Lecture 6, Part I: Effective Degrees of Freedom (motivation, definition, examples)

Text reference: Shalizi Sec 1.5.3.2 and Sec 3.4.3

Review: Kernel Smoothers, Splines and Tuning Parameters

- A linear smoother has the form $\widehat{r}(x) = \sum_i \ell_i(x) Y_i$.
  - In particular, $\widehat{\vec{Y}} = L\,\vec{Y}$ at the observed X_i's.
- For a kernel smoother

$$\ell_i(x) = \frac{K((x - X_i)/h)}{\sum_{i=1}^{n} K((x - X_i)/h)}$$

  - The kernel smoother is a weighted average of the Y_i's in the neighborhood of x.
  - The key choice is the smoothing parameter h.
- A regression or smoothing spline is like linear regression with the X_{ij}'s replaced by the basis functions $B_j(X_i)$

$$\widehat{r}(x) = \sum_j \widehat{\beta}_j B_j(x).$$

  - The basis functions are obtained using B-splines determined by the observed X_i's.
  - For **regression splines**, the key choice is the placement and number of B-splines; $\widehat{\beta}$ is obtained via least squares.
  - For **smoothing splines**, the key choice is the smoothing parameter

λ, which penalizes the function for excess curvature; $\widehat{\beta}$ is obtained via penalized least squares.

R Demo 6.1

(a) Simulate iid data according to the model $Y = X \cdot \sin(X) + \epsilon$, where $\epsilon \sim N(0,1)$ and $X \sim U(0, 6\pi)$.

(b) Fit a *kernel regression* to the data. Use a Gaussian kernel and manually choose the kernel bandwidth; for example, try h=5, 1, and 0.1. What do you see?

(c) Choose six evenly spaced knots. Generate B-spline basis. Fit *regression splines* to the data. Comment on the choice of knots. How about using natural splines?

(d) Fit *smoothing splines* to the data. Choose $\lambda = 0.01$, $2 \cdot 10^{-5}$, and 10^{-20}. What do you see?

To choose the tuning parameters in these estimators, we can use techniques such as K-fold cross-validation, leave-one-out cross-validation (LOOCV), Mallow's C_p-statistics, and generalized cross-validation (GCV). As we shall see, these approaches are closely related for linear smoothers....

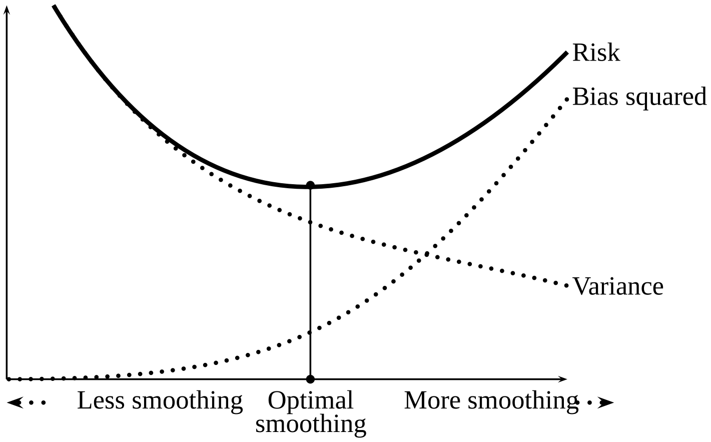

Figure 1: The familiar bias–variance tradeoff in curve estimation. But how do we measure the amount of smoothing?

Effective Degrees of Freedom

Motivation

- So far we have seen several methods for estimating the underlying regression function $r(x) = \mathbb{E}[Y|X = x]$: linear regression, k-nearest-neighbors, kernel smoothing and splines. Often, in a predictive setting, we want to compare (estimates of) test error between such methods, in order to choose the best method.

- We have learned one method for model selection: cross-validation. But in a sense, comparing cross-validation curves between methods is not straightforward, because each curve is parametrized differently. For example: linear regression has no tuning parameters and we report just one error value; k-nearest-neighbors give us an error curve with respect to the number of nearest neighbors k; kernel regression gives us an error curve over the bandwidth h; smoothing splines gives us an error curve over the smoothing parameter λ.

- So what does it actually mean to choose kernel regression with $h = 1.5$, over say, k-nearest-neighbors with $k = 10$ or smoothing splines with $\lambda = 0.417$? That is, does the $h = 1.5$ model from kernel regression correspond to a more of less complex estimate than the $k = 10$ model from k-nearest-neighbors, or the $\lambda = 0.417$ model from smoothing splines?
- The notion of *degrees of freedom* gives us a way of precisely making this comparison. Roughly speaking, the degrees of freedom of a fitting procedure (like kernel regression with $h = 1.5$, or k-nearest-neighbors with $k = 10$) describes the **effective number of parameters** used by this procedure, and hence provides a quantitive measure of the **complexity** of an estimator.
- Keeping track of degrees of freedom therefore saves us from unsuspectingly comparing a procedure that uses say, 10 effective parameters to another one that uses 100.

Definition of Effective Degrees of Freedom

Even though the concept it represents is quite broad, degrees of freedom has a rigorous definition. Consider the following regression setting, where we condition on the values of the random predictors (that is, we treat the predictor measurements x_i, $i = 1, \ldots, n$ as fixed):

We define the **effective degrees of freedom** of the estimator $\widehat{r}$ by:

To reiterate: this covariance treats only Y_i, $i = 1, \ldots n$ as random (and not x_i, $i = 1, \ldots n$).

Intuition: The definition of degrees of freedom above looks at the amount of covariance between each point Y_i and its corresponding fitted values $\widehat{Y}_i$. We sum these contributions over $i = 1, \ldots n$, and divide the result by σ^2 (dividing by σ^2 gets rid of the dependence of the sum on the marginal error variance).

In some situations, it is helpful to write the definition of degrees of freedom *in matrix notation*:

Examples

To get a sense for degrees of freedom, it helps to work through several basic examples.

Example 1: Simple Average Estimator

Consider $\widehat{\vec{Y}}^{\text{ave}} = (\bar{Y}, \ldots \bar{Y})^T$, where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$. Then, the effective number of parameters is given by

Think: does this result make sense to you?

Example 2: Identity Estimator

Consider $\widehat{\vec{Y}}^{\text{id}} = (Y_1, \dots Y_n)^T$. Then, the effective number of parameters is given by

Think: does this result make sense to you?

Example 3: Linear Smoothers

Recall that a **linear smoother** has the form

$$\widehat{r}^{\text{linsm}}(x) = \sum_{j=1}^{n} w(x, x_j) \cdot Y_j.$$

This means that

$$\widehat{Y}_i^{\text{linsm}} = \widehat{r}^{\text{linsm}}(x_i) = \sum_{j=1}^{n} w(x_i, x_j) \cdot Y_j,$$

i.e., we can write

$$\widehat{\vec{Y}}^{\text{linsm}} = L\vec{Y},$$

for the matrix $L \in \mathbb{R}^{n \times n}$ defined as $L_{ij} = w(x_i, x_j)$. Calculating degrees of freedom, in matrix form, we get

This is a very useful formula! Recall that linear regression, k-nearest-neighbors, kernel regression and splines are all linear smoothers. We can calculate the degrees of freedom for all of these estimators by simply summing these weights. As concrete examples, we consider linear regression and k-nearest-neighbors regression:

(3a) Linear Regression

Consider the fitted values from linear regression of Y on X,

$$\widehat{\vec{Y}}^{\text{linreg}} = \mathbb{X}\widehat{\beta} = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\vec{Y} = H\vec{Y}.$$

Here $\mathbb{X}$ is the $n \times q$ predictor or “design” matrix (with observation x_i along its ith row). Then, relying on the above result, we get

Think: does this result make sense to you?

(3b) k-Nearest-Neighbors Regression

Consider k-nearest-neighbors regression with some fixed value of $k \geq 1$. Recall that here

$$w(x, x_j) = \begin{cases} \frac{1}{k} & \text{if } x_j \text{ is one of the } k \text{ closest points to } x \\ 0 & \text{else.} \end{cases}$$

Therefore, the effective number of parameters is given by

Think: what happens for small k? Large k? Does this match your intuition for the complexity of the k-nearest-neighbors fit?

Estimating the Effective Degrees of Freedom

How do we estimate the degrees of freedom in more complex settings?

Degrees of freedom cannot always be calculated analytically, as we did above. In fact, at large, it is rather uncommon for this to be the case. As an extreme example, if the fitting procedure $\widehat{r}$ is just a black box (e.g., just an R function whose mechanism is unknown), then we would have no way of analytically counting its degrees of freedom. However, the expression of degrees of freedom is still well-defined for any fitting procedure $\widehat{r}$; to get an estimate of its degrees of freedom, we can for example estimate the covariance terms $\text{Cov}(\widehat{Y}_i, Y_i)$ via the bootstrap method.

Question: Which bootstrap sampling scheme would be appropriate for our application and why? (Think about what is random and what is fixed in the definition of the effective degrees of freedom)

That is, after fitting $\widehat{Y}_i = \widehat{r}(x_i)$, $i = 1, \ldots n$ using the original sample (X_i, Y_i), $i = 1, \ldots n$, we record the (empirical) residuals

$$\widehat{\epsilon}_i = Y_i - \widehat{Y}_i, \quad i = 1, \ldots n.$$

Then, form a bootstrap sample of size n by resampling the residuals with replacement and recompute the regression fit on the bootstrap sample:

At the end, we approximate the covariance of $\widehat{Y}_i$ and Y_i by the empirical covariance between $\widehat{Y}^*_{i,b}$ and $Y^*_{i,b}$ averaged over $b = 1, \dots B$, i.e.,

$$\mathrm{Cov}(\widehat{Y}_i, Y_i) \approx \frac{1}{B} \sum_{b=1}^{B} \Big(\widehat{Y}^*_{i,b} - \frac{1}{B} \sum_{r=1}^{B} \widehat{Y}^*_{i,r} \Big) \cdot \Big(Y^*_{i,b} - \frac{1}{B} \sum_{r=1}^{B} Y^*_{i,r} \Big),$$

and sum this over $i = 1, \dots n$ to yield our bootstrap estimate for degrees of freedom:

$$\mathrm{df}(\widehat{r}) \approx \frac{1}{\sigma^2} \sum_{i=1}^{n} \left(\frac{1}{B} \sum_{b=1}^{B} \Big(\widehat{Y}^*_{i,b} - \frac{1}{B} \sum_{r=1}^{B} \widehat{Y}^*_{i,r} \Big) \cdot \Big(Y^*_{i,b} - \frac{1}{B} \sum_{r=1}^{B} Y^*_{i,r} \Big) \right).$$

(For simplicity, you can assume that σ^2 is known; otherwise, we would have to estimate it too.)

R Demo 6.2: Estimating Degrees of Freedom via the Residual Bootstrap

a. Create a data example for linear regression. Use simulation to estimate the effective degrees of freedom.

b. Estimate the degrees of freedom using resampling of residuals. Compare with the result in part a and discuss.

c. *(optional)* Estimate the degrees of freedom using resampling of cases. Compare with the results in parts a and b and discuss.