Differential downward bias of variance explained due to the singular value spectrum of the feature matrix

Abstract

Prediction of empirical phenomena of interest is an essential component of a mature science. In this technical note, we characterize one particular pitfall in this endeavor when measuring the variance in a phenomenon that a given model is able to explain. We first note the equivalence between minimizing the squared error and maximizing the explained variance. Then we note the downward bias on measures of variance explained due to measurement noise, and how to correct for it. Importantly, we note that different feature sets can exhibit different extents of downward bias *even when predicting the identical noisy data*. We analytically note the source of this differential downward bias and offer preliminary evidence that it occurs in practice. This work emphasizes the importance of correcting for both the reliability of the data and the reliability of the prediction of that data in order to make appropriate comparisons between models.

1 Introduction: Squared error and measures of variance explained

1.1 Squared error in SLT

Statistical learning theory (SLT) is often concerned with selecting a function f that minimizes the expected risk over the joint distribution for a given problem:

$$\min_{f \in \mathcal{H}} \mathbb{E}_{p(\mathcal{X},y)}[V(y, f(x))]$$

Typically, the joint distribution over $\mathcal{X}$ and y is not available, and in practice often the empirical risk is minimized. A common choice of loss function $V(\cdot)$ is the squared error:

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} V(f(\mathbf{x}_i), y_i) = \min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (f(\mathbf{x}_i) - y_i)^2$$

In many scenarios, the hypothesis space is constrained to the space of linear functions, leaving:

$$\min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (\mathbf{w}^T \mathbf{x}_i - y_i)^2 = \min_{\mathbf{w} \in \mathbb{R}^d} \frac{1}{n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2$$

Note that here $\mathbf{X}$ may refer to the original data or a the data passed through a feature map $\Phi(\cdot)$.

1.2 Variance explained in fields of science

In science, we often test hypotheses about a phenomenon by modeling that phenomenon and evaluating the quality of a model by estimating its error in predicting the phenomenon in left-out data.

*Thanks to Jenelle Feather and Sam Norman-Haignere for useful discussions on issues related to those that I explore here. Thanks to all the course staff and TAs, and particularly to Georgios for help over the course!

MIT 9.520/6.860: Statistical Learning Theory and Applications (Fall 2016).

Consider the case of many samples of some univariate measure of a phenomenon of interest under different conditions or manipulations ($\mathbf{y}$) and a model's prediction of that phenomenon ($\hat{\mathbf{y}}$). There are many ways to generate such a prediction, but a common choice is to create a linear combination (weights: $\mathbf{w}$)of variables (or functions of those variables $\Phi(\cdot)$) that are thought to be relevant to the phenomenon of interest ($\mathbf{X}$). There are many means of evaluating the quality of the fit. A typical one consists of computing the mean squared error (MSE), which is identical to the squared-loss objective often optimized in SLT:

$$\frac{1}{n}\|\hat{\mathbf{y}} - \mathbf{y}\|_2^2 = \frac{1}{n}\|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2$$

In many cases it is useful to normalize this MSE measure in some way. One possibility is to measure the variance of the original variable that you are explaining (ν) by measuring the reduction in the variance of the phenomenon of interest (e.g., Haefner & Cumming 2009; Sahani & Linden, 2003):

$$\nu_{\mathbf{y}|\hat{\mathbf{y}}} = \frac{SS_{tot} - SS_{err}}{SS_{tot}} = 1 - \frac{SS_{err}}{SS_{tot}} = 1 - \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|}{\|\mathbf{y}\|}$$

Another choice is to measure the quality of the prediction by computing correlation between the observed variable and the prediction of the observed variable ($\mathbf{r}$), and then squaring it to obtain an alternative measure of variance explained (David & Gallant, 2005; Nishimoto & Gallant, 2012; Huth et al., 2012; Santoro et al., 2014; Huth et al., 2016.):

$$\mathbf{r} = \frac{\text{cov}[y, \hat{y}]}{\sqrt{\text{var}[y]\text{var}[\hat{y}]}} = \frac{\frac{1}{n}\sum_{i=1}^{n} \hat{y}_i y_i}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 \frac{1}{n}\sum_{i=1}^{n} \hat{y}_i^2}} = \frac{\hat{\mathbf{y}}^T\mathbf{y}}{\|\mathbf{y}\|\|\hat{\mathbf{y}}\|}$$

(Note that throughout I assume variables are demeaned. Also note that boldface font denotes a vector with the exception of the correlation coefficient: $\mathbf{r}$.)

Here we will consider this second notion of variance explained, the squared (Pearson) correlation coefficient. At first glance it's not obvious the relationship between the squared correlation coefficient and the squared error loss function and in the next section I will show that the minimizers of each are identical.

2 The maximizer of the squared correlation coefficient is the minimizer of the squared error

2.1 Minimizer of squared error: $\mathbf{w}^*$

Consider the case where $\mathbf{y}$ is demeaned and unit norm and the columns of $\mathbf{X}$ (either the raw data or features computed from the original data) are demeaned and unit norm. The minimizer of the squared empirical error is:

$$\mathbf{w}^* = \text{argmin}_{\mathbf{w}\in\mathbb{R}^d} \sum_{i=1}^{n} (\mathbf{w}^T\mathbf{x}_i - y_i)^2$$

$$\mathbf{w}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

2.2 Maximizer of squared correlation coefficient: $\mathbf{w}'$

Now consider the maximizer of the squared correlation coefficient:

$$\mathbf{w}' = \text{argmax}_{\mathbf{w}\in\mathbb{R}^d} \frac{(\sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i y_i))^2}{\sum_{i=1}^{n}(\mathbf{w}^T\mathbf{x}_i)^2 \sum_{i=1}^{n} y_i^2} = \text{argmax}_{\mathbf{w}\in\mathbb{R}^d} \frac{(\mathbf{y}^T\mathbf{X}\mathbf{w})^2}{\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w}}$$

Where we note that $\mathbf{y}^T\mathbf{y}$ is unaffected by our choice of $\mathbf{w}$ and thus we can ignore it.

Maximizing this full equation w/r/t $\mathbf{w}$ is equivalent to maximizing the numerator with a constraint place on the norm of $\mathbf{w}$, which we can achieve by constraining $\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} = 1$. We can move from this case of constrained optimization to unconstrained optimization by adding an auxilary variable – i.e., writing down the Lagrangian:

$$L(\mathbf{w}, \lambda) = (\mathbf{y}^T\mathbf{X}\mathbf{w})^2 - \lambda(\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - 1)$$

The squared correlation coefficient will be maximized when the gradient of the Lagrangian equals zero. We can compute the gradient with respect to $\mathbf{w}$:

$$\nabla_{\mathbf{w}} L(\mathbf{w}, \lambda) = \mathbf{0} = 2\mathbf{X}^T\mathbf{y}\mathbf{y}^T\mathbf{X}\mathbf{w}' - 2\lambda\mathbf{X}^T\mathbf{X}\mathbf{w}'$$
$$\mathbf{X}^T\mathbf{y}\mathbf{y}^T\mathbf{X}\mathbf{w}' = \lambda\mathbf{X}^T\mathbf{X}\mathbf{w}'$$
$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\mathbf{y}^T\mathbf{X}\mathbf{w}' = \lambda\mathbf{w}'$$

Here we see that we have an eigenvalue problem: $\mathbf{w}'$ is an eigenvector of $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\mathbf{y}^T\mathbf{X}$ and λ is the corresponding eigenvalue.

Furthermore, note that the matrix $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}\mathbf{y}^T\mathbf{X}$ has a rank of one – it is the outer product of the vector $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ and $\mathbf{X}^T\mathbf{y}$. Therefore there is a single nonzero eigenvalue, which we can solve for.

We can view this problem in a more general case where $\mathbf{x}$ is an eigenvector and λ is an eigenvalue:

$$\mathbf{a}\mathbf{b}^T\mathbf{x} = \lambda\mathbf{x}$$

The only nonzero eigenvalue is the inner product of the two vectors $(\mathbf{a}^T\mathbf{b})$ and its corresponding eigenvector is $\mathbf{a}$:

$$\mathbf{a}\mathbf{b}^T(\mathbf{a}) = (\mathbf{a}^T\mathbf{b})\mathbf{a}$$

When we manipulate the LHS we see that it is equivalent to the RHS:

$$\mathbf{a}\mathbf{b}^T(\mathbf{a}) = \mathbf{a}(\mathbf{b}^T\mathbf{a}) = (\mathbf{b}^T\mathbf{a})\mathbf{a} = (\mathbf{a}^T\mathbf{b})\mathbf{a}$$

In our case:

$$\mathbf{a} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}; \quad \mathbf{b} = \mathbf{X}^T\mathbf{y}$$

And so:

$$\mathbf{w}' = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$
$$\lambda = ((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y})^T\mathbf{X}^T\mathbf{y} = \mathbf{y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

And so we see the connection between the measure of the squared correlation measure of variance explained ($\mathbf{r^2}$) and the minimizer of the squared error – they are identical: $\mathbf{w}' = \mathbf{w}^*$.

2.3 Examining the value of the Lagrange multiplier: λ

Our solution for the eigenvalue/Langrange multiplier can be simplified by taking the SVD of $\mathbf{X}$ ($\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}^T$):

$$\lambda = \mathbf{y}^T\mathbf{U}\mathbf{S}\mathbf{V}^T(\mathbf{V}\mathbf{S}^2\mathbf{V}^T)^{-1}\mathbf{V}\mathbf{S}\mathbf{U}^T\mathbf{y}$$
$$= \mathbf{y}^T\mathbf{U}\mathbf{S}\mathbf{V}^T(\mathbf{V}\mathbf{S}^{-2}\mathbf{V}^T)\mathbf{V}\mathbf{S}\mathbf{U}^T\mathbf{y}$$
$$= \mathbf{y}^T\mathbf{U}\mathbf{U}^T\mathbf{y}$$
$$= \|\mathbf{y}^T\mathbf{U}\|_2^2$$

Thus, the value of the Lagrange multiplier is the norm of the dot product of $\mathbf{y}$ with each of the eigenvectors of $\mathbf{X}\mathbf{X}^T$. The Cauchy-Schwartz inequality indicates tells us that for each column $\mathbf{u}^{(j)}, \forall j = 1, ..., d$:

$$\|\mathbf{y} \cdot \mathbf{u}^{(j)}\|_2^2 \leq \|\mathbf{y}\|_2^2\|\mathbf{u}^{(j)}\|_2^2$$

The norm for both $\mathbf{y}$ and $\mathbf{u}^{(j)}$ is 1 (respectively: by design and because eigenvectors are orthonormal):

$$\|\mathbf{y} \cdot \mathbf{u}^{(j)}\|_2^2 \leq 1$$

The columns $\mathbf{u}^{(j)}$ are orthogonal and so any proportion of $\mathbf{y}$ onto $\mathbf{u}^{(k)}$ does not project onto $\mathbf{u}^{(l)}$ and norms are nonnegative. Therefore:

$$0 \leq \lambda = \sum_{j=1}^{d} \|\mathbf{y} \cdot \mathbf{u}^{(j)}\|_2^2 \leq 1$$

If the data $\mathbf{y}$ lies orthogonal to the subspace of $\mathbb{R}^n$ that the d columns of the eigenvectors of $\mathbf{X}\mathbf{X}^T$ span, then $\lambda = 0$, and if $\mathbf{y}$ lies entirely in this d dimensional subspace of $\mathbb{R}^n$, then $\lambda = 1$. In general, the value of λ is the norm of the vector $\mathbf{y}$ projected onto the d-dimensional subspace of $\mathbb{R}^n$ that the columns of the design matrix $\mathbf{X}$ span. Therefore, as it may be clear by now, the value of the Lagrangian multiplier is the squared correlation coefficient – i.e., the value of $\mathbf{r}^2$ for $\mathbf{w}'$.

3 Downward bias of $\mathbf{r^2}$ due to noise in $\mathbf{y}$

3.1 The problem of downward bias due to measurement noise

Now that we've seen the equivalence between minimizing the squared error and maximizing the squared correlation coefficient between the prediction and the empirical data, let's consider measuing $\mathbf{r^2}$ in practice. Typically, we want to characterize the strength of the relationship between $\mathbf{y}$ and our prediction from a set of features $\mathbf{X}$ in general, but our measurements of a phenomenon is often corrupted by noise (here we'll assume that the measurements of variables $\mathbf{X}$ are noiseless). This measurement noise will downwardly bias our empirically measured $\mathbf{r^2}$ i.e., the ceiling will not actually be one. In order to remove this bias, we can take into account of the reliability of each of these variables.

Let's consider a general case. If we have two variables of interest $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and we receive two measures of each of these variables $\mathbf{x}_a, \mathbf{x}_b, \mathbf{y}_a, \mathbf{y}_b$, each of which is a noisy sample of the true variables – i.e., $x_a^{(1)}, x_a^{(2)}, ..., x_a^{(n)} = x^{(1)} + \delta^{(1)}, x^{(2)} + \delta^{(2)}, ..., x^{(d)} + \delta^{(n)}$, where $\delta^{(i)}$ are all sampled i.i.d. Similarly for y with ϵ instead of δ.

The true correlation is:

$$\mathbf{r_{xy}} = \frac{\sum_{i=1}^n x^{(i)} y^{(i)}}{\sqrt{\sum_{i=1}^n x^{(i)2} \sum_{i=1}^n y^{(i)2}}} = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

Note that after this point, I'm going to drop the i indexing, and use the form in the last equality above.

If we compute this measure for a given set of our observed data, we see that we don't have the same equation:

$$\begin{aligned}\mathbf{r_{x_a y_a}} &= \frac{\sum x_a y_a}{\sqrt{\sum x_a^2 \sum y_a^2}} = \frac{\sum (x+\delta)(y+\epsilon)}{\sqrt{\sum (x+\delta)^2 \sum (y+\epsilon)^2}} \\ &= \frac{\sum xy + \sum x\epsilon + \sum y\delta + \sum \delta\epsilon}{\sqrt{(\sum x^2 + 2\sum x\delta + \sum \delta^2)(\sum y^2 + \sum y\epsilon + \sum \epsilon^2)}} \\ &= \frac{\sum xy}{\sqrt{(\sum x^2 + \sum \delta^2)(\sum y^2 + \sum \epsilon^2)}}\end{aligned}$$

Above invoked the assumption of additive noise (e.g., $x_a = x + \delta$), and the assumption that the noise is uncorrelated with the signal of interest, and therefore the dot product of the noise vector with the signal vector equals zero (i.e., $\sum x\epsilon = \sum y\delta = \sum y\epsilon = \sum x\delta = 0$).

So here we see that we're biased downward $\mathbf{r_{x_a y_a}} \leq \mathbf{r_{xy}}$, because $\sum \delta^2 \geq 0$ and $\sum \epsilon^2 \geq 0$, and the equality holds only in the noiseless case.

To be explicit about the problem in our case, let's replace $\mathbf{x}$ and $\mathbf{y}$ with $\mathbf{y}$ and $\hat{\mathbf{y}}$, our measure of a phenomenon and prediction of that phenomenon, respectively.

$$\mathbf{r_{y_a \hat{y}_a}} = \frac{\sum y_a \hat{y}_a}{\sqrt{(\sum y_a^2 + \sum \delta_y^2)(\sum \hat{y}_a^2 + \sum \epsilon_{\hat{y}}^2)}}$$

Even if $\hat{\mathbf{y}} = \mathbf{y}$, i.e. our model is correct, our measured correlation metric would not reflect this fact. Note that we'd be biased downward by the variance of the noise in the measure of the phenomenon of interest and the variance of the noise of our estimate $\hat{\mathbf{y}}$ compared with the "true" $\hat{\mathbf{y}}$ – i.e., the $\hat{\mathbf{y}}$ estimated with no noise in the training data for estimating $\mathbf{w}$.

3.2 Correcting for this downward bias: Accounting for the reliability of measurements

We can remove this bias by correcting for the reliability of the data. Specifically, by dividing by the root of the product of the reliability.

$$\mathbf{r_{xy}} = \mathbf{r^*_{x_a y_a}} = \frac{\mathbf{r_{x_a y_a}}}{\sqrt{\mathbf{r_{x_a x_b}} \mathbf{r_{y_a y_b}}}}$$

Note that:

$$\mathbf{r_{x_a x_b}} = \frac{\sum x^2}{\sum x^2 + \sum \delta^2} = \frac{\sigma^2_{\mathbf{x}}}{\sigma^2_{\mathbf{x}} + \sigma^2_{\delta}}$$

(Assuming noise of equal variance: $\sum \delta_a^2 = \sum \delta_b^2 = \sum \delta^2$)

And analogously results for $\mathbf{r_{y_a y_b}}$. And therefore:

$$\mathbf{r^*_{x_a y_a}} = \frac{\mathbf{r_{x_a y_a}}}{\sqrt{\mathbf{r_{x_a x_b} r_{y_a y_b}}}} = \frac{\frac{\sum xy}{\sqrt{(\sum x^2 + \sum \delta^2)(\sum y^2 + \sum \epsilon^2)}}}{\sqrt{\frac{\sum x^2}{\sum x^2 + \sum \delta^2} \frac{\sum y^2}{\sum y^2 + \sum \epsilon^2}}}$$

Rewriting this last division out makes it clear that you can just cross out a bunch of terms:

$$\mathbf{r^*_{x_a y_a}} = \frac{\sum xy}{\sqrt{(\sum x^2 + \sum \delta^2)(\sum y^2 + \sum \epsilon^2)}} \frac{\sqrt{(\sum x^2 + \sum \delta^2)(\sum y^2 + \sum \epsilon^2)}}{\sqrt{\sum x^2 \sum y^2}} = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}} = \mathbf{r_{xy}}$$

And thus we've shown that $\mathbf{r^*_{x_a y_a}}$ gets rid of the noise term and exactly equals $\mathbf{r_{xy}}$ if our assumptions hold.

This identity is relatively straightforward to show, but to be clear we did invoke a number of assumptions:

- Noise is additive: $x_a = x + \delta_a$
- Noise is uncorrelated with the underlying variables of interest: $\sum x \delta_a = 0$
- Noise is sampled i.i.d., which means:
  - Different noise samples are uncorrelated: $\sum \delta_a \delta_b = 0$
  - The noise levels (i.e., variances) are the same across different splits: $\sum \delta_a^2 = \sum \delta_b^2$
- Lastly, we assumed that we had enough samples of data (the n that we were summing our vectors over was sufficiently large) that the dot products of these vectors were indeed zero rather than having a non-zero value. In lower-dimensional instances, you'd expect that these values would not in fact be zero and thus this correction would be higher variance with a lower n.

Note that we didn't have to assume anything about the distribution of the noise – e.g., that that the noise is Gaussian!

3.3 Differential downward bias on $\hat{\mathbf{y}}$ for predicting the *same* measured phenomenon

Often when we are evaluating the prediction of a given phenomenon, we compare the quality of the prediction for different models – i.e., different feature sets $\mathbf{X}$. In this section we note the particular problem of not for correcting for the reliability of our prediction: different feature sets $\mathbf{X}$ may exhibit downward biases of different magnitudes when predicting the same data $\mathbf{y}$. (Note that this point is particularly important because it is not common to correct for reliability of the prediction in some fields, e.g., in computational and systems neuroscience.)

To see this potential for differential bias, let's consider where the noise on $\hat{\mathbf{y}}$ comes from. To see this, we need some notation to denote train and test data – and we'll be doing with different measures of the phenomenon. $\mathbf{y}_{train}$ is the training vector of length p elements from $\mathbf{y}$, and $\mathbf{y}_{test}$ is an $n - p$ length vector of the remaining elements of the original $\mathbf{y}$. $\mathbf{F}_{train}$ and $\mathbf{F}_{test}$ are the analogous matrices of the features that are respectively p by d and $n - p$ by d. If we use L-2 regularized linear regression, we compute our linear weights $\mathbf{w}^*$ as follows:

$$\mathbf{w}^* = (\mathbf{F}_{train}^T \mathbf{F}_{train} + n\lambda \mathbf{I}_d)^{-1} \mathbf{F}_{train}^T \mathbf{y}_{train}$$

And our prediction is:

$$\hat{\mathbf{y}}_{test} = \mathbf{F}_{test}\mathbf{w}^* = \mathbf{M}'\mathbf{y}_{train}$$

Where $\mathbf{M}' = \mathbf{F}_{test}\mathbf{M}$. We can expand out $\mathbf{y}_{train}$:

$$\hat{\mathbf{y}}_{test} = \mathbf{M}'(\mathbf{y} + \boldsymbol{\delta}_{train}) = \mathbf{M}'\mathbf{y} + \mathbf{M}'\boldsymbol{\delta}_{train} = \hat{\mathbf{y}}_{noiseless} + \boldsymbol{\epsilon}_{train}$$

Here we can see how the noise in the training data corrupts the predicted response. Namely, it is the sum of the true, noiseless $\hat{\mathbf{y}}$ and a linear transform of the training data noise (i.e., $\boldsymbol{\epsilon}_{train} = \mathbf{M}'\boldsymbol{\delta}_{train}$).

It's worth noting that we invoked that the noise on $\hat{y}$ was additive in section 3.2 and here we see explicitly that this assumption we get for free via our assumption of additive noise on the $\mathbf{y}$. It's also worth pointing out that the matrix $\mathbf{M}$ or $\mathbf{M}'$ is strictly a function of the choice of regressors and regularization parameter you are using – specifically, $\mathbf{M}$ *is not affected by the data of* $\mathbf{y}$ *at all.*

To summarize, the extent of the downward bias for a given feature set (i.e., our reliability of $\hat{\mathbf{y}}_{test}$) will be a function of the relative contribution of the signal, $\hat{\mathbf{y}}_{noiseless}$, and the noise, $\boldsymbol{\epsilon} = \mathbf{M}\boldsymbol{\delta}_{train}$. If we measure in two halves of data, a and b:

$$\hat{\mathbf{y}}_a = \hat{\mathbf{y}}_{noiseless} + \boldsymbol{\epsilon}_a; \;\; \hat{\mathbf{y}}_b = \hat{\mathbf{y}}_{noiseless} + \boldsymbol{\epsilon}_b$$

$$\mathbf{r}_{\hat{\mathbf{y}}_\mathbf{a}\hat{\mathbf{y}}_\mathbf{b}} = \frac{\sum \hat{y}^2_{noiseless}}{\sum \hat{y}^2_{noiseless} + \sum \epsilon^2} = \frac{\sigma^2_{\hat{y}_{noiseless}}}{\sigma^2_{\hat{y}_{noiseless}} + \sigma^2_\epsilon}$$

$\hat{\mathbf{y}}$ captures how well this feature set $\mathbf{X}$ models the phenomenon of interest, and obviously depends on both $\mathbf{X}$ and the target phenomenon, $\mathbf{y}$. In contrast, the noise term, $\boldsymbol{\epsilon}_{train}$, will not be a function of the strength of the predictive power and will simply be a function of properties of the feature matrix $\mathbf{X}$, and we explore these properties of $\mathbf{X}$ in the following section.

To be clear, this differential downward bias is potentially worrisome because models are often compared in how much variance they explain. Failing to account for the reliability of $\mathbf{y}$ will downwardly bias the magnitude of the $\mathbf{r}^2$ equally for all models that are compared – not correcting for $\mathbf{y}$ reliability will not affect your inferences about which model explains more variance. In contrast, this *differential* downward bias could affective the relative variance explained measures across models and could potentially lead to incorrect conclusions about which models best predict the phenomenon of interest.

3.4 The source of this differential downward bias: singular values of X

Where does this differential downward bias come from? Consider $\mathbf{w}^*$ in the case of vanilla linear regression:

$$\mathbf{w}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}_{train} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T(\mathbf{y} + \boldsymbol{\delta}) = \mathbf{w}_{noiseless} + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\delta}$$

First, note that $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\delta}$ is zero mean:

$$\mathbb{E}_{p(\boldsymbol{\delta})}[\mathbf{A}\boldsymbol{\delta}] = \mathbf{A}\mathbb{E}_{p(\boldsymbol{\delta})}[\boldsymbol{\delta}] = \mathbf{A}\mathbf{0} = \mathbf{0}$$

Next let's consider the total noise that $\boldsymbol{\delta}$ introduces to our estimate of $\mathbf{w}^*$:

$$\frac{1}{d}\sum_{i=1}^{d} \text{var}[\mathbf{w}^{*(i)}] = \frac{1}{d}\sum_{i=1}^{d} \text{var}[((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\delta})^{(i)}]$$

Start by considering the covariance, and we know that $\text{cov}[\mathbf{A}\mathbf{x}] = \mathbf{A}\text{cov}[\mathbf{x}]\mathbf{A}^T$:

$$\text{cov}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\delta}] = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\text{cov}[\boldsymbol{\delta}]((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)^T$$

Two observations to simplify the equation above. First:

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = ((\mathbf{U}\mathbf{S}\mathbf{V}^T)^T(\mathbf{U}\mathbf{S}\mathbf{V}^T))^{-1}(\mathbf{U}\mathbf{S}\mathbf{V}^T)^T = (\mathbf{V}\mathbf{S}^{-2}\mathbf{V}^T)(\mathbf{V}\mathbf{S}\mathbf{U}^T) = \mathbf{V}\mathbf{S}^{-1}\mathbf{U}^T$$

Second, we'll introduce the diagonal matrix $\mathbf{D}$ where each element on the diagonal is σ^2, the variance of the additive noise on our measurement $\mathbf{y}$. We are left with:

$$\text{cov}[(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{\delta}] = \mathbf{V}\mathbf{S}^{-1}\mathbf{U}^T\mathbf{D}(\mathbf{V}\mathbf{S}^{-1}\mathbf{U}^T)^T = \mathbf{V}\mathbf{S}^{-2}\mathbf{D}\mathbf{V}^T = \mathbf{S}^{-2}\mathbf{D}$$

(Note that $n > d$ so $\mathbf{V}$ is square and so not only: $\mathbf{V}^T\mathbf{V} = \mathbf{I}_d$, but also: $\mathbf{V}\mathbf{V}^T = \mathbf{I}_d$.)

So we see that the noise is uncorrelated across elements of $\mathbf{w}^*$ (i.e., the covariance is diagonal), and the cumulative variance is:

$$\frac{1}{d}\sum_{i=1}^{d}\text{var}[\mathbf{w}^{*(i)}] = \frac{1}{d}\sum_{i=1}^{d}\frac{\sigma^2}{s_i^2} = \frac{\sigma^2}{d}\sum_{i=1}^{d}\frac{1}{s_i^2}$$

Two things to note here. First, intuitively, as the noise on your training data increases, the noise on your inferred weights increases. Second, the source of the downward bias is clear: different feature sets $\mathbf{X}$ may have different singular value spectra, and these different singular value spectra will differentially affect the noise through the summation term.

What can we say about these singular values? Well, first note that the columns of $\mathbf{X}$ are zero mean and unit norm, so we know that the Frobenius norm for any choice of matrix $\mathbf{X}$ will be the same: $\|\mathbf{X}\|_F = \sqrt{\sum_{j=1}^{d}\sum_{i=1}^{n}X_{ij}^2} = \sqrt{\sum_{j=1}^{d}1} = \sqrt{d}$. Further, to characterize the singular values of a matrix we can turn to the definition of the Frobenius norm: $\|\mathbf{A}\|_F = \sqrt{\text{Tr}[\mathbf{A}^T\mathbf{A}]}$.

$$\|\mathbf{X}\|_F^2 = \text{Tr}[\mathbf{X}^T\mathbf{X}] = \text{Tr}[\mathbf{V}\mathbf{S}^2\mathbf{V}^T] = \text{Tr}[\mathbf{S}^2] = \sum_{i=1}^{d}s_i^2 = d$$

So the sum of the square of the singular values is identical for any choice of matrix $\mathbf{X}$ is identical: d.

Given this constraint on the sum of the squared singular values, we can consider how the singular value spectrum of $\mathbf{X}$ will affect the noise on $\mathbf{w}^*$. If the singular value spectrum is flat – i.e., $s_i^2 = \frac{1}{d}$ then the cumulative noise on $\mathbf{w}^*$ will be be:

$$\frac{1}{d}\sum_{i=1}^{d}\text{var}[\mathbf{w}^{*(i)}] = \frac{\sigma^2}{d}\sum_{i=1}^{d}\frac{1}{1/d} = d\sigma^2$$

In contrast, if the singular value spectrum has one quite large squared singular value d' and $d-1$ very small squared singular values ϵ, we see:

$$\frac{1}{d}\sum_{i=1}^{d}\text{var}[\mathbf{w}^{*(i)}] = \frac{\sigma^2}{d}\left(\frac{1}{d'} + \sum_{i=1}^{d-1}\frac{1}{e}\right) = \frac{\sigma^2}{\epsilon}\frac{d-1}{d} + \frac{\sigma^2}{dd'}$$

We see that very small squared singular values will render the cumulative noise on $\mathbf{w}^*$ to increase substantially.

Based on this analysis, we see that the flatter a singular value spectrum is, the less noise will corrupt our estimate of $\mathbf{w}^*$ and thus the less our noise will corrupt our predictions $\hat{\mathbf{y}} = \mathbf{X}\mathbf{w}^*$. Intuitively, this finding may make sense, considering that small singular values will become quite large in the matrix inversion and thus amplify the noise that was in the original signal.

3.5 Initial empirical results suggesting that the differential downward bias occurs in practice

How much of a problem is this differential downward bias in practice? In this final section we will examine the effects in one empirical case to show that in fact this differential downward bias is a problem in real data.

Specifically, we'll take a case from my research where I am using a hierarchical convolutional neural network trained to recognize words and musical genres to predict auditory cortical responses to natural sounds measured with fMRI. For each voxel, I have a 165-dimensional vector of that voxel's mean response to each of 165 different natural sounds ($\mathbf{y}$). To predict this phenomenon of interest, I am passing those same 165 natural sounds through this neural network and extracting each model

units response in each of 17 different layers to each of the sounds, and averaging that model unit's response over time (obtaining a $165 \times d_l$ matrix $\mathbf{X}_l$ for each layer $l = 1, ..., 17$). I then am using cross-validated L2-regularized linear regression to predict each voxel from the features extracted from each layer of the network $\mathbf{y} = \mathbf{X}_l\mathbf{w}$. I am evaluating the variance explained with the squared correlation between the predicted responses and the measured responses to left-out sounds.

To empirical test the extent to which the choice of feature matrix $\mathbf{X}_l$ affects the reliability of the prediction, we can compute a summary measure of the eigenspectra of $\mathbf{X}_l$ for each feature set and compare it with how *reliable* that feature set's predictions are, measured by correlating predictions (different $\hat{\mathbf{y}}$) across different presentations of the same stimuli.

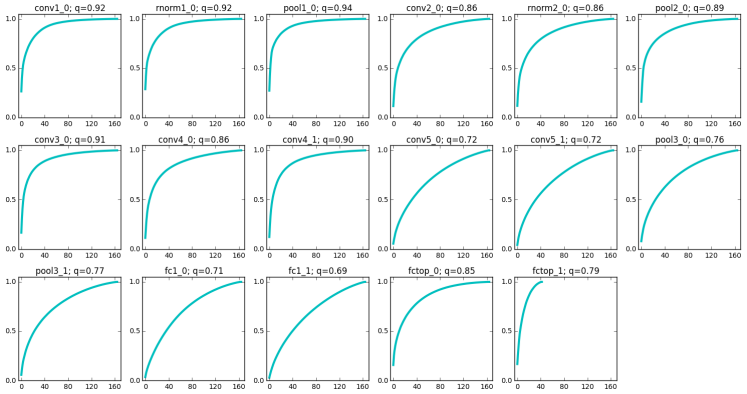

Figure 1: Cumulative squared singular value spectra for all 17 feature sets.

The summary measure of the singular values that we'll use is the area under the curve of the cumulative squared sorted singular values, i.e.: $q = \frac{1}{165^2} \sum_{j=1}^{165} \sum_{i=1}^{j} s_i^2$. Note that because $\sum_{i=1}^{165} s_i^2 = 165$, $q \leq 1$, with $q = 1$ iff there were one nonzero singular value (we'd have a "square"). If the singular value spectrum was flat, then $q = 0.5$ (we'd have a "triangle"), so $q \in [0.5, 1.0]$.

When we examine plots of the squared sorted cumulative singular values (see Fig. 1), we see that there is variance across different layers' responses (i.e., the feature sets $\mathbf{X}_l$).

We can then examine how the area under each of these curves (q) relates to the reliability of the predictions. Recall that the reliability of the prediction $\hat{\mathbf{y}}$ is a measure of the signal-to-noise of the prediction versus the noise. We've focused on how the absolute amount of noise may be modulated by the singular value spectrum of the feature set – i.e., we've focused on the noise portion of this signal-to-noise ratio. Obviously, the amount of signal also will affect these reliabilities, so it will be useful also to note how good of a model a given feature set is for a given voxel's response (i.e., how much signal there is). The proxy for this that we'll use is the corrected $\mathbf{r^2}$.

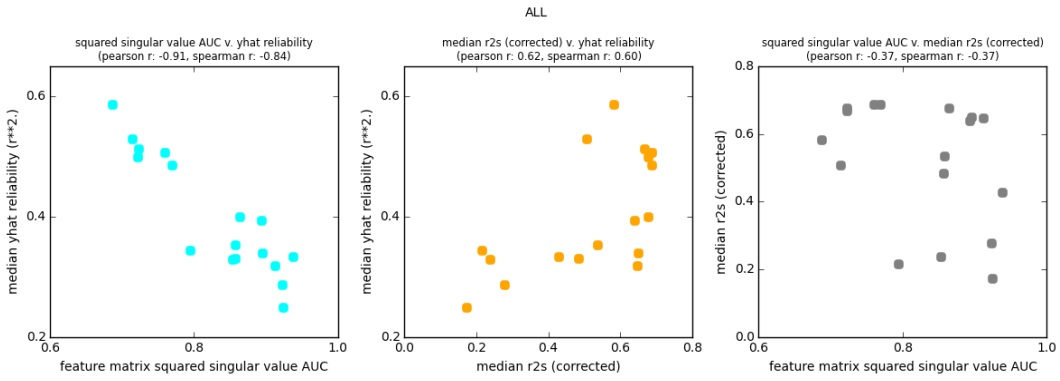

Figure 2: Scatters relating: (1) $\hat{\mathbf{y}}$ reliability, (2) AUC of cumulative squared singular values (q) for different layers, and (3) uncorrected r^2 values. Note the strong correlation between squared singular value AUC and $\hat{\mathbf{y}}$ reliability (leftmost plot).

Empirically, there's a decent amount of variance in the $\hat{\mathbf{y}}$ reliabilities. We address which of the two components of the $\hat{\mathbf{y}}$ reliability are affecting this – the contribution of the $\hat{\mathbf{y}}_{noiseless}$ versus the contribution of the term due to the differences in the feature matrices $\boldsymbol{epsilon}$. We compare the $\hat{\mathbf{y}}$ for 17 different feature sets with either: (1) the AUC of the singular value spectra for each feature set; or (2) the corrected r2s of those features for the data (as a proxy for the $\hat{\mathbf{y}}_{noiseless}$). You see 80% of variance of $\hat{\mathbf{y}}$ reliability is explained by singular value AUC, in comparison with only 40% of variance explained by the $\mathbf{r^2}$. Thus, it appears that at least in these data there is a strong relationship between the reliability of the predictions and the measure of the flatness of the singular value spectrum.

4 Conclusions, limitations, and future directions

Here we've:

- Noted the connection between squared error in SLT and variance explained measures.
- Shown that the maximizer of the squared correlation is the minimizer of the squared error, seeing along the way that maximizing the squared correlation reduces to an eigenvalue problem where the $\mathbf{r}^2$ is the only nonzero eigenvalue and the optimal weights are the corresponding eigenvector.
- Seen that $\mathbf{r}^2$ values computed from data with noise are downwardly biased and how to correct for that bias.
- Noticed that the magnitude of the downward bias due to the reliability of the prediction $\hat{\mathbf{y}}$ can vary across feature sets, and this variance is related to the homogeneity or heterogeneity of the singular values of the features matrix $\mathbf{X}$.
- Observed that this differential downward bias due to the singular value spectrum of the feature set appears to influence the reliability of predictions in practice.

In terms of future directions, it would be nice to extend the discussion of the singular values from the vanilla regression case to the more common case of regularized regression. (Extending to L2-regularization would be straightforward.) It would also be useful to run simulations to further characterize the size of this differential downward bias when we can independently vary the quality of the noiseless fit and the shape of the singular value spectrum for a given feature matrix.

References

[1] Haefner & Cumming. NIPS, 2009. [2] Sahani & Linden. NIPS, 2003. [3] David & Gallant. Network, 2005. [4] Nishimoto & Gallant. J Neuro, 2011. [5] Huth et al., Neuron, 2012. [6] Santoro et al., PLoS Comp Bio, 2014. [7] Huth et al., Nature, 2016.