Guided kernel density estimator and the gamma kernel estimator

Lule Hallaçi, Llukan Puka*

Abstract

Parametrically guided nonparametric estimation is a method that allows improving the bias of a nonparametric estimator by using a parametric pilot estimator. Talamakrouni (2016) generalize the parametrically guided nonparametric estimation to randomly right-censored data. The basic idea is to start with any parametric density estimator and then to adjust this first stage parametric approximation using a nonparametric kernel-type estimator of a particular correction factor. However, in many situations, using the classical kernel leads to the well-known boundary effect problem, that is, the estimator has a large bias near the endpoints. Bouezmarni (2011) proposed a gamma kernel estimator that corrects for the boundary effects. In this paper we perform a comparison between the guided kernel density estimator, based on Kaplan-Meier (1958) estimator and the gamma kernel estimator, for both the density and the hazard function via a Monte Carlo simulation, the finite sample performance of the estimators is investigated under various scenarios.

Keywords: Parametrically Guided Nonparametric Estimation, Guided Kernel Density, Gamma Kernel Estimator.
JEL Classification: C02, C14, C15.

Introduction

Censored data arise in many contexts, for example, in medical follow-up studies in which the occurrence of the event times (called survival) of individuals may be prevented by the previous occurrence of another competing event (called censoring). The estimation of the probability density and hazard function has received considerable attention in such studies, as it allows visualizing and exploring the distribution of data. There is a large variety of approaches to estimate the density and the hazard functions that are parametric, nonparametric, semi parametric and method which use aspects from both the nonparametric and the parametric school. Few of this method have been investigated in the presence of censoring mechanism.

The parametric approach has the advantage of being powerful by its $\sqrt{n}$ rate of convergence and also precise when the chosen family is correctly specified. However, a major complication that is emphasized in parametric modeling is the risk of biased and inconsistent parameter estimation due to misspecification problem. In the fully nonparametric approach, the estimators suffer from the curse of dimensionality and have in general a slower rate of convergence. However, despite its drawbacks, nonparametric approach provides more flexibility since the estimation is not based on any parameterized family of functions and remains more robust and applicable in practice.

Based on the Kaplan-Meier estimator, several nonparametric density estimators have been proposed in the literature. A popular approach for estimating the density function and the hazard rate function is done using a fixed symmetric kernel density with bounded support and a band width parameter, Blum and Susarla (1980). The kernel determines the shape of the local neighborhood while the bandwidth controls the degree of smoothness. Sabine and Stute (1988) investigated the kernel-type nonparametric estimator in the presence of right-censoring. However, when the density function of the data has a bounded support, using the classical kernel leads to an estimator with a large bias near the

* Lule Hallaçi, Llukan Puka, Department of Applied Mathematics, Faculty of Natural Science, University of Tirana, Tirana, Albania, *E-mail:*lule.hallaci@fshn.edu.al

endpoints. The problem of bias is called also the boundary effect. Boundary effects are well known to be a disturbing nuisance for applications as well as for global measures of performance of kernel estimators. The reason that boundary effects occur for unmodified kernel estimators is that the curve to be estimated has a discontinuity at an endpoint, so that the usual bias expansion which depends on smoothness assumptions cannot be carried out anymore. This is especially the case in survival analysis, since the survival time is assumed to be nonnegative variable. There have been various efforts to modify kernel estimators near boundaries in order to reduce the impact of these boundary effects. Bouezmarni (2011) proposed a gamma kernel (GK) estimator that corrects for the boundary effects.

In the fully nonparametric approach, the estimators have a slower rate of convergence. The parametric approach has the advantage of being powerful by its $\sqrt{n}$ rate of convergence but in parametric modeling is the risk of biased and inconsistent parameter estimation due to misspecification problem. Usually, even when the proposed model is misspecified, parametric estimation can provide valuable information about the phenomenon under study. This motivates the consideration of an approach called parametrically guided nonparametric estimation that contains both a parametric and a nonparametric component. The idea is to multiply an initial parametric density estimate with a kernel type estimate of the necessary correction factor. A guided nonparametric estimator is completely nonparametric in the sense that it does not rely on any assumed global structure. On the other hand, a guided nonparametric estimator takes advantage of both parametric and nonparametric methods: In the complete data case, considerable attention has recently been paid to parametrically guide nonparametric estimation in the literature. The starting point for this method was Hjort and Glad (1995), who introduced the parametric guided kernel (PGK) scheme and proved the bias reduction property of their guided estimator in the context of density estimation. Talamakrouni, Keilegom and Ghouch (2016) adapt and generalize the parametrically guided nonparametric estimation to the censored data case.

The paper is organized as follows. Section 2 introduces the gamma kernel estimators and parametrically guided nonparametric estimation for the density and the hazard rate function for right-censored data. In Section 3 we show the asymptotic properties. Via a Monte Carlo simulation, the finite sample performance of the estimators is investigated under various scenarios in Section 4.

Methodology

Let $T_1,.....,T_n$ (survival times) be independent and identically distributed (i.i.d) nonnegative random variables with density f and common distribution function F. Let $C_1,.....,C_n$ be a censoring variable with continuous distribution function G. Under random right censoring, instead of observing T_i, one can only observe (X_i,δ_i) where $X_i=\min(T_i,C_i)$ and $\delta_i=I(T_i\le C_i)$. Based on Kaplan-Meier estimator proposed by Kaplan and Meier (1958)several nonparametric density estimators have been proposed.

$$1-\hat{F}(t)=\prod_{i:X_i\le x}\left(1-\frac{1}{\sum_{j=1}^{n}1_{\{X_j\ge X_i\}}}\right)^{\delta_i} \tag{1}$$

Blum and Susarla (1980) extended the traditional kernel-type nonparametric estimator to censored data

$$\hat{f}(x)=\frac{1}{n}\int_{-\infty}^{+\infty}K(\frac{t-s}{h})d\hat{F}(s) \tag{2.2}$$

where K is a kernel function generally chosen to be a symmetric probability density function, $0 < h \equiv h_n$ is a bandwidth sequence and $\hat{F}(\square)$ is the Kaplan-Meier estimator.This method is totally nonparametric and admirably impartial to special types of shapes of the underlying density. However classical kernel leads to an estimator with a large bias near the endpoints. Bouezmarni (2011) proposed a gamma kernel estimator that corrects for the boundary effects, defined as follows:

$$\hat{f}_h(x) = \int K(x,h)(t)d\hat{F}(t) = \sum_{i=1}^{n} K(x,h)(X_{(i)})W_i \tag{2}$$

where the kernel K is given by

$$K(x,h)(t) = \frac{t^{\rho_h(x)-1}\exp(-t/h)}{h^{\rho_h(x)}\Gamma(\rho_h(x))}, \qquad \rho_h(x) = \begin{cases} \dfrac{x}{h} & \text{if } x \geq 2h \\ \dfrac{1}{4}\left(\dfrac{x}{h}\right)^2 + 1 & \text{if } x \in [0,2h) \end{cases} \tag{3}$$

The weights W_i are the jumps of $\hat{F}$ at X_i (Suzukawa et al. (2001)).

$$W_i = \frac{\delta_{[i]}}{n-i+1}\prod_{j=1}^{i-1}\left(\frac{n-j}{n-j+1}\right)^{\delta_{[j]}}, \quad i = 1,2,....,n \tag{4}$$

The gamma kernel estimator for the hazard rate is

$$h_h(x) = \frac{\hat{f}_h(x)}{1-\hat{F}(x)} \text{ (2.6)}$$

In the fully nonparametric approach, the estimators suffer from the curse of dimensionality and have in general a slower rate of convergence. However, nonparametric approach provides more flexibility since the estimation is not based on any parametrized family of functions and remains more robust and applicable in practice. The kernel estimator has a rate of convergence of $\sqrt{nh}$, (Lo eta al. (1989)) which is slower compared with the $\sqrt{n}$ rate of convergence established in the parametric approach. However, a major complication that is emphasized in parametric modeling is the risk of biased and inconsistent parameter estimation due to misspecification problem.

Hjort and Glad (1995) proposed a new scheme that contains both a parametric and a nonparametric component, called parametrically guided kernel density estimator. The essential idea behind the guided estimation is to start with a crude parametric estimator which is not necessarily well specified, then to correct this parametric guide using a particular type of correction and a nonparametric estimator. A guided nonparametric estimator takes advantage of both parametric and nonparametric methods. It always converges to the true model no matter if the parametric part is correct or not, and it adapt automatically to the parametric model if the latter is locally or globally close to the true underlying curve.

Talamakrouni, Keilegom and Ghouch (2016) adapt and generalize the parametrically guided kernel estimator to the censored data case defined as follows:

$$\hat{f}_{\hat{\theta}}(x) = \frac{1}{h}\int_{-\infty}^{+\infty} K\left(\frac{x-s}{h}\right)\frac{f_{\hat{\theta}}(x)}{f_{\hat{\theta}}(s)}d\hat{F}(s) = \frac{1}{h}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)\frac{f_{\hat{\theta}}(x)}{f_{\hat{\theta}}(X_i)}W_i \tag{5}$$

The parametrically guided kernel estimator for the hazard function is

$$\lambda_{\hat{\theta}}(x) = \hat{f}_{\hat{\theta}}(x)/(1-\hat{F}(x)) \tag{6}$$

Asymptotic Properties

In this section the performance of the guided kernel density estimator (5) is compared to that of the gamma kernel estimator (2). Both estimators have the bias reduction property and allows for a theoretically unbiased estimator. The multiplicative correction used in guided kernel density and hazard function does not affect the variance, the same for gamma kernel estimator. For parametrically guided kernel density estimator, the asymptotic bias and optimal bandwidth are:

$$B_{\theta^*}(x) = \frac{1}{2}h^2\mu_K^2\left(\frac{f(x)}{f_{\theta^*}(x)}\right)'' f_{\theta^*}(x) \text{ and } h_{opt} = \left(\frac{\int\sigma^2(x)dx}{\mu_K^4\int(r''(x)f_{\theta^*}(x))^2dx}\right)^{1/5} n^{-1/5}$$

For gamma kernel estimator the asymptotic bias and optimal bandwidth are:

$$B = \begin{cases} \frac{1}{2}xf''(x)h + o(h) + o(n^{-1/2}h^{-1/4}) & \text{if } x \geq 2h \\ \frac{(1-x)(\rho_h(x)-x/h)}{1+h\rho_h(x)-x} f'(x)h + o(h) + o(n^{-1/2}h^{-1/2}) & \text{if } x \in [0,2h) \end{cases}$$

$$\text{and } h_{opt} = \left(\frac{1}{4}\frac{\int\frac{1}{2\sqrt{\pi}}\frac{x^{-1/2}f(x)}{\overline{G}(x)}dx}{\int\frac{1}{2}xf''(x)dx}\right)^{2/5} n^{-2/5}$$

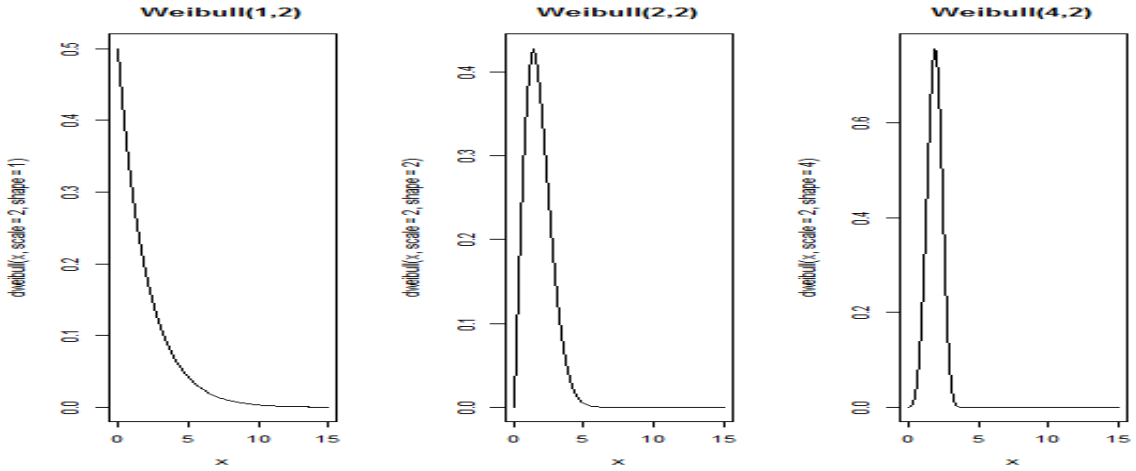
In practice, the choice of the bandwidth is a crucial issue in kernel-based density estimation. To select the bandwidth h in our case, we use unbiased cross validation method (Scott D. W. and Terrell G. R., 1987), adapted to the censoring case.

Simulations Results

In this section is studied the finite sample performance of the guided kernel density estimator and the gamma kernel estimator. Our goal is to compare the performance of the guided kernel density estimator (5) with that of the gamma kernel estimator (2) and traditional kernel. The comparison is based on Bias, MSE and the optimal bandwidth h.

The model considered is: the survival times follow a Weibull distribution with scale parameter *b=2* and shape parameter *a=1, 2, 4*. The graphs of the resulting densities are plotted in Figure 1.

Figure 1: Weibull density with shape parameters a = 1, 2, 4 and scale parameter b = 2.



The censoring times are also generated from a Weibull distribution with shape parameter *a* and a scale parameter given by $b((1-p)/p)^{1/a}$, ensuring a degree of censoring equal to *p*. We consider two censoring rates *p=10%* and *p=40%* and sample size n=200. For parametrically guided kernel density estimator, as a parametric guide we use the exponential density $f_\theta(t) = \theta \exp(-\theta t)$ where θ is estimated using the approximated maximum likelihood estimator. The only situation where the guide is correctly specified is the case a=1, in the other cases the parametric guide deviates gradually from the true density.

*Table 1: Squared bias (*10^5), MSE (*10^5), the optimal bandwidth h, for the estimators of several* Weibull densities for a= (1, 2, 4), two censoring rates p= (10%, 40%) and sample size n=200.

	10%					40%		
a	Method	Bias	MSE	h		Bias	MSE	h
1	PGK	0.09	110.7	8		1.67	189.7	7.9
	GK	0.009	112.5	8		0.67	187.6	8.01
	TK	22.98	99.2	9		28.65	260.4	9.01
2	PGK	89	260.7	4.2		89	445.1	4.2
	GK	87.99	244.3	4.5		87.99	446	4.46
	TK	110.14	270.1	4		110.14	589.23	4
4	PGK	87.2	344.2	3		87.2	587	3
	GK	86.51	304.3	2.28		86.51	524.2	3.2
	TK	210.5	689.99	3.9		210.5	890.98	4

We get the best results for the PGK estimator when a=1(a correct parametric guide). The bias of the PGK estimator is significantly reduced compared to that of TKestimator, but GK estimator gives a smaller bias compared to both of them. Regarding the MSE, it is also reduced for the PGK estimator compared to the MSE of the TK estimator. MSE is also reduced for the GK estimator compared to the MSE of the TK and PGK estimator. For a=2 and a=4, even if the parametric guide is incorrect, the PGK estimator remains significantly better than the TK estimator, while the GK estimator has a significantly smaller bias than both the PGK and TK estimator.

Along the simulations we consider the Gaussian kernel function K (Hansen B, 2009) and, for every estimator, we only show the results corresponding to the optimal tuning parameters, i.e. those which minimize the empirical mean squared error (MSE). The choice of the bandwidth is made by unbiased cross-validation bandwidth selection method, adapted to the censoring case.

Conclusion

In this paper, we investigated a parametrically guided kernel and a gamma kernel estimator for censored data. The PGK estimatoris obtained by multiplying an initial parametric estimator by a nonparametric kernel type estimator of a suitable correction function. The simulation results confirm the bias reduction property. We showed that the bias of the PGK estimator and GK estimator can be reduced compared to that of the traditional kernel estimator, while the GK estimator has a significantly smaller bias than both the PGK and TK estimator.

References

Blum, J. R., Susarla, V. (1980). Maximal deviation theory of density and failure rate estimates based on censored data. *J.Multiv.Anal. 5, pp.213-222*.

Bouezmarni, T., El Ghouch, A., Mesfioui, M. (2011, April 18). Gamma kernel estimators for density and hazard rate of right-censored data. *J. Prob. Stat. 2011*. DOI 10.1155/2011/937574

Glad, I. K., Hjort, N. L., Ushakov, N.G. (2003). Correction of density estimators that are not densities. *Scand. J. Stat. 30(2), pp. 415-427*.

Hansen B., (2009). *Kernel Density Estimation*. New York, NY: Springer.

Hjort, N.L. (1992). On inference in parametric survival data models. *Int. Stat. Rev. 60, pp. 355-387*.

Hjort, N. L., Glad, I. K. (1994). Nonparametric density estimation with a parametric start. Statistical research report. *Dept. Mathematics, Univ. Oslo.*

Hjort, N. L., Glad, I. K. (1995). Nonparametric density estimation with a parametric start. *Ann. Statist. 23, pp. 882-904*

Kaplan, E., Meier, P. (1958). "Nonparametric estimation from incomplete observations," *Journal of the American Statistical Association, vol. 53, pp. 457–481*.

Klein, J. P., Moeschberger, M. L. (1997). *Survival Analysis. Techniques for Censored and Truncated Data*. New York, NY: Springer.

Scott D. W.,Terrell G. R., 1987. Biased and unbiased cross-validation in density estimation. *J. Amer. Statist. Assoc. 82, pp. 1131–1146.*

Suzukawa, A., Imai, H., Sato, Y. (2001). Kullback-Leibler information consistent estimation for censored data. *Ann. Inst. Statist. Math. 53, pp. 262-276.*

Talamakrouni, M., Van Keilegom, I., El Ghouch, A. (2016). Parametrically guided nonparametric density and hazard estimation with censored data. *Comput. Statist. Data Anal. 93, pp. 308-323*