Estimation the Hazard Rate Function using Gumbel Type 2 Kernel

Ola A. Elsamadony
Department of Mathematics, Faculty of Science,
Tanta University, Tanta, Egypt

Ahmed Mattar
Department of Mathematics, Faculty of Science,
Tanta University, Tanta, Egypt

ABSTRACT

Survival analysis is a branch of statistics where it is a collection of statistical procedures for data analysis where the outcome variable of interest is time until an event occurs, such as failure in mechanical systems. Hazard rate estimation for the lifetime event is a basic tool for processing survival analysis. Kernel estimators are boundary effects near the endpoints of the support of the hazard rate and some solutions have been proposed to solve this problem, including the use of asymmetric kernel functions like the Gumbel type 2 kernel function. We study the non-parametric estimation of the hazard rate function using the Gumbel type 2 kernel function for identically independent data. The bias, variance and optimal bandwidth will be investigated, then AMSE of the proposed estimator were obtained.

Keywords

Nonparametric estimation, Kernel functions, Gumbel Type 2, Bandwidth, Hazard rate function

1. INTRODUCTION

Kernel density estimation is a useful statistical tool. Often shortened to KDE, it’s a technique that let’s you create a smooth curve given a set of data. It appeared in early 1950 by Parzen and Rosenblatt . Although the method was introduced in the middle of the last century until recently it remained unpopular because of its computationally intensive nature. The kernel function is a good alternative to the histogram, where the histogram is an estimator of the probability function but the histogram can not provide us with the probability of the event of interest, While the kernel function treats these disadvantages. Therefore, the kernel function can be used as an estimator for the unknown probability density function. For more details in this regard, we may refer the reader to [4], [10], [12] and [11]. In this regard, the non-parametric estimation of hazard rates for lifetime data has become a common tool for statisticians, as it is an essential tool for processing survival analysis. Focusing on kernel estimators, it was observed that bias problems occur when estimating near the endpoints of the data. These are called a boundary effect. Boundary effects are a major complication when smoothing hazard rate. Estimators of the hazard function based on kernel smoothing have been studied widely. For more details, see [6] and [8]. Recently, [3] suggested a nice way to circumvent the well-known boundary bias or edge effect that appears in standard kernel density estimation. There are many symmetric parametric kernel estimators in the literature. Specifically, [8] introduced the inverse Gaussian kernel estimator and examined some properties, such as bias and variance. It is proved to be boundary bias-free and achieve a significant rate of convergence for the asymptotic mean integrated squared error (AMSE). Besides, prove these properties to hazard rate function estimation. On the same methodological basis,[9] and [1] developed the Weibull and inverse Gaussian hazard rate kernel estimators, respectively.

Some example of kernel functions such as Gaussian, Linear, Cosine and others in figures 1, 2 and 3.

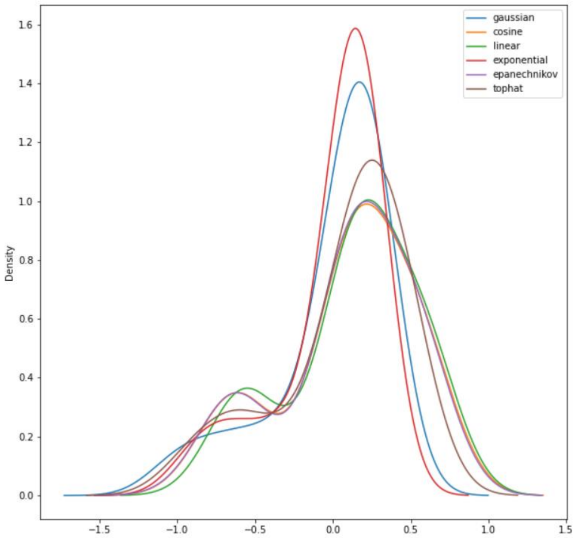

Figure 1: Plots of the kernel functions

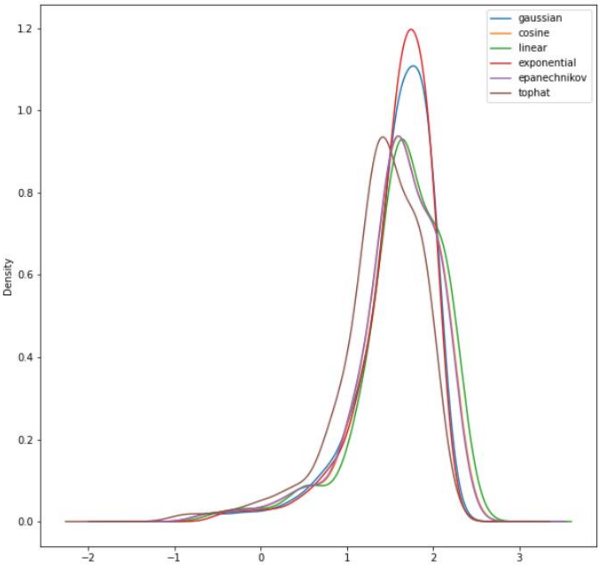

Figure 2: Plots of the kernel functions

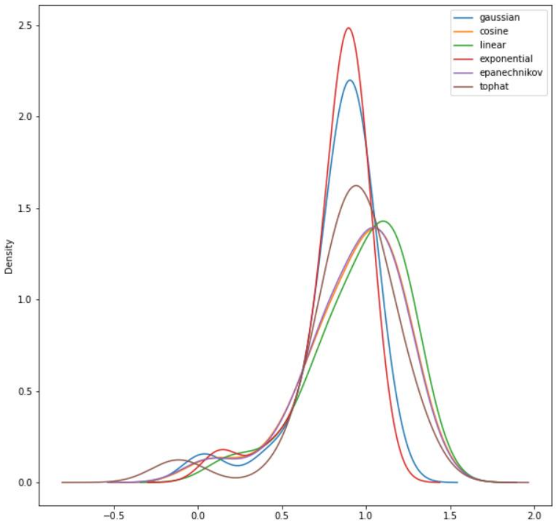

Figure 3: Plots of the kernel functions

The remainder of the paper is outlined as follows. In Section 2, we introduce the cumulative and hazard rate of Gumbel type 2 kernel function. In Section 3, we define the Hazard rate function of Gumbel type 2 kernel estimator. In Section 4, we get the theoretical properties, such as bias, variance and optimal bandwidth. In Section 5, we conclude the paper.

2. THE CUMULATIVE AND HAZARD RATE OF GUMBEL TYPE 2 KERNEL FUNCTION

Let X1, . . . , Xn be n independent and identically distributed random variable with a common pdf denoted by f(x). Hence, a general form of a kernel estimation of f(x) is given by:

$$\tilde{f}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{x,h}\left(X_i\right), \quad x \in \mathbb{R}, \tag{1}$$

where $k_{x,h}$(t) = $\frac{1}{h}$k ($\frac{x-t}{h}$), k (t) is a kernel function and h is called a bandwidth. We use the Gumbel type 2 kernel function, which was introduced by [2]. The following Gumbel type 2 is selected as the kernel function in this paper:

$$K_{G;x,h}(t) = \frac{1}{h}\left[\frac{x}{\Gamma(1-h)}\right]^{\frac{1}{h}} t^{-\frac{1}{h}-1} \exp\left(-\left[\frac{x}{t\Gamma(1-h)}\right]^{\frac{1}{h}}\right),$$

where t, x > 0, h ∈ (0, 1) and Γ(x) = $\int_0^{+\infty} t^{x-1}\, e^{-t}$ dt is the standard gamma function. x and h refer to the secondary parameters and the main one being the variable t.

2.1 This kernel function is defined by the following cumulative kernel function:

$$K^*{}_{G;x,h}(t) = \mathrm{e}^{-t^{-1/h}\left(\frac{x}{\mathrm{Gamma[1\text{-}h}}\right)^h}. \tag{2}$$

The flexibility of the shapes for K∗ G;x,h(t) is illustrated in Figure4.

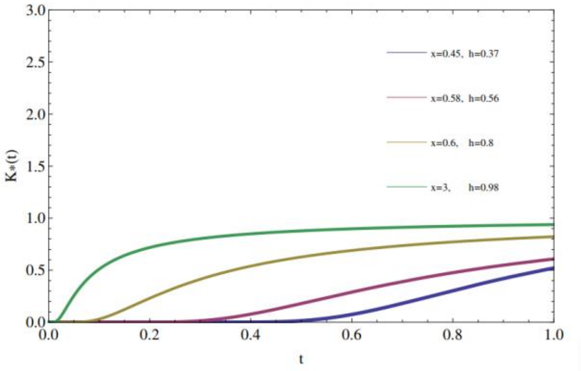

Figure 4: Plots of the Gumbel type 2 cumulative kernel function for some values of the parameters.

2.2 The hazard rate of Gumbel type 2 kernel function is defined by:

$$r(K(t)) = \frac{K(t)}{1-K^*(t)} = \frac{t^{-\frac{1+h}{h}}\left(\frac{x}{\mathrm{Gamma}[1-h]}\right)^h}{\left(1-\mathrm{e}^{t^{-1/h}}\left(\frac{x}{\mathrm{Gamma[1\text{-}h]}}\right)^h\right)h}. \tag{3}$$

The flexibility of the shapes for rG;x,h(t) is illustrated in Figure 5.

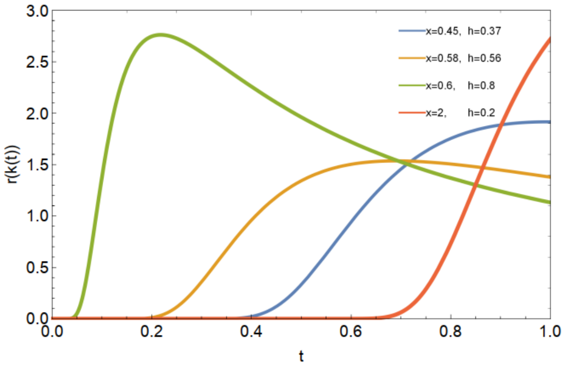

Figure 5: Plots of hazard rate of the Gumbel type 2 kernel function for some values of the parameters.

3. HAZARD RATE FUNCTION OF GUMBEL TYPE 2 KERNEL ESTIMATOR

Place Tables/Figures/Images in text as close to the reference as possible (see Figure 1). It may extend across both columns to a maximum width of 17.78 cm (7”).

$$r(x) = \lim_{\Delta x \to 0} \frac{P(X \le x + \Delta x \mid X > x)}{\Delta x}, x > 0.$$

The hazard rate function can be written as the ratio of the density function and the survivor function, as the following:

$$\mathrm{r(x)} = \frac{f(x)}{s(x)} \tag{4}$$

where, S(x) = 1−F(x) and the kernel estimator for the survivor function can defined as $\hat{S}$(x) = 1 − $\hat{F}$(x) and $\hat{F}$(x) = $\int_0^x$ $\hat{f}$ˆ(t)dt = $\frac{1}{n}$ $\sum_{i=1}^{n}$ $\int_0^x$

K$G_{x,h}$(t)dt. Through equation (4), the proposed estimator for the hazard rate function is given by

$$\hat{r}(x) = \frac{\hat{f}(x)}{\hat{S}(x)} = \frac{\hat{f}(x)}{1-\hat{F}(x)}. \tag{5}$$

As preliminary assumptions, in the whole paper, we suppose that the pdf f(x) is defined on (0, +∞) such that f(x) has a continuous second derivative. Among others, this implies that ‖ f ´´(.) ‖ ∞ is bounded, where ‖ g(.) ‖ ∞ = $sub_{t\in}$(0,+∞) |g(t)| denotes the supremum norm of a function g(t). These assumptions will be used in the calculations related to the bias and the variance. The bias and variance of the Gumbel type 2 kernel estimator was proven in the paper Bakouch et al. (2021).

3.1 Lemma

$$\mid \ Bias\ [\hat{f}(x)] \mid \leq \frac{1}{2}x^2\left[\frac{\Gamma(1-2h)}{\Gamma(1-h)^2}-1\right]\left\|f''(.)\right\|_\infty .$$

3.2 Lemma

$$\mathrm{Var}[\hat{f}(x)] \leq \frac{1}{4\,\mathrm{nh}}\left[\frac{1}{\Gamma(1-h)}\right]^{\frac{1}{5}}\left\{\left\|.^{-1}f(.)\right\|_\infty + Cx^2 2^{2h}h^{-2}\left[\frac{\Gamma(1-2h)}{\Gamma(1-h)^2}-1\right]\right\}.$$

4. BIAS, VARIANCE AND RATE OF CONVERGENCE OF HAZARD RATE KERNEL ESTIMATOR

The bias and variance of the hazard rate Gumbel type 2 kernel estimator $\hat{r}$(x) defined by Equation (5) are the objects of this section. Thus, we aim to provide a theoretical performance of $\hat{r}$(x).

4.1 Bias

We recall that its mathematical definition is Bias[$\hat{r}$(x)] = E[$\hat{r}$(x)] − r(x).

4.2 Proposition 4.1. The bias of $\hat{r}$(x) satisfies

$$\mid \ Bias\ [\hat{f}(x)] \mid \leq \frac{1}{2}x^2\left[\frac{\Gamma(1-2h)}{\Gamma(1-h)^2}-1\right]\left\|f''(.)\right\|_\infty .$$

Proof. The following proposition based on lemma 3.1, it is noteworthy that

$$E[\hat{r}(x)] = \frac{E[\hat{f}(x)]}{E[\hat{S}(x)]} \simeq \frac{f(x)+\frac{1}{2}\left\|f''(.)\right\|_\infty x^2\left[\frac{\Gamma(1-2h)}{\Gamma(1-h)^2}-1\right]}{S(x)}$$
$$\simeq r(x) + \frac{\left\|f''(.)\right\|_\infty x^2\left[\frac{\Gamma(1-2h)}{\Gamma(1-h)^2}-1\right]}{2S(x)}.$$

Hence

$$\mid \mathrm{Bias}[\hat{r}(x)]\mid = \frac{1}{2S(x)}x^2\left[\frac{\Gamma(1-2h)}{\Gamma(1-h)^2}-1\right]\left\|f''(.)\right\|_\infty .$$

This ends the proof of Proposition 4.1.

From Proposition 4.1, by applying h → 0 when n → +∞, we can use the equivalence $\frac{\Gamma(1-2h)}{(\Gamma(1-h))^2}$ − 1 ∼ π 2 6 h 2 → 0, we get

$$\mid \mathrm{Bias}[\hat{r}(x)]\mid = \frac{1}{12S(x)}\left\|f''(.)\right\|_\infty (x\pi h)^2.$$

Implying that

Bias(ˆr(x)) → 0.

Hence, the estimator ˆr(x) is asymptotically unbiased, Proposition 4.1 implies the existence of a constant C∗ > 0 such that

$$\mathrm{Bias}[\hat{r}(\mathrm{x})]| \leq C{*}h^2 \tag{6}$$

Thus, this shows that the bias depends on the bandwidth h and x, and it goes to zero as h → 0.

4.3 Variance

The following proposition is based on lemma 3.2, to investigate the variance of $\hat{r}$(x).

4.4 Proposition 4.2. suppose that ‖.$^{-1}$f(.)‖ ∞ exists. Then, the variance of rˆ(x) satisfies

$$\mathrm{Var}[\hat{r}(x)] \leq \frac{1}{S^2(x)}\frac{1}{4nh}\left[\frac{1}{\Gamma(1-h)}\right]^{\frac{1}{h}}\left\{\left\|.^{-1}f(.)\right\|_\infty + Cx^2 2^{2h}h^{-2}\left[\frac{\Gamma(1-2h)}{\Gamma(1-h)^2}-1\right]\right\}.$$

Proof.

$$\mathrm{Var}[\hat{r}(x)] = \mathrm{Var}\left[\frac{\hat{f}(x)}{\hat{S}(x)}\right] = \frac{1}{S^2(x)}\frac{1}{4nh}\left[\frac{1}{\Gamma(1-h)}\right]^{\frac{1}{h}}\left\{\left\|^{-1}f(.)\right\|_\infty + Cx^2 2^{2h}h^{-2}\left[\frac{\Gamma(1-2h)}{\Gamma(1-h)^2}-1\right]\right\}.$$

This ends the proof of Proposition 4.2.

From Proposition 4.2, By applying h → 0 and nh → +∞ when n → +∞, since $\frac{\Gamma(1-2h)}{(\Gamma(1-h))^2}$ − 1 ∼ $\frac{\pi^2}{6}$ h 2 → 0, and $\left[\frac{1}{\Gamma(1-h)}\right]$ $\frac{1}{h}$ ∼ e −γ where γ denotes the Euler constant, i.e., γ ≈ 0.5772, we get

$$\mathrm{Var}[\hat{r}(x)] \leq \frac{1}{nh}\frac{e^{-\gamma}}{4S^2(x)}\left\{\left\|^{-1}f(.)\right\|_\infty + Cx^2 2^{2h}\frac{\pi^2}{6}\right\}.$$

Hence

Var[$\hat{f}$(x)] → 0.

Proposition 4.2 implies the existence of a constant C_{**} > 0 such that

$$\mathrm{Var}[\hat{f}(\mathrm{x})] \leq \frac{C{**}}{nh}. \tag{7}$$

This result will be important for determining the optimal bandwidth for ˆr(x) in the next section.

4.5 Bandwidth selection

Bandwidth selection plays an important role in kernel estimation, the estimate will be important when the bandwidth is very small. For more details see [7] and [5]. In order to get optimal bandwidth (hopt), we define the AMSE of ˆr(x) as follows AMSE[$\hat{r}$(x)] = E $[(\hat{r}(x) - r(x))^2]$. Based on Equations (6) and (7), and the underlying assumptions, we have

$$\text{AMSE}[\hat{r}(\text{x})] = \{\text{Bias}[\hat{r}(\text{x})]\}\ 2 + \text{Var}[\hat{r}(\text{x})] \leq C_*^2 \quad h^4 + \frac{C_{**}}{nh}$$

By choosing h such that

$$\text{h} = \left(\frac{c_{**}}{c_*^2}\right)^{\frac{1}{5}} n^{-\frac{1}{5}} \qquad (8)$$

5. CONCLUSION

In this paper, estimation the hazard rate function based on the Gumbel type 2 kernel function has been introduced, hazard rate of the Gumbel type 2 kernel estimator is characterized as free of boundary bias. We have then derived the bias, variance, and optimal bandwidth for hazard rate estimation. The theory proves the convergence of these quantities to zero under some conditions.

6. REFERENCES

[1] N. ALjayeh. On the inverse gaussian kernel estimator of the hazard rate function. The Islamic University of Gaza, 2016.

[2] H. S. Bakouch, C. Chesneau, and O. A. Elsamadony. The gumbel kernel for estimating the probability density function with application to hydrology data. Journal of Data, Information and Management, 3(4):261–269, 2021.

[3] S. X. Chen. Probability density function estimation using gamma kernels. Annals of the Institute of Statistical Mathematics, 52(3):471–480, 2000.

[4] R. A. Davis, K.-S. Lii, and D. N. Politis. Remarks on some nonparametric estimates of a density function. In Selected Works of Murray Rosenblatt, pages 95–100. Springer, 2011.

[5] A. C. Guidoum. Kernel estimator and bandwidth selection for density and its derivatives. Department of Probabilities and Statistics, University of Science and Technology, Houari Boumediene, Algeria, 2015.

[6] J. Rice and M. Rosenblatt. Estimation of the log survivor function and hazard function. Sankhy¯a: The Indian Journal of Statistics, Series A, pages 60–78, 1976.

[7] J. P. Romano. On weak convergence and optimality of kernel density estimates of the mode. The Annals of Statistics, pages 629–647, 1988.

[8] R. B. Salha. Estimating the density and hazard rate functions using the reciprocal inverse gaussian kernel. In International Conference, Matar´o (Barcelona), Spain 25-28 June 2013, number Proceedings, 15th Applied Stochastic Models and Data Analysis (ASMDA2013), 2013.

[9] R. B. Salha, H. I. El Shekh Ahmed, and I. M. Alhoubi. Hazard rate function estimation using weibull kernel. Open Journal of Statistics, 4(08), 2014.

[10] R. B. Salha, H. I. El Shekh Ahmed, and I. M. Alhoubi. Hazard rate function estimation using weibull kernel. Open Journal of Statistics, 4(08), 2014.

[11] M. P. Wand and M. C. Jones. Kernel smoothing. CRC press, 1994.

[12] W. Zucchini, A. Berzel, and O. Nenadic. Applied smoothing techniques. Part I: Kernel Density Estimation, 15:1–20, 2003.