Nonparametric Density Estimation (one dimension)

Härdle, Müller, Sperlich, Werwarz, 1995, *Nonparametric and Semiparametric Models, An Introduction*

Nonparametric kernel density estimation

Tine Buch-Kromann

February 7, 2007

Derivation

Idea of the histogram:

$$\frac{1}{n \cdot \text{interval length}} \#\{\text{obs that fall into a small interval } \textbf{containing } x\}$$

Idea of the kernel density estimator:

$$\frac{1}{n \cdot \text{interval length}} \#\{\text{obs that fall into a small interval } \textbf{around } x\}$$

Note: That the estimator does not depends on a origin of a bin grid.

When we want to estimate the density in x, we consider the interval $[x-h, x+h)$:

$$\hat{f}_h(x) = \frac{1}{2hn} \#\{X_i \in [x-h, x+h)\}$$

Note: Interval length is $2h$.

Derivation

$$
\begin{aligned}
\hat{f}_h(x) &= \frac{1}{2hn}\#\{X_i \in [x-h, x+h)\} \\
&= \frac{1}{2hn}\sum_{i=1}^{n} I\left(|x - X_i| \le h\right) \\
&= \frac{1}{hn}\sum_{i=1}^{n} \frac{1}{2} I\left(\left|\frac{x - X_i}{h}\right| \le 1\right) \\
&= \frac{1}{hn}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)
\end{aligned}
$$

where $K(u) = \frac{1}{2}I(|u| \le 1)$ is the **uniform kernel function**, which assigns weight $1/2$ to each observation in the interval around x. Points outside the interval assigns the weight 0.

Derivation

Improvement:

More weight to observations very close to x and less weight to observations farther away x (The Epanechnikov kernel function)

$$K(u) = \frac{3}{4}(1 - u^2)I(|u| \leq 1)$$

Kernel	$K(u)$
Uniform	$\frac{1}{2}I(\|u\| \leq 1)$
Triangle	$(1 - \|u\|)I(\|u\| \leq 1)$
Epanechnikov	$\frac{3}{4}(1 - u^2)I(\|u\| \leq 1)$
Quartic	$\frac{15}{16}(1 - u^2)^2 I(\|u\| \leq 1)$
Triweight	$\frac{35}{32}(1 - u^2)^3 I(\|u\| \leq 1)$
Gaussian	$\frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}u^2)$
Cosine	$\frac{\pi}{4} \cos\left(\frac{\pi}{2}u\right) I(\|u\| \leq 1)$

Kernel Density Estimator

General form of the **kernel density estimator** of a probability density f, based on a sample $X_1, ..., X_n$

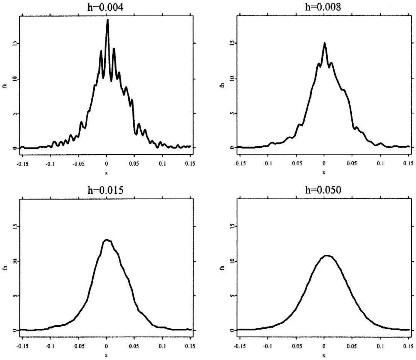
$$\hat{f}_h(x) = \frac{1}{n} \sum_{i=1}^{n} K_h(x - X_i)$$

where $K_h(\cdot) = \frac{1}{h} K(\cdot/h)$ and h is the bandwidth.

Bandwidth

Bandwidth, *h*:

h controls the **smoothness** of the estimate (similar to the histogram) and the choice of *h* is a crucial problem.



The problem of how to determine the value of *h* is a reasonable way is handled later.

The Kernel Function

Properties:

Kernel functions are usually **probability density function**:

$$\int K(u)\,du = 1$$

$$K(u) \geq 0 \;\forall u \text{ in the domain of } K$$

Consequences:

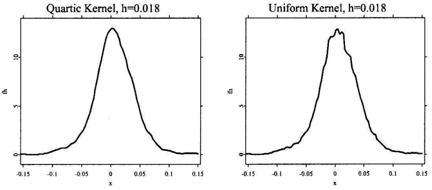
The **kernel density estimator** is a **pdf**

$$\int \hat{f}_h(x)\,dx = 1$$

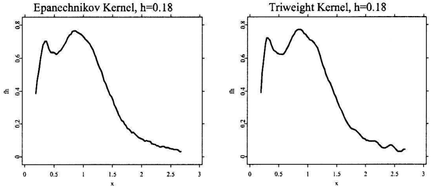
Moreover, $\hat{f}_h(x)$ **inherits** all the continuity and differentiability **properties of** K: If K is ν times continuously differentiable then also $\hat{f}_h(x)$ is ν times cont. diff.

The Kernel Function

The smoothness of $\hat{f}_h(x)$ (for the same value of h) depends of kernel function.



...even if both kernel functions are continuous



Kernel Density Estimator as a Sum of Bumps

An alternative view of the kernel density estimation.
Rescaled kernel function

$$\frac{1}{nh}K\left(\frac{x-X_i}{h}\right)=\frac{1}{n}K_h\left(x-X_i\right)$$

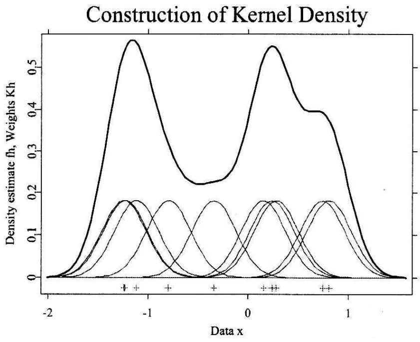
Note: The area under the rescaled kernel function is

$$\begin{aligned}\int \frac{1}{nh}K\left(\frac{x-X_i}{h}\right)\,dx &= \frac{1}{nh}\int K(u)h\,du\\ &= \frac{1}{nh}h\int K(u)\,du\\ &= \frac{1}{n}\end{aligned}$$

Rescaled kernel function

Rewrite the **kernel density function** as the **sum of rescaled kernel functions**

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i) = \sum_{i=1}^{n} \frac{1}{nh} K\left(\frac{x - X_i}{h}\right)$$



Statistical Properties: Bias

Bias:

$$\mathrm{Bias}(\hat{f}_h(x)) = \frac{h^2}{2} f''(x)\mu_2(K) + o(h^2), \qquad h \to 0$$

where $\mu_2(K) = \int s^2 K(s)\, ds$.

The bias is proportional to h^2:
Choose a small h to reduce the bias.

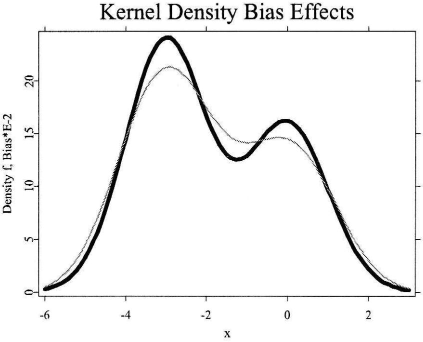
Statistical Properties: Bias

Bias depends on $f''(x)$ (curvature):

In **peaks** of f: The bias < 0 since $f'' < 0$ around a local maximum of f.

In **"valleys"** of f: The bias > 0 since $f'' > 0$ around a local minimum of f.

The **magnitude** of the bias depends of the absolute value of f''.



Statistical Properties: Variance

Variance:

$$\mathbb{V}(\hat{f}_h(x)) = \frac{1}{nh}||K||_2^2 f(x) + o\left(\frac{1}{nh}\right), \qquad nh \to \infty$$

where $||K||_2^2 = \int K^2(s)\,ds$, the squared L_2 norm of K.

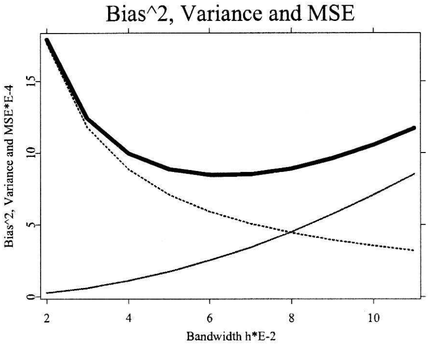
The variance is proportional to $(nh)^{-1}$:

- Choose large h to reduce the variance.
- Increase n to reduce the variance.
- The variance increases in $||K||_2^2$: Flat kernels reduce the variance.

Statistical Properties: MSE

Trade-off between bias and variance:
Increasing h will lower the variance but increases the bias and vice versa.

Minimizing the MSE is a compromise between over- and undersmoothing.



Statistical Properties: MSE

Mean Squared Error:

$$\mathrm{MSE}(\hat{f}_h(x)) = \frac{h^4}{4} f''(x)^2 \mu_2(K)^2 + \frac{1}{nh} ||K||_2^2 f(x) + o(h^4) + o\left(\frac{1}{nh}\right)$$

Note: MSE $\rightarrow 0$ as $h \rightarrow 0$ and $nh \rightarrow \infty$,
ie. the kernel density estimator is a **consistent estimator.**

Statistical Properties: MISE and AMISE

MISE: Mean Integrated Squared Error

$$\mathrm{MISE}(\hat{f}_h) = \frac{1}{nh}||K||_2^2 + \frac{h^4}{4}\mu_2(K)^2||f''||_2^2 + o\left(\frac{1}{nh}\right) + o(h^4),$$

as $h \to 0, nh \to \infty$.
AMISE: Approx. formula for MISE, ignoring higher order terms

$$\mathrm{AMISE}(\hat{f}_h) = \frac{1}{nh}||K||_2^2 + \frac{h^4}{4}\mu_2^2(K)||f''||_2^2$$

Statistical Properties: Optimal bandwidth

Bandwidth: Optimal wrt. AMISE

$$h_{opt} = \left(\frac{||K||_2^2}{||f''||_2^2 \, \mu_2(K)^2 \, n} \right)^{1/5} \sim n^{-1/5}$$

Depends on $||f''||_2^2$ - unknown.

Convergence of AMISE:

$$\mathrm{AMISE}(\hat{f}_{h_{opt}}) = \frac{5}{4} \left(||K||_2^2\right)^{4/5} \left(\mu_2(K)||f''||_2\right)^{2/5} n^{-4/5} \sim n^{-4/5}$$

Note: AMISE converges at the rate $n^{-4/5}$.

Statistical Properties: Comparison with the histogram

AMISE: Histogram

$$\mathrm{AMISE}(\hat{f}_h) = \frac{1}{nh} + \frac{h^2}{12}||f'||_2^2$$

Bandwidth: Optimal wrt. AMISE (histogram)

$$h_0 = \left(\frac{6}{n}||f'||_2^2\right)^{1/3} \sim n^{-1/3}$$

AMISE converges at the rate $n^{-2/3}$ in the histogram.

Slower rate of convergence in the histogram compared to the kernel density estimator.

Bandwidth selection

The two most frequently used method of banwidth selection:

- The plug-in method,
- Cross-validation methods.

Bandwidth selection: Silverman's Rule of Thumb

Plug-in methods:
Replace unknown parameters with estimates.

AMISE optimal bandwidth

$$h_{opt} = \left(\frac{||K||_2^2}{||f''||_2^2 \, \mu_2(K)^2 \, n} \right)^{1/5}$$

Unknown parameter is $||f''||_2^2$.
Assume f is a normal(μ, σ^2)-distribution, then

$$||f''||_2^2 = \sigma^{-5} \frac{3}{8\sqrt{\pi}} \approx 0.212\sigma^{-5}$$

Replace the σ with $\hat{\sigma}$.

Bandwidth selection: Silverman's Rule of Thumb

Choose kernel function: **Gaussian kernel**.

Then the **"Rule-of-Thumb"** bandwidth

$$
\begin{aligned}
\hat{h}_{rot} &= \left(\frac{||\varphi||_2^2}{||\widehat{f''}||_2^2\, \mu_2^2(\varphi)\, n} \right)^{1/5} \\
&= \left(\frac{4\hat{\sigma}^5}{3n} \right)^{1/5} \\
&\approx 1.06\hat{\sigma}\, n^{-1/5}
\end{aligned}
$$

Applicable formula for bandwidth selection.
If X is normally distributed, then $\hat{h}_{rot}$ gives the optimal bandwidth. If X is not normally distributed, then $\hat{h}_{rot}$ will give a bandwidth not too far from the optimal bandwidth, if the distribution of X is not too different from the normal distribution.

Bandwidth selection: Silverman's Rule of Thumb

Practical problem:

The Rule-of-Thumb bandwidths is sensitive to outliers:
A single outlier may cause a too large estimator of σ and hence a too large bandwidth.

Robust estimator:

Calculate

$$R = \underbrace{X_{[0.75n]}}_{75\%-quantile} - \underbrace{X_{[0.25n]}}_{25\%-quantile}$$

We assume $X \sim N(\mu, \sigma^2)$ and $Z \sim N(0, 1)$, therefore

$$\begin{aligned} R &= X_{[0.75n]} - X_{[0.25n]} \\ &= (\mu + \sigma Z_{[0.75n]}) - (\mu + \sigma Z_{[0.25n]}) \\ &= \sigma(Z_{[0.75n]} - Z_{[0.25n]}) \\ &\approx \sigma(0.67 - (-0.67)) \\ &= 1.34\sigma \end{aligned}$$

Bandwidth selection: Silverman's Rule of Thumb

Therefore

$$\hat{\sigma} = \frac{R}{1.34}$$

Plug it into the "Rule-of-Thumb" bandwidth

$$\begin{aligned} \hat{h}_{rot} \approx &= 1.06\hat{\sigma} n^{-1/5} \\ &= 1.06 \frac{R}{1.34} n^{-1/5} \\ &\approx 0.79\hat{R} n^{-1/n} \end{aligned}$$

"Better-Rule-of-Thumb":

Combine the fist and the robust Rule-of-Thumb

$$\hat{h}_{rot} = 1.06 \min\left(\hat{\sigma}, \frac{R}{1.34}\right) n^{-1/5}$$