1 Small Denominators in Nadaraya-Watson

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 \( x_0 \).

2 \( L^2 \)-Regularized Nadaraya-Watson Estimator

A new approach is to define the \( L^2 \)-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
\[ \hat{\beta}[x_0] = \arg \min_{\beta_0} \|W_{1/2}^{-1/2}(Y - 1_n \beta_0)\|_2^2 + \lambda \|\beta_0\|_2^2 = \arg \min_{\beta_0} \sum_{i=1}^{n} K(\|x_i - x_0\|/h) \cdot (y_i - \beta_0)^2 + \lambda \beta_0^2. \]
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\}.

3 Partially Linear Model

Recall the following partially linear model
\[ y_i = x_i' \beta + g(z_i) + \epsilon_i = f(x_i, z_i; \beta) + \epsilon_i, \]
where \( y_i \in \mathbb{R}, x_i, \beta \in \mathbb{R}^p, z_i \in \mathbb{R}^q, g(\cdot) \) is an unknown nonlinear function, and \( \epsilon_i \) are noise. The data \( x_i, z_i \) are i.i.d., and the noise has conditionally zero mean \( \mathbb{E}[\epsilon_i | x_i, z_i] = 0 \) with unknown and bounded conditional variance \( \mathbb{E}[\epsilon_i^2 | x_i, z_i] = \sigma^2(x_i, z_i) \). This model is known as a partially linear model because it consists of a (parametric) linear part \( x_i' \beta \) and a nonparametric part \( g(z_i) \). One can think of the \( g(\cdot) \) as an infinite-dimensional nuisance parameter, but in some situations this function can be of interest.

4 Nonparametric Approach

Suppose we were to compute a LLR of this model at an arbitrary point \( x_0, z_0 \) within the support of the \( x_i, z_i \):
\[
\begin{bmatrix} \hat{\beta}_0[x_0, z_0] \\ \hat{\beta}[x_0, z_0] \\ \hat{\eta}[x_0, z_0] \end{bmatrix} = \arg \min_{\hat{\beta}_0, \hat{\beta}, \hat{\eta}} \left\| W_h^{1/2} \left( Y - [1_n \ X_0 \ Z_0] \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta} \\ \hat{\eta} \end{bmatrix} \right) \right\|_2^2,
\]
where \( X_0 = X - x_0' 1_n, Z_0 = Z - z_0' 1_n \), and
\[
W_h = \text{diag} \left( K \left( \frac{1}{h} \left\| \begin{bmatrix} x_1 \\ z_1 \end{bmatrix} - \begin{bmatrix} x_0 \\ z_0 \end{bmatrix} \right\| \right), \ldots, K \left( \frac{1}{h} \left\| \begin{bmatrix} x_n \\ z_n \end{bmatrix} - \begin{bmatrix} x_0 \\ z_0 \end{bmatrix} \right\| \right) \right).
\]
By noting that \( \nabla_x f = \beta \), one estimate of the parametric coefficients is \( \hat{\beta} = \hat{\beta}[x_0, z_0] \). That is, in principle, we can use a purely nonparametric approach to estimate the parameters of this partially linear model. However, the rate of convergence will be \( O_p(n^{-2/(p+q+4)}) \). This is much slower than the parametric rate \( O_p(1/\sqrt{n}) \).

5 Semiparametric Approach

Ideally, our estimates of \( \beta \) should converge at the parametric rate \( O_p(1/\sqrt{n}) \), but the \( g(z_i) \) term causes difficulties in being able to achieve this. But if we could somehow subtract out
this term, then we would be able to estimate $\beta$ at the parametric rate. This is the intuition behind the semiparametric approach. Observe that

$$E[y_{i} | z_{i}] = E[x_{i}' \beta + g(z_{i}) + \epsilon_i | z_i] = E[x_{i} | z_{i}]' \beta + g(z_{i}),$$

and so

$$y_{i} - E[y_{i} | z_{i}] = (x_{i}' \beta + g(z_{i}) + \epsilon_i) - E[x_{i} | z_{i}]' \beta - g(z_{i}) = (x_{i} - E[x_{i} | z_{i}])' \beta + \epsilon_i.$$

Now if we define

$$\hat{Y} = \begin{bmatrix} E[y_{1} | z_{1}] \\ \vdots \\ E[y_{n} | z_{n}] \end{bmatrix}$$

and

$$\hat{X} = \begin{bmatrix} E[x_{1} | z_{1}]' \\ \vdots \\ E[x_{n} | z_{n}]' \end{bmatrix}$$

then we can define an estimator

$$\hat{\beta} = \arg \min_{\beta} \| (Y - \hat{Y}) - (X - \hat{X}) \beta \|^2 = ((X - \hat{X})'(X - \hat{X}))^{-1}((X - \hat{X})'(Y - \hat{Y})).$$

The only question is how can we compute $E[x_{i} | z_{i}]$ and $E[y_{i} | z_{i}]$? It turns out that if we compute those values with the trimmed version of the Nadaraya-Watson estimator, then the estimate $\hat{\beta}$ converges at the parametric rate under reasonable technical conditions. Intuitively, we would expect that we could alternatively use the $L^2$-regularized Nadaraya-Watson estimator, but this has not yet been proven to be the case.