1 Kernel Estimators

1.1 Convergence Rate

There is one point of caution to note regarding the use of kernel density estimation (and any other nonparametric density estimators like the histogram). Suppose we have data \( x_i \) from a multivariate jointly Gaussian distribution \( \mathcal{N}(\mu, \Sigma) \). The we can use the sample mean vector estimate

\[
\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

and sample covariance matrix estimate

\[
\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu})(x_i - \hat{\mu})^t.
\]

to estimate the distribution as \( \mathcal{N}(\hat{\mu}, \hat{\Sigma}) \). The estimation error of this approach is \( O_p(d/n) \), and so the error increases linearly in dimension. In contrast, the estimation error of the kernel density estimate is \( O_p(n^{-2/(d+4)}) \). This means that the estimation error is exponentially worse in terms of dimension \( d \), and this is an example of the curse of dimensionality in the statistics context. In fact, the estimation error of the kernel density estimate is typically \( O_p(n^{-2/(d+4)}) \) when applied to a general distribution.

1.2 Nadaraya-Watson Estimator

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

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

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.

1.3 Example: Telephone Call Data

Suppose the lengths of calls at a call center are

\[
x_i = \{0.66, 0.05, 0.27, 1.26, 1.51, 0.38, 1.79, 0.94, 0.48, 0.89\}.
\]
And imagine that we conduct a survey after each call where we ask the customer to rate their satisfaction with the call. Suppose the corresponding satisfaction levels (1 = very dissatisfied, 2 = somewhat dissatisfied, 3 = neutral, 4 = somewhat satisfied, and 5 = very satisfied) are

\[ y_i = \{3, 5, 4, 1, 1, 3, 2, 5, 4, 2\}. \]

Q: Suppose we choose \( h = 0.5 \), and that we use the Epanechnikov kernel. Estimate the satisfaction level for a telephone call of length 0.7 using the Nadaraya-Watson estimator.

A: We first compute the quantities \( K \left( \frac{u - x_i}{h} \right) \) for each data point. We have

\[
\begin{align*}
K((0.7 - 0.66)/0.5) &= K(0.08) = 3/4 \cdot (1 - 0.08^2) = 0.7452 \\
K((0.7 - 0.05)/0.5) &= K(1.3) = 0 \\
K((0.7 - 0.27)/0.5) &= K(0.86) = 3/4 \cdot (1 - 0.86^2) = 0.1953 \\
K((0.7 - 1.26)/0.5) &= K(-1.12) = 0 \\
K((0.7 - 1.51)/0.5) &= K(-1.62) = 0 \\
K((0.7 - 0.38)/0.5) &= K(0.64) = 3/4 \cdot (1 - 0.64^2) = 0.4428 \\
K((0.7 - 1.79)/0.5) &= K(-2.18) = 0 \\
K((0.7 - 0.94)/0.5) &= K(-0.48) = 3/4 \cdot (1 - 0.48^2) = 0.5772 \\
K((0.7 - 0.48)/0.5) &= K(0.44) = 3/4 \cdot (1 - 0.44^2) = 0.6048 \\
K((0.7 - 0.89)/0.5) &= K(-0.38) = 3/4 \cdot (1 - 0.38^2) = 0.6417
\end{align*}
\]

Finally, we compute

\[
\hat{g}(0.7) = \frac{\sum_{i=1}^{n} K(\|x_i - x_0\|/h) \cdot y_i}{\sum_{i=1}^{n} K(\|x_i - x_0\|/h)} = \frac{0.7452 \cdot 3 + 0.1953 \cdot 4 + 0.4428 \cdot 3 + 0.5772 \cdot 5 + 0.6048 \cdot 4 + 0.6417 \cdot 2}{0.7452 + 0.1953 + 0.4428 + 0.5772 + 0.6048 + 0.6417} = 3.41.
\]

For reference, the scatter plot and the full curve \( \hat{g}(x) \) estimated by Nadaraya-Watson are shown below. The full curve is generally calculated using a computer.
The Epanechnikov kernel is not differentiable, and so the estimated curve is not differentiable. An example of a kernel function that is differentiable is the biweight kernel, which is defined as

$$K(u) = \frac{15}{16} \cdot (1 - u^2)^2 1(\vert u \vert \leq 1).$$

If we use the biweight kernel, then the full curve estimated by Nadaraya-Watson is differentiable.
1.4 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 \( \hat{f}(x_0) \) 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 \textit{trimming}, in which small denominators are eliminated. The trimmed version of the Nadaraya-Watson estimator is

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

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

1.5 Example: Telephone Call Data

Suppose the lengths of calls at a call center are

\[ x_i = \{0.66, 0.05, 0.27, 1.26, 1.51, 0.38, 1.79, 0.94, 0.48, 0.89\}. \]

And imagine that we conduct a survey after each call where we ask the customer to rate their satisfaction with the call. Suppose the corresponding satisfaction levels (1 = very dissatisfied, 2 = somewhat dissatisfied, 3 = neutral, 4 = somewhat satisfied, and 5 = very satisfied) are

\[ y_i = \{3, 5, 4, 1, 1, 3, 2, 5, 4, 2\}. \]

Then the trimmed Nadaraya-Watson estimator using the biweight kernel with bandwidth \( h = 0.5 \) and threshold \( \mu = 0.01 \) is:
1.6 \textit{L2-Regularized Nadaraya-Watson Estimator}

A new approach is to define the \textit{L2}-regularized Nadaraya-Watson estimator

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

where \( \lambda > 0 \). If the kernel function is differentiable, then the function \( \hat{g}(x_0) \) is always differentiable in \( x_0 \). The reason for the name of this estimator is that we have

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

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 \( \{0, y_1, \ldots, y_n\} \).

1.7 \textbf{Example: Telephone Call Data}

Suppose the lengths of calls at a call center are

\[ x_i = \{0.66, 0.05, 0.27, 1.26, 1.51, 0.38, 1.79, 0.94, 0.48, 0.89\}. \]
And imagine that we conduct a survey after each call where we ask the customer to rate their satisfaction with the call. Suppose the corresponding satisfaction levels (1 = very dissatisfied, 2 = somewhat dissatisfied, 3 = neutral, 4 = somewhat satisfied, and 5 = very satisfied) are

\[ y_i = \{3, 5, 4, 1, 3, 2, 5, 4, 2\}. \]

Then the trimmed Nadraya-Watson estimator using the biweight kernel with bandwidth \( h = 0.5 \) and regularization \( \lambda = 0.2 \) is:

This curve is differentiable

## 2 Partially Linear Model

Consider the following model

\[ y_i = x_i' \beta + g(z_i) + \epsilon_i, \]

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 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.

### 2.1 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

$$\hat{\beta} = \arg \min_{\beta} \| (Y - \hat{Y}) - (X - \hat{X}) \beta \|_2^2 = ((X - \hat{X})'(X - \hat{X}))^{-1}((X - \hat{X})'(Y - \hat{Y})).$$

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.

### 2.2 Example: Telephone Call Data

Suppose the lengths of calls at a call center are

$$x_i = \{0.66, 0.05, 0.27, 1.26, 1.51, 0.38, 1.79, 0.94, 0.48, 0.89\}.$$ 

And imagine that we conduct a survey after each call where we ask the customer to rate their satisfaction with the call. Suppose the corresponding satisfaction levels (1 = very dissatisfied, 2 = somewhat dissatisfied, 3 = neutral, 4 = somewhat satisfied, and 5 = very satisfied) are

$$y_i = \{3, 5, 4, 1, 1, 3, 2, 5, 4, 2\}.$$ 

Furthermore, suppose we also record the time of day for each call:

$$t_i = \{18, 19, 17, 13, 11, 19, 16, 12, 16, 10\}.$$
Now imagine that we believe that the model relating the satisfaction level to the length of call is

\[ y = m \cdot x + g(t), \]

where \( m \) is an unknown constant, and \( g(\cdot) \) is an unknown function. Suppose we are interested in estimating \( m \), which gives the sensitivity of satisfaction to the length of call. Then, one natural approach is to use semiparametric estimation.

Suppose we use the L2-regularized Nadaraya-Watson estimator with an Epanechnikov kernel, \( h = 0.5 \), and \( \lambda = 0.2 \). Then we get

\[ \hat{x}_i = \{0.5204, 0.1902, 0.2110, 0.9982, 1.1916, 0.1902, 1.0015, 0.7384, 1.0015, 0.7002\} \]
\[ \hat{y}_i = \{2.3684, 3.5294, 3.1579, 0.7895, 0.7895, 3.5294, 2.6471, 3.9474, 2.6471, 1.5789\}. \]

Computing \( \tilde{x}_i = x_i - \hat{x}_i \) and \( \tilde{y}_i = y_i - \hat{y}_i \), we get

\[ \tilde{x}_i = \{0.1388, -0.1357, 0.0563, 0.2662, 0.3178, 0.1864, 0.7874, 0.1969, -0.5203, 0.1867\} \]
\[ \tilde{y}_i = \{0.6316, 1.4706, 0.8421, 0.2105, 0.2105, -0.5294, -0.6471, 1.0526, 1.3529, 0.4211\}. \]

In this case, \( \hat{m} = \frac{\left(\bar{X} - \hat{X}\right)^{\prime}(X - \hat{X})^{-1}(X - \hat{X})^{\prime}(Y - \hat{Y})}{\bar{x}^2} \). Computing these quantities, we have:

\[ \bar{x}^2 = 0.1212 \]
\[ \bar{xy} = -0.0968. \]

Thus, we get

\[ \hat{m} = \frac{\bar{xy}}{\bar{x}^2} = \frac{-0.0968}{0.1212} = -0.80. \]

For reference, if we had identified a model

\[ y = m \cdot x + b, \]

then the estimate would have been \( \hat{m} = -1.92 \) and \( \hat{b} = 4.6 \). In this example, adjusting the model for the time of day makes a significant difference in our estimate.