Parzen Windows

Suppose we assume that the region

$$\mathcal{R}_n$$

is some m-dimensional hypercube whose volume is

$$V_n = l_n^m$$

where l_n is the length of an edge of the hypercube. We define the *window function* ϕ as follows:-

$$\phi(\mathbf{u}) = \begin{cases} 1 & |u_j| \leq 1/2, \quad j = 1, \ldots, m \\ 0 & \text{otherwise} \end{cases}$$

Hence ϕ defines a *unit hypercube* centered at the origin. Evidently,

$$\phi\left(\frac{(\mathbf{x} - \mathbf{x}_i)}{l_n}\right)$$

will be equal to 1 whenever $\mathbf{x}_i$ falls within the hypercube of volume V_n centered at $\mathbf{x}$. Then the number of samples in this hypercube is

$$k_n = \sum_{i=1}^{n} \phi\left(\frac{(\mathbf{x} - \mathbf{x}_i)}{l_n}\right)$$

From before, we had

$$p_n(\mathbf{x}) = \frac{k_n/n}{V_n}$$

whence

$$p_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{V_n} \phi\left(\frac{(\mathbf{x} - \mathbf{x}_i)}{l_n}\right)$$

The essence of the Parzen strategy lies in the choice of window functions ϕ.

Suppose we widen the definition of our window functions to include a more general class of functions (whose properties we might wish to condition). An examination of

$$p_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{V_n} \phi \left(\frac{(\mathbf{x} - \mathbf{x}_i)}{l_n} \right)$$

shows that the window function is essentially an *interpolatory* function, weighting the contribution of each sample $\mathbf{x}_i \in \mathcal{R}_n$ *according to its distance from* $\mathbf{x}$.

What conditions might have to be placed on ϕ? Evidently, we require that $p(\mathbf{x})$ be nonnegative and integrate to 1. Now suppose we keep the requirement that

$$V_n = l_n^m$$

i.e., that the volume be a hypercube. *Then provided*

$$\phi(\zeta) \geq 0$$

and

$$\int \phi(\zeta) d\zeta = 1$$

it will hold that

$$p_n(\mathbf{x}) \geq 0$$

and

$$\int p_n(\zeta) d\zeta = 1$$

In other words,

ϕ is a density function.

The width l_n of the window plays an important role in the estimation p_n of p. (If you are familiar with the Fourier theory, this argument can be formalized). To see this, define

$$\delta_n(\mathbf{x}) = \frac{1}{V_n}\phi\left(\frac{\mathbf{x}}{l_n}\right)$$

Then

$$p_n(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}\delta_n(\mathbf{x}-\mathbf{x}_i)$$

Now if l_n is *large*, then the amplitude of δ_n is small (for

$$\int \delta_n = 1$$

) and $\mathbf{x}_i$ will have to be a (relatively) large distance from $\mathbf{x}$ before $\delta_n(\mathbf{x}-\mathbf{x}_i)$ changes much from $\delta_n(0)$. Thus,

whenever l_n is large, $p_n(\mathbf{x})$ is the superposition of n wide, slowly changing functions i.e., $p_n(\mathbf{x})$ is a smooth, blurred, version of $p(\mathbf{x})$.

On the other hand, whenever l_n is *small*, δ_n will have a large amplitude, peaking at $\mathbf{x} = \mathbf{x}_i$ and decaying *very rapidly* away from $\mathbf{x}_i$. In this case,

$p_n(\mathbf{x})$ is the superposition of n sharp spikes centered at the samples, and is now a noisy version of $p(\mathbf{x})$.

How can we tell if p_n is ever converging to p? First we define what is meant by convergence in circumstances such as this.

Clearly, for any $\mathbf{x}$, the value of $p_n(\mathbf{x})$ depends on the n *random variables* $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$. Hence $p_n(\mathbf{x})$ has some *mean*

$$\mu[p_n(\mathbf{x})]$$

and some variance

$$\sigma^2[p_n(\mathbf{x})]$$

Then p_n is said to converge to p by *mean square convergence* provided

$$\lim_{n \to \infty} p_n(\mathbf{x}) = p(\mathbf{x})$$

and

$$\lim_{n \to \infty} \sigma^2[p_n(\mathbf{x})] = 0$$

Convergence requires that other conditions be placed on $p(\mathbf{x})$, $\phi(\zeta)$ and l_n. As far as p is concerned, the most important consideration is that p be *continuous* at $\mathbf{x}$. Aside from the density conditions on ϕ, the following conditions will assure convergence:-

- $\sup \phi(\zeta) < \infty$
- $\lim_{||\zeta|| \to \infty} \phi(\zeta) \Pi_{i=1}^{m} \zeta_i = 0$
- $\lim_{n \to \infty} V_n = 0$
- $\lim_{n \to \infty} n V_n = \infty$

The first two items ensure that ϕ is a well-behaved function. The other items ensure that V_n approaches zero, *but does so at a rate slower than 1/n*. Why?

We see the reasons for the above conditions by considering the required behavior of the convergences of the mean and variance.

Convergence of the Mean

Consider the mean $\mu[p_n(\mathbf{x})]$. Now the samples $\mathbf{x}_i$ are independent and identically distributed (iid); thus

$$\begin{aligned}
\mu[p_n(\mathbf{x})] &= E[p_n(\mathbf{x})] \\
&= \frac{1}{n}\sum_{i=1}^{n} E\left[\frac{1}{V_n}\phi\left(\frac{(\mathbf{x}-\mathbf{x}_i)}{l_n}\right)\right] \\
&= \int \frac{1}{V_n}\phi\left(\frac{(\mathbf{x}-\zeta)}{l_n}\right) p(\zeta)d\zeta \\
&= \int \delta_n(\mathbf{x}-\zeta)p(\zeta)d\zeta
\end{aligned}$$

Thus,

the mean of the estimate is the convolution of the unknown density with the window function

Hence, $\mu[p_n(\mathbf{x})]$ is a blurred version of $p(\mathbf{x})$ blurred by the window ϕ. But we require V_n to approach zero; some handwaving leads us to

$$\lim_{V_n \to 0} \delta_n(\mathbf{x}-\zeta) = \delta(\mathbf{x})$$

i.e., the sequence δ_n converges to a Dirac delta at $\mathbf{x}$ as V_n goes to zero. Now,
provided p is continuous at $\mathbf{x}$, $\mu[p_n(\mathbf{x})]$ will approach $p(\mathbf{x})$ as $n \to \infty$

Convergence of the Variance

It is important to understand what is meant by the equation

$$\mu[p_n(\mathbf{x})] = \int \delta_n(\mathbf{x}-\zeta)p(\zeta)d\zeta$$

It is saying that $\mu[p_n(\mathbf{x})]$ can be made to approach $p(\mathbf{x})$ for *any* number of samples n (not just an infinite number) provided we let V_n go to zero.

But we need more than one instance of a set of samples of size n!!!!

Any *single* estimate p_n will be (statistically) useless, and one can see this by considering how the *variance* of the estimate behaves. Now p_n is a sum of functions of statistically independent random variables. Hence its variance is linear through the sum

$$\begin{aligned}
\sigma^2[p_n(\mathbf{x})] &= \sum_{i=1}^{n} E\left[\left\{\frac{1}{nV_n}\phi\left(\frac{(\mathbf{x}-\mathbf{x}_i)}{l_n}\right) - \frac{1}{n}\mu[p_n(\mathbf{x})]\right\}^2\right] \\
&= nE\left[\frac{1}{n^2V_n^2}\phi^2\left(\frac{(\mathbf{x}-\mathbf{x}_i)}{l_n}\right)\right] - \frac{1}{n}\mu^2[p_n(\mathbf{x})] \\
&= \frac{1}{nV_n}\int \frac{1}{V_n}\phi^2\left(\frac{(\mathbf{x}-\zeta)}{l_n}\right)p(\zeta)d\zeta - \frac{1}{n}\mu^2[p_n(\mathbf{x})]
\end{aligned}$$

But

$$\delta_n(\mathbf{x}) = \frac{1}{V_n}\phi\left(\frac{\mathbf{x}}{l_n}\right)$$

Hence

$$\sigma^2[p_n(\mathbf{x})] = \int \delta_n(\mathbf{x}-\zeta)p(\zeta)\phi\left(\frac{(\mathbf{x}-\zeta)}{l_n}\right)d\zeta - \frac{1}{n}\mu^2[p_n(\mathbf{x})]$$

Dropping the second term, and using a little analysis, we obtain the bound

$$\sigma^2[p_n(\mathbf{x})] \leq \frac{\sup(\phi)\mu[p_n(\mathbf{x})]}{nV_n}$$

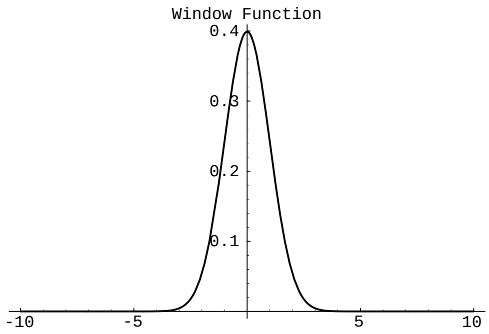
Evidently, the bound seems to imply that a small variance is best achieved when V_n gets *large* (not small). What it is really saying is that

a zero variance can be achieved even when V_n goes to zero by ensuring that nV_n goes to infinity

Hence the requirement that V_n approach zero at a rate slower than $1/n$.

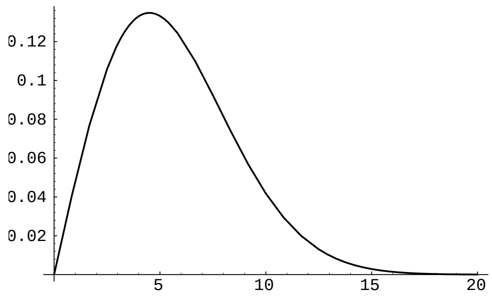
Here are examples using the window function

$$\phi(\zeta) = \frac{1}{\sqrt{2\pi}} \exp[-\frac{1}{2}\zeta^2]$$



The target distribution is the Rayleigh Distribution

$$p(\mathbf{x}) = \mathrm{Rayleigh}[4.5]$$

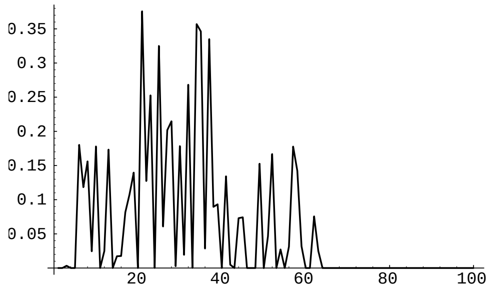


We let

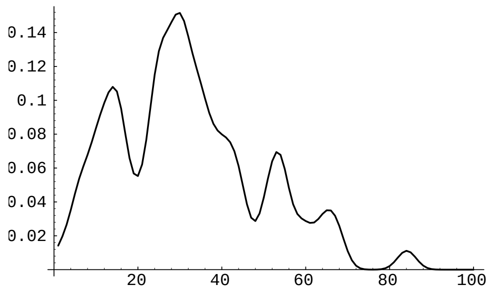
$$l_n = l_1/\sqrt{n}$$

where l_1 is a length parameter we vary.

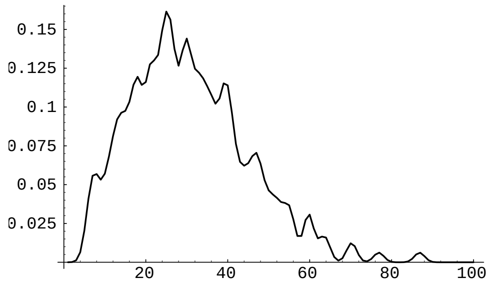
Here is the Parzen estimate for n=80, $l_1 = 0.25$. Note that the range $[0, 20]$ has been divided into 100 equal increments).



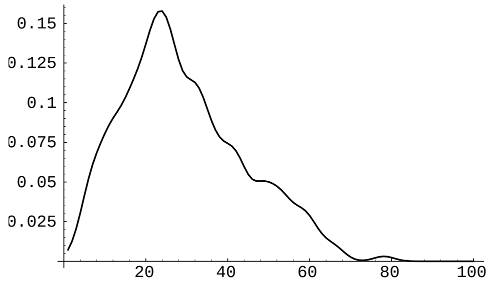
The estimate is noisy (l_n is small with respect to the Rayleigh distribution). When $l_1 = 4$ the following is obtained:-



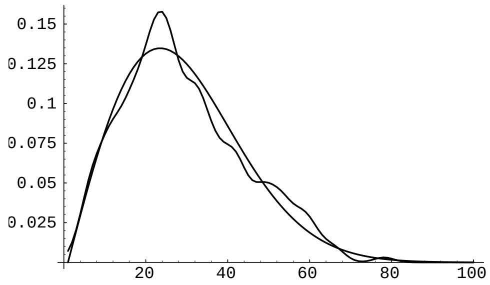
When n=256 ($l_1 = 4$) one gets the following:-



and when $l_1 = 8$ (n=256) convergence is even better:-



Superimposed with the true distribution, one has



Here, for comparison, is the histogram of the 256 samples:-