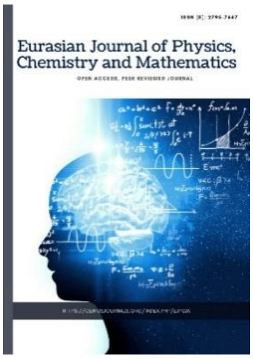

Exploring the Impact of Kernel Function and Bandwidth Parameters on Covariate Estimates

Mansurov Dilshod Ravilovich
Navoi State Pedagogical Institute, Faculty of Mathematics and Informatics, Department of Mathematics, Doctor of Philosophy (PhD) in Physics and Mathematics, Navoi, Uzbekistan. e-mail: mathematicianmd@gmail.com

Sherkulova Sabina Lochin kizi
1nd year master's student of the specialty “Mathematics”, Navoi State Pedagogical Institute, Navoi, Uzbekistan

ABSTRACT

In this article we are exploring the impact of kernel function and bandwidth parameters on covariate estimates

Keywords: Kernel function

Suppose that $\pi(t)$ is a known density function (kernel), $\{h_n\}$ is a sequence of positive numbers (“bandwidth”), $h_n \to 0$ in $n \to \infty$. Let $\{\omega_{ni}(x;h_n)\}_{i=1}^{n}$ be the Hesser-Müller weight function. In that case, the following assessment is appropriate:

$$\omega_{ni}(x;h_n)=\frac{1}{C_n(x;h_n)}\int\limits_{x_{i-1}}^{x_i}\frac{1}{h_n}\pi\left(\frac{x-y}{h_n}\right)dy, i=1,\ldots,n;$$

$$C_n(x;h_n)=\int\limits_{x_0}^{x_n}\frac{1}{h_n}\pi\left(\frac{x-y}{h_n}\right)dy, \sum_{i=1}^{n}\omega_{ni}(x;h_n)=1.$$

Here we use an analogue of Stone [1] statistics to estimate the distribution function $H_x(t)$ and the probability $\left\{p_x^{(m)}=P\left(X_x^{(m)}=1\right), m=0,1,2\right\}$:

$$H_{xh}(t)=\sum_{i=1}^{n}I(\xi_i\le t)\omega_{ni}(x;h_n), \qquad (0.1)$$

$$P_{xh}^{(m)}=\sum_{i=1}^{n}X_i^{(m)}\omega_{ni}(x;h_n), m=0,1,2. \qquad (0.2)$$

In particular, $\omega_{ni}(x;h_n)=\frac{1}{n}$ (or in the absence of a covariate) (1) and (2) becomes a simple empirical estimate. (2) putting the values in equality into the expression $P\left(\delta_x^{(0)}=1\right)=1-\lambda_x, P\left(\delta_x^{(1)}=1\right)=\tau_x\lambda_x, P\left(\delta_x^{(2)}=1\right)=(1-\tau_x)\lambda_x$, we obtain the following values for the parameters λ_x and γ_x:

$$\lambda_{xh} = 1 - p_{xh}^{(0)}, \quad \gamma_{xh} = p_{xh}^{(1)}\left(1 - p_{xh}^{(0)}\right)^{-1}. \quad (0.3)$$

Now putting values (1) and (3) into the equation $1 - F_x(t) = \left[1 - \left(H_x(t)\right)^{\lambda_x}\right]^{\tau_x}, t \ge 0,$ where $\lambda_x = \dfrac{1}{1+\beta_x}$ and $\tau_x = \dfrac{1}{1+\theta_x}$. For the F_x distribution function, we have the following estimate:

$$F_{xh}(t) = 1 - \left\{1 - \left[H_{xh}(t)\right]^{\lambda_{xh}}\right\}^{\gamma_{xh}}, t \ge 0. \quad (0.4)$$

Our subsequent primary focus of investigation (4) pertains to the domain of evaluation. We systematically examine diverse attributes of this estimation through the application of statistical modeling methodologies. To facilitate this endeavor, the ensuing kernel functions serve as instrumental elements:

Table 1
Kernel functions

	Kernel functions	$\int t^2\pi(t)dt$	$\int \pi^2(t)dt$
Uniform ("rectangular window")	$\pi(t) = \begin{cases} \frac{1}{2}, \lvert t\rvert \le 1; \\ 0, \lvert t\rvert > 1. \end{cases}$	$\frac{1}{3}$	$\frac{1}{2}$
Triangular	$\pi(t) = \begin{cases} 1-\lvert t\rvert, \lvert t\rvert \le 1; \\ 0, \lvert t\rvert > 1. \end{cases}$	$\frac{1}{6}$	$\frac{2}{3}$
Epanechnikov (parabolic)	$\pi(t) = \begin{cases} 0,75\left(1-t^2\right), \lvert t\rvert \le 1; \\ 0, \lvert t\rvert > 1 \end{cases}$	$\frac{1}{5}$	$\frac{3}{5}$
Quartic (biweight)	$\pi(t) = \begin{cases} \frac{15}{16}\left(1-t^2\right)^2, \lvert t\rvert \le 1; \\ 0, \lvert t\rvert > 1. \end{cases}$	$\frac{1}{7}$	$\frac{5}{7}$
Triweight	$\pi(t) = \begin{cases} \frac{35}{32}\left(1-t^2\right)^3, \lvert t\rvert \le 1; \\ 0, \lvert t\rvert > 1. \end{cases}$	$\frac{1}{9}$	$\frac{350}{429}$
Gaussian	$\pi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}, \lvert t\rvert < \infty;$	1	$\frac{1}{2\sqrt{\pi}}$
Cosine	$\pi(t) = \begin{cases} \frac{\pi}{4}\cos\left(\frac{\pi}{2}t\right), \lvert t\rvert \le 1; \\ 0, \lvert t\rvert > 1. \end{cases}$	$1 - \frac{8}{\pi^2}$	$\frac{\pi^2}{16}$

Now we begin to study the properties of the estimate. For this, the minimum number of repetitions of experiments to achieve $\varepsilon = 0,01$ accuracy should be equal to $N = 16000$. As a measure of how good an estimate is, we study the $\sup_{t\in\mathbb{R}} \left|F_x(t) - F_{xh}(t)\right|$ "distance" between the estimate and the theoretical distribution function. The smaller the distance, the better the score. Below is a table of average values of

$\sup\limits_{t\in\Box}\left|F_x(t)-F_{xh}(t)\right|$ in $N=20000$ experiments. To compile the table, we use a sample with a size of $n=300$ and an average censoring rate of 10%. We choose the Cox model $F(t)=1-\left(1-F_0(t)\right)^{r(x,\beta)}$ as the theoretical distribution function, and the exponential distribution as the base distribution function.

Table 2
The variance between the estimate and the theoretical distribution. (Cox model $S(t)=\left(S_0(t)\right)^{r(x,\beta)}$, $S_0(t)-$ exponential, $r(x,\beta)=e^{\beta_0+\beta_1\log x}$, $\beta_0=1,\beta_1=2$, $n=1000$, censorship 10%**)**

x / h_n	0,1	0,2	0,3	0,5	0,7	0,9	1
Uniform ("rectangular window")							
0,1	0,08368	0,08458	0,08422	0,08394	0,08334	0,08498	0,11643
0,3	0,06046	0,05330	0,05046	0,04932	0,04827	0,06094	0,06848
0,5	0,04897	0,04548	0,04189	0,03778	0,04285	0,04935	0,05378
0,7	0,04263	0,03994	0,03813	**0,03669**	0,03813	0,04194	0,04507
0,9	0,03803	**0,03643**	0,03774	0,03800	**0,03687**	0,03806	0,03982
1	**0,03647**	0,03820	**0,03669**	0,03829	0,03793	**0,03750**	0,03839
3	0,03783	0,03837	0,03804	0,03838	0,03855	0,03808	**0,03725**
5	0,03824	0,03845	0,03818	0,03793	0,03877	0,03882	0,03814
Triangular							
0,1	0,09642	0,09678	0,09668	0,09672	0,09655	0,09739	0,13451
0,3	0,06800	0,05877	0,05755	0,05654	0,05566	0,06747	0,07873
0,5	0,05532	0,05050	0,04665	0,04311	0,04739	0,05565	0,06177
0,7	0,04868	0,04513	0,04178	0,03956	0,04240	0,04798	0,05145
0,9	0,04297	0,04097	0,03960	**0,03691**	0,03979	0,04348	0,04579
1	0,04156	**0,03782**	**0,03716**	0,03796	**0,03748**	**0,03622**	0,04401
3	**0,03609**	0,03810	0,03814	0,03844	0,03865	0,03820	**0,03741**
5	0,03832	0,03848	0,03856	0,03898	0,03896	0,03884	0,03815
Epanechnikov (parabolic)							
0,1	0,09185	0,09226	0,09216	0,09190	0,09182	0,09273	0,12820
0,3	0,06519	0,05647	0,05496	0,05360	0,05267	0,06499	0,07467
0,5	0,05275	0,04857	0,04487	0,04106	0,04551	0,05322	0,05875
0,7	0,04627	0,04323	0,04032	0,03832	0,04076	0,04547	0,04897
0,9	0,04100	0,03923	0,03844	0,03819	0,03865	0,04137	0,04353

1	0,03945	0,03832	**0,03618**	**0,03741**	**0,03643**	0,03930	0,04190
3	**0,03785**	**0,03780**	0,03805	0,03839	0,03756	**0,03708**	**0,03624**
5	0,03824	0,03845	0,03847	0,03894	0,03836	0,03882	0,03813
Quartic (biweight)							
0,1	0,09994	0,10013	0,10004	0,09992	0,10001	0,10079	0,13955
0,3	0,06942	0,05997	0,05954	0,05843	0,05762	0,06887	0,08161
0,5	0,05658	0,05141	0,04749	0,04455	0,04816	0,05690	0,06418
0,7	0,04988	0,04614	0,04261	0,03970	0,04323	0,04924	0,05335
0,9	0,04399	0,04176	0,03979	**0,03671**	0,04004	0,04457	0,04740
1	0,04238	0,04117	**0,03619**	0,03874	0,03946	0,04204	0,04554
3	**0,03790**	**0,03784**	0,03806	0,03879	**0,03756**	**0,03710**	**0,03728**
5	0,03825	0,03845	0,03817	0,03894	0,03836	0,03881	0,03813
Triweight							
0,1	0,10654	0,10658	0,10642	0,10662	0,10651	0,10743	0,14903
0,3	0,07274	0,06317	0,06321	0,06245	0,06165	0,07189	0,08717
0,5	0,05966	0,05361	0,04972	0,04749	0,05047	0,05981	0,06848
0,7	0,05277	0,04832	0,04442	0,04139	0,04511	0,05211	0,05695
0,9	0,04643	0,04386	0,04122	0,03949	0,04147	0,04716	0,05063
1	0,04507	0,04304	0,04037	**0,03723**	0,04060	**0,03757**	**0,03658**
3	**0,03798**	**0,03791**	**0,03709**	0,03840	**0,03758**	0,03814	0,03836
5	0,03826	0,03846	0,03876	0,03894	0,03836	0,03881	0,03893
Gaussian							
0,1	0,07232	0,06524	0,06251	0,06360	0,06343	0,07257	0,08896
0,3	0,04902	0,04486	0,04345	0,04075	0,04268	0,04891	0,05133
0,5	0,04095	0,03994	0,03871	0,03813	0,03927	0,04075	0,04200
0,7	0,03933	0,03849	0,03831	0,03781	0,03804	0,03835	0,03913
0,9	0,03826	0,03770	0,03786	**0,03704**	**0,03707**	0,03806	0,03832
1	**0,03675**	**0,03644**	**0,03679**	0,03832	0,03814	**0,03711**	**0,03716**
3	0,03784	0,03778	0,03804	0,03838	0,03879	0,03807	0,03824
5	0,03824	0,03845	0,03818	0,03894	0,03890	0,03882	0,03856
Cosine							
0,1	0,09307	0,09345	0,09335	0,09314	0,09308	0,09398	0,12991
0,3	0,06588	0,05702	0,05565	0,05433	0,05344	0,06561	0,07574
0,5	0,05335	0,04903	0,04530	0,04159	0,04593	0,05382	0,05957

0,7	0,04684	0,04372	0,04068	0,03853	0,04117	0,04610	0,04962
0,9	0,04147	0,03965	0,03866	0,03827	0,03888	0,04187	0,04412
1	0,03992	0,03962	0,03834	**0,03646**	0,03860	0,03973	0,04245
3	**0,03686**	**0,03681**	**0,03605**	0,03839	**0,03656**	**0,03708**	**0,03624**
5	0,03825	0,03845	0,03817	0,03794	0,03836	0,03882	0,03813

Upon meticulous scrutiny of the tables above, it becomes apparent that the accuracy of our estimate is profoundly influenced by the choice of the kernel function, the covariate selected during estimation construction, and the parameters governing the “bandwidth”. While the selection of sample and covariate might be predetermined, the responsibility for choosing the optimal kernel function and “bandwidth” parameters lies squarely with us during the estimation process. Thus, in our pursuit of enhancing estimation accuracy, our focus centers on identifying the most advantageous kernel function and fine-tuning the “bandwidth” parameter.

A closer examination of each column reveals a discernible optimal “bandwidth”. For instance, in the first column, a “bandwidth” of 1 emerges as the smallest relative to its neighboring values, emphasized by bold markings in the table. This observation underscores the presence of a distinct optimal “bandwidth” that careful scrutiny can unveil. Further investigation across other columns unveils a strong dependence of the “bandwidth” parameter on the covariate's value.

In summary, after subjecting the covariate estimator to rigorous testing involving diverse covariates, various kernels, and “bandwidth”, a consistent pattern emerges. The Gaussian kernel function consistently outperforms others, establishing itself as the optimal choice for achieving superior estimation performance across all scenarios.

References

1. Stone C. Consistent Nonparametric Regression // The Annals of Statistics. 1977. № 4 (5). C. 595–645. https://doi.org/10.1214/aos/1176343886.