5 Conditional Density Estimation

5.1 Estimators

The conditional density of $y_i$ given $X_i = x$ is $f(y \mid x) = f(y, x) / f(x)$. An natural estimator is

$$
\hat{f}(y \mid x) = \frac{\hat{f}(y, x)}{f(x)} = \frac{\sum_{i=1}^{n} K(H^{-1}(X_i - x)) k_{h_0}(y_i - y)}{\sum_{i=1}^{n} K(H^{-1}(X_i - x))}
$$

where $H = \text{diag}(h_1, ..., h_q)$ and $k_h(u) = h^{-1}k(u/h)$. This is the derivative of the smooth NW-type estimator $\tilde{F}(y \mid x)$. The bandwidth $h_0$ smooths in the $y$ direction and the bandwidths $h_1, ..., h_q$ smooth in the $X$ directions.

This is the NW estimator of the conditional mean of $Z_i = k_{h_0}(y - y_i)$ given $X_i = x$.

Notice that

$$
E(Z_i \mid X_i = x) = \int \frac{1}{h_0} k\left(\frac{v - y}{h_0}\right) f(v \mid x) \, dv \\
= \int k(u) f(y - uh_0 \mid x) \, du \\
\simeq f(y \mid x) + \frac{h_0^2 \kappa_2}{2} \frac{\partial^2}{\partial y^2} f(y \mid x).
$$

We can view conditional density estimation as a regression problem. In addition to NW, we can use LL and WNW estimation. The local polynomial method was proposed in a paper by Fan, Yao and Tong (Biometrika, 1996) and has been called the “double kernel” method.

5.2 Bias

By the formula for NW regression of $Z_i$ on $X_i = x$,

$$
E \hat{f}(y \mid x) = E(Z_i \mid X_i = x) + \kappa_2 \sum_{j=1}^{q} h_j^2 B_j(y \mid x) \\
= f(y \mid x) + \kappa_2 \sum_{j=0}^{q} h_j^2 B_j(y \mid x)
$$

where

$$
B_0(y \mid x) = \frac{1}{2} \frac{\partial^2}{\partial y^2} f(y \mid x)
$$

$$
B_j(y \mid x) = \frac{1}{2} \frac{\partial^2}{\partial x_j^2} f(y \mid x) + f(x)^{-1} \frac{\partial}{\partial x_j} f(y \mid x) \frac{\partial}{\partial x_j} f(x), \quad j > 0
$$
as $B_j$ are the curvature of $E(Z_i \mid X_i = x) \approx f(y \mid x)$ with respect to $x_j$. For LL or WNW

$$B_j(y \mid x) = \frac{1}{2} \frac{\partial^2}{\partial x_j^2} f(y \mid x), \quad j > 0$$

The bias of $\hat{f}(y \mid x)$ for $f(y \mid x)$ is $\kappa_2 \sum_{j=0}^q h_j^2 B_j(y \mid x)$.

For the bias to converge to zero with $n$, all bandwidths must decline to zero.

### 5.3 Variance

By the formula for NW regression of $Z_i$ on $X_i = x$,

$$\text{var} \left( \hat{f}(y \mid x) \right) \approx \frac{R(k)^q}{nh_1 \cdots h_q f(x)} \text{var} \left( Z_i \mid X_i = x \right)$$

We calculate that

$$\text{var} \left( Z_i \mid X_i = x \right) = E \left( Z_i^2 \mid X_i = x \right) - (E \left( Z_i \mid X_i = x \right))^2$$

$$\approx \frac{1}{h_0^2} \int k \left( \frac{v - y}{h_0} \right)^2 f(v \mid x) \, dv$$

$$= \frac{1}{h_0} \int k(u)^2 f(y - uh \mid x) \, du$$

$$\approx \frac{R(k) f(y \mid x)}{h_0}.$$ 

Substituting this into the expression for the estimation variance,

$$\text{var} \left( \hat{f}(y \mid x) \right) \approx \frac{R(k)^q}{nh_1 \cdots h_q f(x)} \text{var} \left( Z_i \mid X_i = x \right)$$

$$= \frac{R(k)^q f(y \mid x)}{nh_0 h_1 \cdots h_q f(x)}$$

What is key is that the variance of the conditional density depends inversely upon all bandwidths.

For the variance to tend to zero, we thus need $nh_0 h_1 \cdots h_q \to \infty$.

### 5.4 MSE

$$\text{AMSE} \left( \hat{f}(y \mid x) \right) = \kappa_2^2 \left( \sum_{j=0}^q h_j^2 B_j(y \mid x) \right)^2 + \frac{R(k)^q f(y \mid x)}{nh_0 h_1 \cdots h_q f(x)}$$

In this problem, the bandwidths enter symmetrically. Thus the optimal rates for $h_0$ and the other bandwidths will be equal. To see this, let $h$ be a common bandwidth and ignoring constants, then

$$\text{AMSE} \left( \hat{f}(y \mid x) \right) \sim h^4 + \frac{1}{nh^{1+q}}$$
with optimal solution

\[ h \sim n^{-1/(5+q)}. \]

Thus if \( q = 1 \), \( h \sim n^{-1/6} \) or \( q = 2 \), \( h \sim n^{-1/7} \). This is the same rate as for multivariate density estimation (estimation of the joint density \( f(y, x) \)). The resulting convergence rate for the estimator is the same as multivariate density estimation.

5.5 Cross-validation

Fan and Yim (2004, Biometrika) and Hall, Racine and Li (2004) have proposed a cross-validation method appropriate for nonparametric conditional density estimators. In this section we describe this method and its application to our estimators. For an estimator \( \hat{f}(y \mid x) \) of \( f(y \mid x) \) define the integrated squared error

\[
I(h) = \int \int \left( \tilde{f}(y \mid x) - f(y \mid x) \right)^2 M(x)f(x)dydx
\]

\[
= \int \int \tilde{f}(y \mid x)^2 M(x)f(x)dydx - 2 \int \int \tilde{f}(y \mid x)M(x)f(y \mid x)f(x)dydx + \int \int f(y \mid x)^2 M(x)f(x)dydx
\]

\[
= \mathbb{E} \left( \int \tilde{f}(y \mid X_i)^2 M(X_i) \right) - 2\mathbb{E} \left( \tilde{f}(y_i \mid x_i) M(X_i) \right) + \mathbb{E} \left( \int f(y \mid X_i)^2 M(X_i)dy \right)
\]

\[
= I_1(h) - 2I_2(h) + I_3.
\]

Note that \( I_3 \) does not depend on the bandwidths and is thus irrelevant.

Let \( \tilde{f}_{-i}(y \mid X_i) \) denote the estimator \( \tilde{f}(y \mid x) \) at \( x = X_i \) with observation \( i \) omitted. For the NW estimator this equals

\[
\tilde{f}_{-i}(y \mid X_i) = \frac{\sum_{j \neq i} K(H^{-1}(X_i - X_j)) k_{h_0}(y_i - y)}{\sum_{j \neq i} K(H^{-1}(X_i - X_j))}
\]

The cross-validation estimators of \( I_1 \) and \( I_2 \) are

\[
\hat{I}_1(h) = \frac{1}{n} \sum_{i=1}^{n} M(X_i) \int \tilde{f}_{-i}(y \mid X_i)^2 dy
\]

\[
\hat{I}_2(h) = \frac{1}{n} \sum_{i=1}^{n} M(X_i) \int \tilde{f}_{-i}(y \mid X_i).\]

The cross-validation criterion is

\[
CV(h) = \hat{I}_1(h) - 2\hat{I}_2(h).
\]

The cross-validated bandwidths \( h_0, h_1, \ldots, h_q \) are those which jointly minimize \( CV(h) \)
For the case of NW estimation

\[
\hat{I}_1 = \frac{1}{n} \sum_{i=1}^{n} M(X_i) \sum_{j \neq i} \sum_{k \neq i} K \left( H^{-1}(X_i - X_j) \right) K \left( H^{-1}(X_i - X_k) \right) \int k_{ho}(y_j - y) k_{ho}(y_k - y) \, dy \\
\left( \sum_{j \neq i} K \left( H^{-1}(X_i - X_j) \right) \right)^2
\]

\[
\hat{I}_2(h) = \frac{1}{n} \sum_{i=1}^{n} M(X_i) \sum_{j \neq i} K \left( H^{-1}(X_i - X_j) \right) k_{ho}(y_i - y_j) \\
\sum_{j \neq i} K \left( H^{-1}(X_i - X_j) \right)
\]

where \( \bar{k} \) is the convolution of \( k \) with itself, and