1 Definition of Nadaraya-Watson Estimator

Consider the nonlinear model \( y_i = f(x_i) + \epsilon_i \), where \( f(\cdot) \) is an unknown nonlinear function. Suppose that given \( x_0 \), we would like to only estimate \( f(x_0) \). One estimator that can be used is

\[
\hat{f}_0[x_0] = \frac{\sum_{i=1}^{n} K(\|x_i - x_0\|/h) \cdot y_i}{\sum_{i=1}^{n} K(\|x_i - x_0\|/h)},
\]

where \( K(\cdot) \) is a kernel function. This estimator is known as the Nadaraya-Watson estimator, and it was one of the earlier techniques developed for nonparametric regression.

2 Alternative Characterizations

It turns out that we can characterize this estimator through multiple formulations. The first is as the following M-estimator

\[
\hat{f}_0[x_0] = \min_{\beta_0} \left\| W_{1/2}^{-1} (Y - 1_n \beta_0) \right\|^2 = \min_{\beta_0} \sum_{i=1}^{n} K(\|x_i - x_0\|/h) \cdot (y_i - \beta_0)^2.
\]

A second characterization is as the mean with weights \( \{K(\|x_1 - x_0\|/h), \ldots, K(\|x_n - x_0\|/h)\} \) of points \( \{y_1, \ldots, y_n\} \).

3 Small Denominators

The denominator of the Nadaraya-Watson estimator is worth examining. Define

\[
\hat{g}(x_0) = \frac{1}{nh^p} \sum_{i=1}^{n} K(\|x_i - x_0\|/h),
\]

and note that \( \hat{g}(x_0) \) is an estimate of the probability density function of \( x_i \) at the point \( x_0 \). This is known as a kernel density estimate (KDE), and the intuition is that this is a smooth version of a histogram of the \( x_i \).
The denominator of the Nadaraya-Watson estimator is a random variable, and technical problems occur when this denominator is small. This can be visualized graphically. The traditional approach to dealing with this is trimming, in which small denominators are eliminated. The trimmed version of the Nadaraya-Watson estimator is

\[
\hat{\beta}_0[x_0] = \begin{cases} 
\frac{\sum_{i=1}^n K(||x_i - x_0||/h) \cdot y_i}{\sum_{i=1}^n K(||x_i - x_0||/h)}, & \text{if } \sum_{i=1}^n K(||x_i - x_0||/h) > \mu \\
0, & \text{otherwise}
\end{cases}
\]

One disadvantage of this approach is that if we think of \( \hat{\beta}_0[x_0] \) as a function of \( x_0 \), then this function is not differentiable in \( x_0 \).

4 \textbf{L2-Regularized Nadaraya-Watson Estimator}

A new approach is to define the L2-regularized Nadaraya-Watson estimator

\[
\hat{\beta}_0[x_0] = \frac{\sum_{i=1}^n K(||x_i - x_0||/h) \cdot y_i}{\lambda + \sum_{i=1}^n K(||x_i - x_0||/h)},
\]

where \( \lambda > 0 \). If the kernel function is differentiable, then the function \( \hat{\beta}[x_0] \) is always differentiable in \( x_0 \).

The reason for this name is that under the M-estimator interpretation of Nadaraya-Watson estimator, we have that

\[
\hat{\beta}[x_0] = \arg \min_{\beta_0} \|W_h^{1/2}(Y - 1_n \beta_0)\|_2^2 + \lambda \|\beta_0\|_2^2 = \arg \min_{\beta_0} \sum_{i=1}^n K(||x_i - x_0||/h) \cdot (y_i - \beta_0)^2 + \lambda \beta_0^2.
\]

Lastly, note that we can also interpret this estimator as the mean with weights

\[
\{\lambda, K(||x_1 - x_0||/h), \ldots, K(||x_n - x_0||/h)\}
\]

of points \{0, y_1, \ldots, y_n\}.