7.5 Smoothing Splines

Zongyi Liu

2023-06-03

7.5 Smoothing Splines

7.5.1 An Overview of Smoothing Splines

In fitting a smooth curve to a set of data, what we really need to do is to find some functions, say $g(x)$, that fits the observed data well (which means we want the $RSS = \sum\_i = 1^n(y_i - g(x_i))^2$ to be small.

There are many ways to help we ensure that g is smooth, and a natural approach is to find the function g that minimizes

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt$$

where λ is a nonnegative tuning parameter. The function g that minimizes the equation is known as a smoothing spline.

The meaning of the equation above takes the "Loss+Penalty" formulation that we encounter in the context of ridge regression and the lasso before.

The term $\sum_{i=1}^{n}(y_i - g(x_i))^2$ is a **loss function** that encourages g to fit the data well, and the term $\lambda \int g''(t)^2 dt$ is a penalty term that penalizes the variability in g.

We should notice that there is are derivative terms. The first derivative $g'(t)$ measures the slope of a function at t, and the second derivative corresponds to the amount by which the slope is changing. Hence, broadly speaking, the second derivative of a function is a measure of its **roughness**: it is large in absolute value if $g(t)$ is very wiggly near t, and it is close to zero otherwise.

When $\lambda = 0$, then the penalty term in equation above has no effect, and so the function g will be very jumpy and will exactly interpolate the training observations. When λß∞, g will be perfectly smooth—it will just be a straight line that passes as closely as possible to the training points.

7.5.2 Choosing the Smoothing Parameter λ

Choosing the parameter λ is important since it controls the roughness of the smoothing spline, and hence the **effective degrees of freedom**. It is possible to show that as λ increases from 0 to ∞, the effective degrees of freedom, which we write df_λ, decrease from n to 2.

Here we have a new definition of effective degrees of freedom as a measure if the flexibility of the smoothing spline.

We can write

$$\hat{\mathbf{g}}_\lambda = \mathbf{S}_\lambda \mathbf{y},$$

where $\hat{g}_\lambda$ is the solution to the minimization equation for a particular choice of
lambda, in other words, it is a n-vector containing the fitted values of the smoothing spline at the training points $x_1, \ldots, x_n$.

The equation of $\hat{g}_\lambda$ indicates that the vector of fitted values when applying a smoothing spline to the data can be written as a $n * n$ matrix S_λ (for which there is a formula) times the response vector $\mathbf{y}$. Then we can get the dof to be defined as

$$df_\lambda = \sum_{i=1}^{n} \{\mathbf{S}_\lambda\}_{ii},$$

the sum if the diagonal elements of the matrix S_λ.

In fitting a smoothing spline, we do not need to select the number or location of the knots, instead, we have another problem, we need to choose the value of
lambda. We need to choose the λ to make the CV-RSS as small as possible. It turns out that the leave-one-out cross-validation error (LOOCV) can be computed very efficiently for smoothing splines, with essentially the same cost as computing a single fit, using this formula:

$$\text{RSS}_{cv}(\lambda) = \sum_{i=1}^{n} (y_i - \hat{g}_\lambda^{(-i)}(x_i))^2 = \sum_{i=1}^{n} \left[\frac{y_i - \hat{g}_\lambda(x_i)}{1 - \{\mathbf{S}_\lambda\}_{ii}} \right]^2.$$

The notation $\hat{g}_\lambda^{(-i)}(x_i)$ indicates that the fitted value for this smoothing spline evaluated at x_i, where the fit uses all of the training observations except for the ith one (x_i, y_i). We can get a smoothing spline fits to the `Wage` data, and plot it as below: