Kernel Density Estimation

on Euclidean and more general metric spaces

Adam Furlong

August 2025

Contents

- The method
- A classical result
- Some recent research

Setting

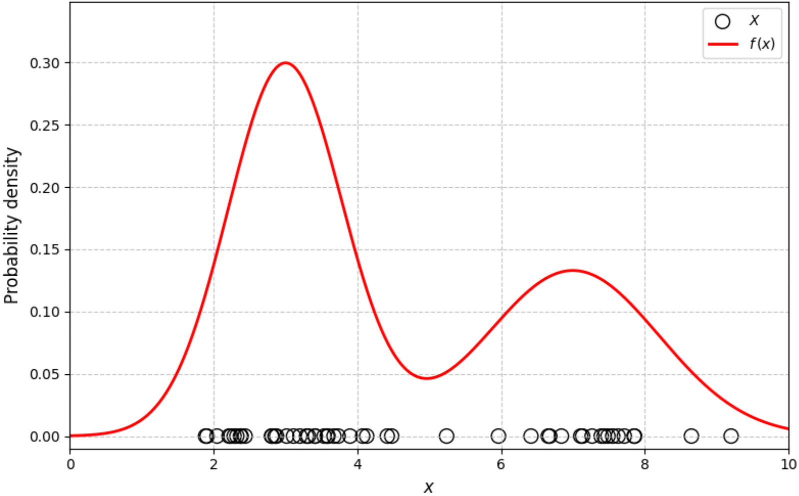
Suppose that f is a **probability density function** (pdf) over $\mathbb{R}$. That is, $f : \mathbb{R} \to [0, \infty)$ with $\int f(x)dx = 1$.

Suppose that $X = (X_1, \cdots, X_n)$ are independent random variables identically distributed according to f.

If we do not have access to f, can we estimate it from the data X?

As n grows, can our estimates improve?

In red: an example pdf.
Black circles: the data X, $n = 50$ numbers generated from f.



Rosenblatt's estimator

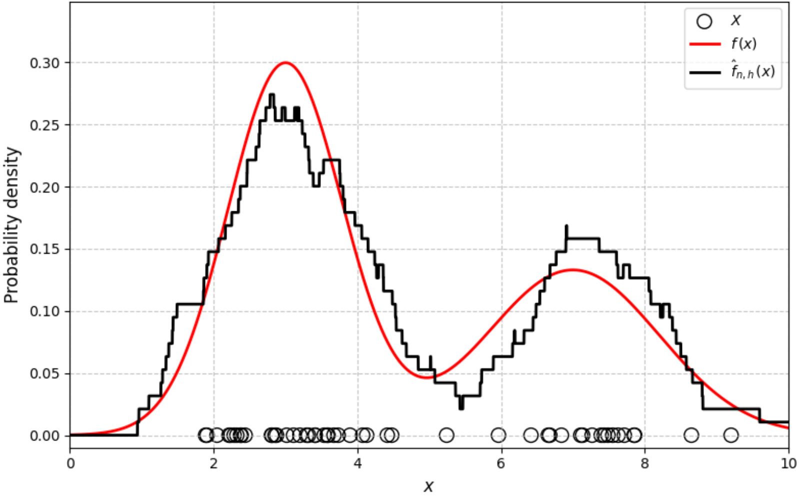
Rosenblatt (1956) suggested the following estimator:

1. Use $K(u) = \begin{cases} \frac{1}{2} & |u| \leq 1 \\ 0 & \text{otherwise} \end{cases}$, the rectangular kernel.
2. Choose a bandwidth $h > 0$. (Think of h as small.)
3. Dilate K by h: $K_h(u) := h^{-1}K(u/h)$.
4. Centre a copy of K_h at each data point X_i and average.

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

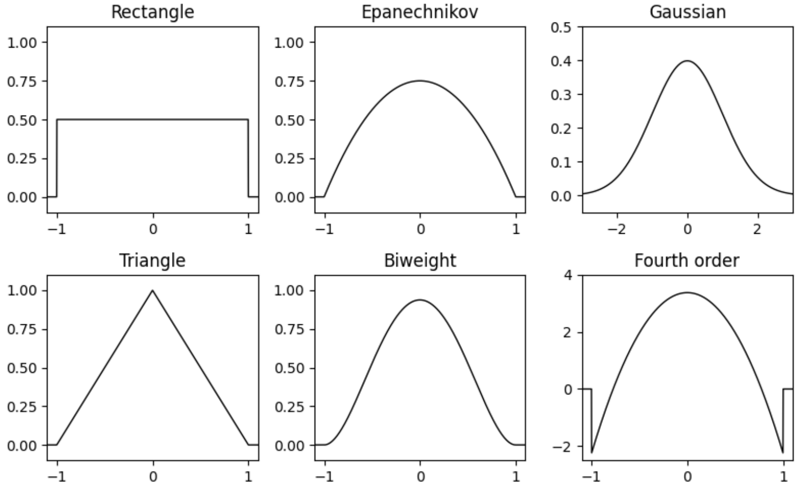
Rosenblatt's estimator

Same pdf f and data X as before. $h = 0.95$.



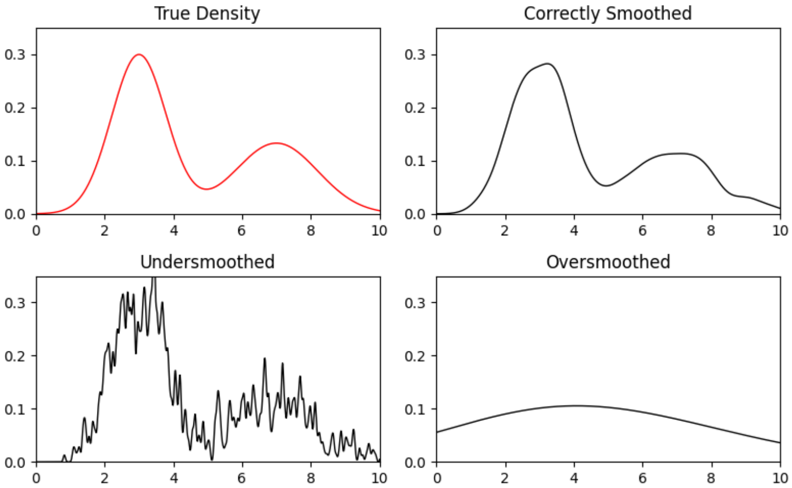
Can we use a different kernel?

Yes! A function $K : \mathbb{R} \to \mathbb{R}$ with $\int K(u)du = 1$, called a **kernel**, will work. Some commonly used examples include:



The choice of bandwidth

$n = 500$ data sampled from f. Clockwise from top left: true density f, estimator $\hat{f}_{n,h}$ with $h = 0.3$, $h = 3$, $h = 0.03$.



Bandwidth

We must carefully choose the bandwidth h.
h too large leads to oversmoothing, large bias.
h too small leads to undersmooting, large variance.
We must choose a value between these extremes.

We expect that as n grows, the best choice for h will change. So, we allow h to be a sequence h_n. We anticipate $h_n \xrightarrow{n \to \infty} 0$.

Kernel Density Estimators

1. Choose a kernel K.
2. Choose a bandwidth $h = h_n > 0$.
3. Dilate K by h: $K_h(u) := h^{-1}K(u/h)$.
4. Centre a copy of K_h at each data point X_i and average.

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

This is a **kernel density estimator**.

Mean Squared Error

The **Mean Squared Error** is defined as

$$
\begin{aligned}
\mathrm{MSE}(x) :=& \mathbb{E}\left(\hat{f}_{n,h}(x) - f(x)\right)^2 \\
=& \int \cdots \int \left(\hat{f}_{n,h}(x, x_1, \cdots, x_n) - f(x)\right)^2 \left(\prod_{i=1}^{n} f(x_i) dx_i\right).
\end{aligned}
$$

We hope that $\mathrm{MSE}(x) \xrightarrow{n\to\infty} 0$.

In particular, we hope there exists some **convergence rate** $R > 0$ and constant $c > 0$ such that for each $x \in \mathbb{R}$, $n \in \mathbb{N}$

$$
\mathrm{MSE}(x) \leq cn^{-R}.
$$

Bias and variance

The Mean Squared Error has a well-known decomposition

$$\mathrm{MSE}(x) = b^2(x) + \sigma^2(x),$$

where the **bias** is given by

$$b(x) := \mathbb{E}\left(\hat{f}_{n,h}(x)\right) - f(x),$$

and the **variance** is given by

$$\sigma^2(x) := \mathbb{E}\left[\left(\hat{f}(x) - \mathbb{E}\left(\hat{f}_{n,h}(x)\right)\right)^2\right].$$

Estimating the Variance

Proposition

Suppose f is bounded and K is square integrable. Then for every $x \in \mathbb{R}, n \in \mathbb{N}$ and $h > 0$,

$$\sigma^2(x) \leq \frac{C_1}{nh},$$

where $C_1 = \|f\|_\infty \int K(u)^2 du$.

We want $h_n \xrightarrow{n\to\infty} 0$, but this result tells us that cannot happen too fast. In order for the variance to vanish as $n \to \infty$, we require $nh \xrightarrow{n\to\infty} \infty$.

Assumptions on f

Let $0 < \alpha \leq 1$. A function $g : \mathbb{R} \to \mathbb{R}$ is Hölder α-continuous if there exists some constant $C > 0$ such that for every $x, y \in \mathbb{R}$

$$|g(y) - g(x)| \leq C|x - y|^{\alpha}.$$

Let $s > 0$, $\ell = \lfloor s \rfloor$ the greatest integer strictly less than s. An ℓ-times differentiable function $f : \mathbb{R} \to \mathbb{R}$ belongs to the Hölder space, $\mathcal{H}^s$, if

$$\|f\|_{\mathcal{H}^s} := \|f\|_\infty + \sup_{x \neq y} \frac{\left|f^{(\ell)}(y) - f^{(\ell)}(x)\right|}{|y - x|^{s-\ell}} < \infty.$$

Estimating the Bias

Using a Taylor expansion, where $R_\ell(x,t)$ is the remainder term,

$$f(x+t) = f(x) + \sum_{m=1}^{\ell-1} \frac{f^{(m)}(x)}{m!} t^m + R_\ell(x,t)$$

$$\begin{aligned} b(x) &= \mathbb{E}\left(\hat{f}_{n,h}(x)\right) - f(x) \\ &= \int K(u) f(x+uh) du - f(x) \\ &= f(x)\left(\int K(u) du - 1\right) + \sum_{m=1}^{\ell-1} \frac{f^{(m)}(x)}{m!} h^m \int u^m K(u) du \\ &+ \int K(u) R_\ell(x, uh) du \end{aligned}$$

Estimating the Bias

A function $K : \mathbb{R} \to \mathbb{R}$ is a **kernel of order s** if it satisifes:

1. $\int K(u)du = 1$
2. $\int u^m K(u)du = 0$ for $1 \leq m \leq \ell = \lfloor s \rfloor$
3. $\int |u|^s |K(u)| du < \infty$

Proposition

Suppose $s > 0$, $f \in \mathcal{H}^s$ *and* K *is a kernel of order* s. *Then*

$$|b(x)| \leq C_2 h^s,$$

where $C_2 = \|f\|_{\mathcal{H}^s} \frac{1}{\ell!} \int |u|^s |K(u)| du.$

Kernels of high order

If K is even, that is $K(u) = K(-u)$, then for each odd m
$$\int u^m K(u) du = 0.$$

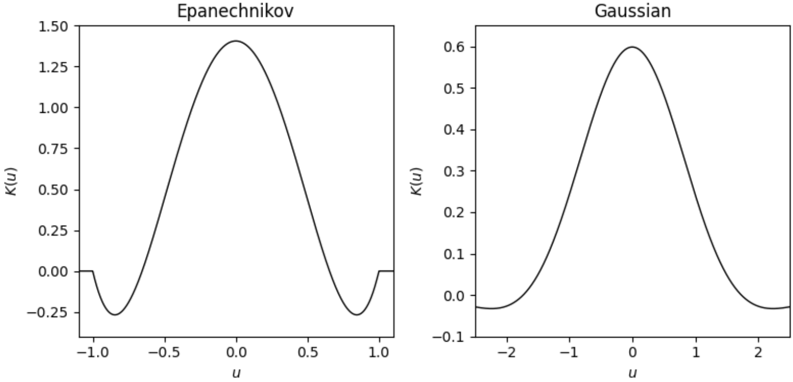
The even m are more difficult. To have both $\int K(u)du = 1$ and $\int u^2 K(u) du = 0$, the kernel must be negative in places.

These exist for every $s > 0$, and may be generated, for example, by multiplying a gaussian by an even polynomial of order $\ell - 2$.

Eg. $K(u) = \frac{3 - u^2}{2} \frac{1}{\sqrt{2\pi}} \exp\left(-u^2/2\right)$ is a kernel of order 4.

Kernels of high order

Examples of fourth order kernels.



Estimating the MSE

If $s > 0$, $f \in \mathcal{H}^s$, K is a square-integrable kernel of order s, then

$$\mathrm{MSE}(x) = \sigma^2(x) + b^2(x) \leq \frac{C_1}{nh} + C_2^2 h^{2s}$$

We see that $h_n = cn^{-q}$ for any $c > 0$ and $0 < q < 1$ will work.

This estimate is minimised by

$$h_n^* = \left(\frac{C_1}{2sC_2^2} \right)^{1/(2s+1)} n^{-1/(2s+1)}.$$

Theorem

Suppose $s > 0$, $f \in \mathcal{H}^s$ and K is a square-integrable kernel of order s. Choose $h = h_n = n^{-1/(2s+1)}$. Then there exists a constant $c = c(s, K, \|f\|_{\mathcal{H}^s}) > 0$ such that

$$\mathrm{MSE}(x) \leq Cn^{-2s/(2s+1)},$$

for every $x \in \mathbb{R}$ and $n \in \mathbb{N}$.

That is, under the assumption $f \in \mathcal{H}^s$, we can attain a convergence rate of $R = 2s/(2s+1)$.

Can we do better?

Theorem

Suppose $s > 0$, $f \in \mathcal{H}^s$ and K is a square-integrable kernel of order s. Choose $h = h_n = n^{-1/(2s+1)}$. Then there exists a constant $c = c(s, K, \|f\|_{\mathcal{H}^s}) > 0$ such that

$$\mathrm{MSE}(x) \leq Cn^{-2s/(2s+1)},$$

for every $x \in \mathbb{R}$ and $n \in \mathbb{N}$.

That is, under the assumption $f \in \mathcal{H}^s$, we can attain a convergence rate of $R = 2s/(2s+1)$.

Can we do better?

No! This is optimal.

Recent Research

1. On what other geometric spaces can we perform kernel density estimation?

2. On such spaces, what convergence rates can we reach?

Recent Research

1. On what other geometric spaces can we perform kernel density estimation?
   A great many metric measure spaces! For example, any Riemannian manifold of non-negative Ricci curvature. This includes $\mathbb{R}^n$, $\mathbb{S}^n$, $\mathbb{B}^n$ and many more.

2. On such spaces, what convergence rates can we reach?

Recent Research

1. On what other geometric spaces can we perform kernel density estimation?

   A great many metric measure spaces! For example, any Riemannian manifold of non-negative Ricci curvature. This includes $\mathbb{R}^n$, $\mathbb{S}^n$, $\mathbb{B}^n$ and many more.

2. On such spaces, what convergence rates can we reach?

   On a d-dimensional space, we can attain a rate of $R = 2s/(2s + d)$ over the analagous Hölder space. This is as good as on Euclidean space $\mathbb{R}^d$!