 1$ parameters (including β_0); the larger p is, the larger the variance of the estimated β_i. Thus in turn we get larger variance to our predicted values. For predicting a new case, different samples will give us different predictions, and larger p will give us higher variance in our predicted value for that case.

Let n and m denote the number of rows and columns in A. Then W and H will be of dimensions $n \times k$ and $k \times m$. Well, then, how many parameters are we estimating? It's

$$nk + km = k(n + m) \tag{5.9}$$

So, the larger we make k, the larger the variance.

In other words, in predicting a specific A_{ij}, our predicted value $\widehat{A}_{ij}$ will experience this tradeoff:

- Larger k means lesser bias in our estimate of $\widehat{A}_{ij}$.

- Larger k means greater variance in $\widehat{A}_{ij}$.