Bias-Variance Decomposition

Mohammad Emtiyaz Khan
EPFL

Oct 6, 2015

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Motivation

In ridge regression, we observe a typical behaviour for train and test errors with respect to model complexity. Bias-variance decomposition explains this behaviour.

IMPORTANT: Throughout this lecture, we assume that we have a training dataset of size N.

Training errors

Given a training dataset $\mathcal{D}_{tr}$ of size N, we define train error as follows:

$$trErr(\mathcal{D}_{tr}) := \frac{1}{N}\sum_{n=1}^{N}(y_n - f(\mathbf{x}_n))^2$$

where f is the *learned* regression function from $\mathcal{D}_{tr}$. This is the train error we can compute in practice.

We can define the expected train error by taking expectation over all possible training datasets of size N.

$$\overline{trErr} := \mathbb{E}_{\mathcal{D}_{tr}}\ [trErr(\mathcal{D}_{tr})]$$

Since we have finite data, we do not know this error.

Test errors

Given a test-pair $\mathcal{D}_{te} = \{y_*, \mathbf{x}_*\}$, we can define the in-sample test error,

$$teErr_*(\mathcal{D}_{te}, \mathcal{D}_{tr}) := [y_* - f(\mathbf{x}_*)]^2$$

We can compute this error for samples in the validation set.

The test error is obtained by taking expectation over the test-data.

$$teErr(\mathcal{D}_{tr}) := \mathbb{E}_{\mathcal{D}_{te}}[\{y_* - f_{lse}(\mathbf{x}_*)\}^2]$$

We do not know this quantity in practice and our goal is to estimate it as accurately as possible.

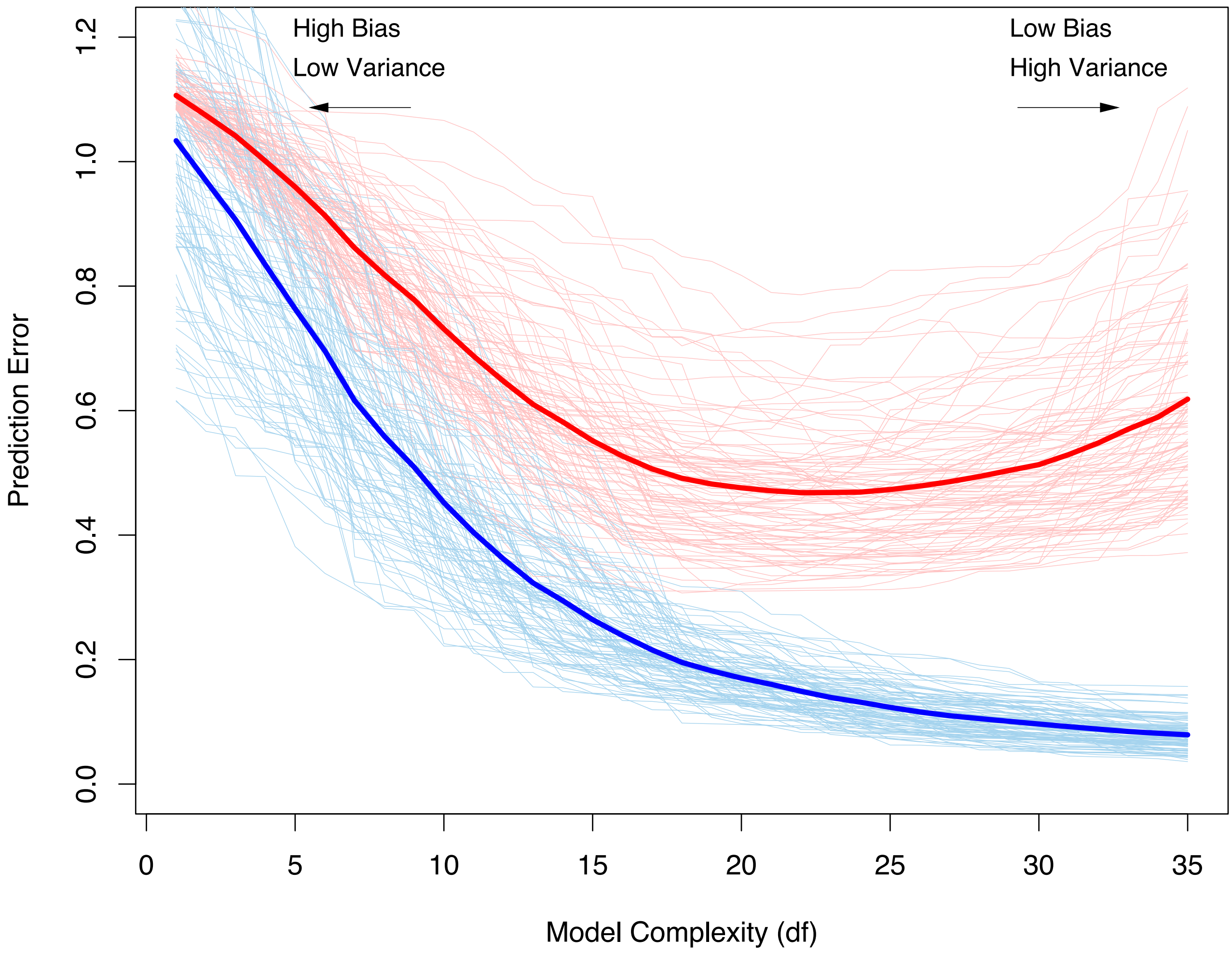
Note that the test error depends on the $\mathcal{D}_{tr}$. We can further average over all possible training data of size N to get the expected test error.

$$\overline{teErr} := \mathbb{E}_{\mathcal{D}_{tr}}\mathbb{E}_{\mathcal{D}_{te}}\ [\{y_* - f_{lse}(\mathbf{x}_*)\}^2]$$

This error tells us the optimal error obtained with a dataset of size N. Obviously, this is not known either.

Error vs Model Complexity

This figure is taken from the book by Hastie, Tibshirani, and Friedman (HTF) Chapter 7. A total of 100 training sets $\mathcal{D}_{tr}$ with $N = 50$ were used. Light blue curves show the train error and the thick blue curve shows the expected train error. Light red curves show the test error and the thick red curve shows the expected test error.



Bias-variance decomposition

We will show four key results using Bias-variance decomposition.

Let us assume $f_{true}(\mathbf{x}_n)$ is the true model and the observations are given as follows:

$$y_n = f_{true}(\mathbf{x}_n) + \epsilon_n$$

where the ϵ_n are i.i.d. with zero mean and variance σ^2.

Note that f_{true} can be nonlinear and ϵ_n doesn't have to be Gaussian.

We denote the least-square estimator by $f_{lse}(\mathbf{x}_*) = \widetilde{\mathbf{x}}_*^T \boldsymbol{\beta}_{lse}$. For this derivation, we will assume that $\mathbf{x}_*$ is fixed, although it is straightforward to generalize this.

Both bias and variance contribute to expected test error

The expected test error for the least-squares estimate can be written as follows,

$$\begin{aligned}
\overline{teErr} &:= \mathbb{E}_{\mathcal{D}_{tr},\mathcal{D}_{te}}[(y_* - f_{lse})^2] \\
&= \mathbb{E}_{y_*,\beta_{lse}}[(y_* - f_{lse})^2] \\
&= \sigma^2 + \mathbb{E}_{\beta_{lse}}(f_{lse} - f_{true})^2 \\
&= \sigma^2 + \mathbb{E}_{\beta_{lse}}\{[f_{lse} - \mathbb{E}_{\beta_{lse}}(f_{lse})]^2\} \\
&\qquad + [f_{true} - \mathbb{E}(f_{lse})]^2
\end{aligned}$$

Both model bias and estimation bias are important

Ridge regression increases estimation bias while reducing variance

Increasing model complexity increases test error

For the least-squares estimate, you can show that

$$\overline{teErr} := \sigma^2 + \frac{D}{N}\sigma^2 + \mathbb{E}_{\beta_{lse}}[f_{true} - \mathbb{E}(f_{lse})]^2$$

With increasing D, the variance of the estimator increases. See HTF Page 224.

Additional Notes

A figure to illustrate bias-variance tradeoff

This is taken from HTF Chapter 7.

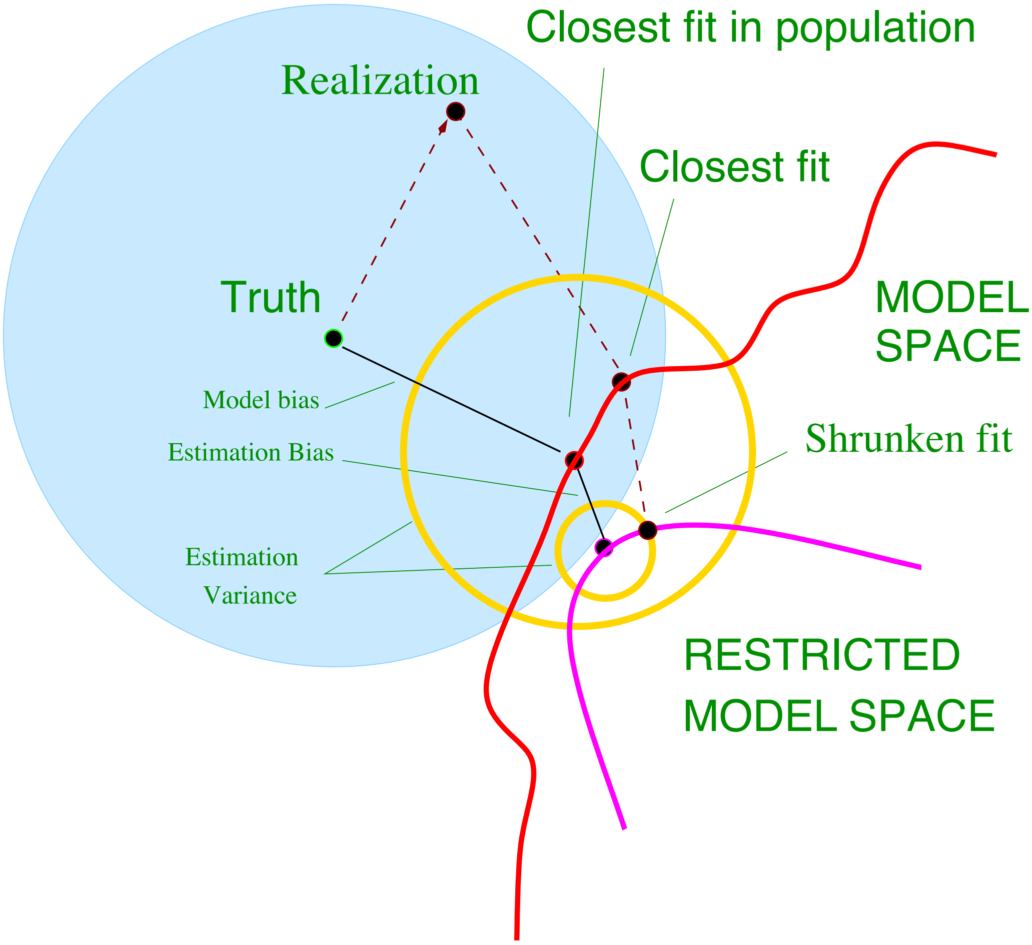


FIGURE 7.2. *Schematic of the behavior of bias and variance. The model space is the set of all possible predictions from the model, with the "closest fit" labeled with a black dot. The model bias from the truth is shown, along with the variance, indicated by the large yellow circle centered at the black dot labeled "closest fit in population." A shrunken or regularized fit is also shown, having additional estimation bias, but smaller prediction error due to its decreased variance.*

Cross validation and generalization error

Cross validation estimates the expected train and test error. If the learning curve is steep, then cross validation overestimates the true objective function. Please read HTF Section 7.10.

Testing regression methods

You can learn about the following from HTF Page 47-49 and Chapter 3 of JWHT (see course infromation for book details).

- R^2 and RMSE goodness of fit.
- Significance and hypothesis testing.
- Confidence interval, standard error, p-value, t-statistics etc.
- Feature engineering: transformations of input variables, adding interactions, dummy encoding of binary and categorical variables, missing values.

To do

- Clearly understand the definition of expected test and train errors (do the exercise).
- Read HTF Section 7.2 and 7.3. This may not be an easy read.
- Revise the derivation of bias-variance decomposition.
- Visualize bias-variance decomposition during labs.
- For testing regression methods, read Page 47-49 from HTF and Chapter 3 of JWHT (see course infromation for book details).