MTH516A: Non-Parametric Inference Project Report

Kernel density estimation for Skewed data

Udayan Nath(191168)

Vinay Kumar Sharma(191171)

Vishnudatt Nagar(181172)

Department of Mathematics and Statistics , IIT KANPUR

(DATE : 15 April 2021)

Abstract.

In probability and statistics, density estimation is the construction of an estimate, based on observed data, of an unobservable underlying probability density function. There are two kind of density estimation, parametric and nonparametric. In parametric the data are drawn from one of a known parametric family of distributions, the underlying density f is estimated by estimating the parameters. In nonparametric less rigid assumptions are made. In statistics, kernel density estimation (KDE) is a non-parametric way to estimate the probability density function of a random variable. Kernel density estimation is a fundamental data smoothing problem where inferences about the population are made, based on a finite data sample. If the data is Gaussian(or approximately Gaussian) in shape the usual kernel density estimation method(if we use Normal Or Eepanechnikov kernel function) gives good estimate of the density. But in real life data like financial or insurance data is heavy tailed, in this data usual estimation method performs poorly. Using transformations we can solve the problem.

Introduction

In most situations,losses are small and extreme losses are rarely observed, but the number and the size of extreme losses can have a substantial influence on the profit of a insurance company. The proposed estimator is obtained by transforming the original data set with a parametric transformation and afterwards estimating the density of the transformed data set using the classical kernel density estimator described in the previous section. For a given data point kernel density estimator is,

$$f_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{(x - x_i)}{h}$$

Here K is the kernel function. K has several properties like non negative even function and if we intregrate over R it will give 1. There are several kernel function like uniform, triangular, gaussian and epanechnikov etc. Here we use epanechnikov kernel in this project. h is the bandwidth and it is chossen in such a way that there is a trade of between bias and variance of the estimator. Here h is choosen using Silverman rule of thumb((($h = ((\frac{4\hat{\sigma}^5}{3n})^{\frac{1}{5}}$),$\hat{\sigma}$ is sample standard deviation).

Mobius-like transformation

$$T_{\alpha,R}(x) = \frac{x^\alpha - R^\alpha}{x^\alpha + R^\alpha}, \forall x \in [0, \infty] \tag{1}$$

where $\alpha > 0, R > 0$, as We used the Möbius like transformation, scale parameter R ,
for **highly skewed data** Let $x_{(1)} < x_{(2)} < ... < x_{(n)}$ a sample of n iid observations of a random variable X taking values in $[0, \infty)$. For positive α and R, the function (1) $Z = T_{\alpha,R}(x)$maps the

original sample into $z_1 < z_2 < ... < z_n$ where $z_i \in [-1, 1]$. The probability density function $f(x)$ of X and $f(z)$ of Z are connected by the formula,

$$f_X(x) = f_Z(z)\frac{2\alpha R^\alpha x^{\alpha-1}}{(x^\alpha + R^\alpha)^2} = \frac{\alpha}{2R} f_Z(z)(1+z)^{1-\frac{1}{\alpha}}(1-z)^{1+\frac{1}{\alpha}} \tag{2}$$

Algorithm for Optimal α and R

The algorithm is as follows:
(i) Arrange the observations $x_1, x_2, \ldots, x_n$ in increasing order.
(ii) Provided a starting value of R for a one dimensional minimization algorithm. Start with the median of the original data and also take a small positive parameter $\in = 10^{-5}$ to be used to terminate the contraction mapping to estimate α.
(iii) Provide a starting value of α, say $\alpha_{new} = 1$ choice of α is insignificant as long as it is small.
(iv) Set $\alpha_{old} = \alpha_{new}$; generate the data $z_i = T_{\alpha,R(x_i)}, i = 1, 2, \ldots, n$ and compute σ_z, the standard deviation of z_i's.
(v) Calculate,

$$\alpha_{new} = \frac{(ln\beta\sigma_z) - (ln2 - \beta\sigma_z)}{ln\frac{R}{x_{(n)}}}$$

,where $\beta = \frac{1.059\sqrt{5}}{n^{\frac{1}{5}}}$. If $|\alpha_{old} - \alpha_{new}| > \epsilon$ i.e if the contraction has not yet converged return to step (iv). Otherwise take $\hat{\alpha} = \alpha_{new}$ and go to step (vi).
(vi) The minimization algorithm chooses another value of R and return to step (iii) or the minimization is complete and optimal value of α and R is obtained. Compute $z_1, z_2, \ldots, z_n$ by the transforming the original data using these optimal estimates.
(vii) Given any points x, compute z using the optimal transformation. Estimate density at z using Epanechnikov kernels bandwidth based on $z_1, z_2, \ldots, z_n$, The density at x is now found by scaling this estimate using the equation(2) as described above.

The mapping in fact is a contraction and such $\hat{\alpha}$ exists with taking a small value as starting value of α is proved in the paper by Clements et al. The algorithm provided above is also discussed in the same paper. As, median is a robust estimator we took the value of R as the static value as the median of the original data which also gives fine result. This above discussed algorithm is used later for simulation studies and also to analyze the two data sets based on income and car insurance. The form of the final transformed estimator for the original data $x_1, x_2, \ldots, x_n$ is as follows,

$$f(x) = \frac{1}{n}\sum_{i=1}^{n}(K_h(T(x) - T_i(x))T^{'}(x)$$

Where, T(.) is the transformation function, $T^{'}$ is the first differential and K_h is the kernel function with bandwidth h.

Asymptotic Theory for Transformation

For the $x_1, x_2, \ldots, x_n$ transformed density estimator

$$f(x) = \frac{1}{n}\sum_{i=1}^{n}(K_h(T(x) - T_i(x))T^{'}(x)$$

Where, T(.) is the transformation function, $T^{'}$ is the first differential and K_h is the kernel function with bandwidth h. Now, the bias and the variance of $f(x)$ are given by,

$$Bias = E[f(x)] - f(X) = \frac{1}{2}\mu_2(K)h^2((\frac{f(x)}{T^{'}(x)})\frac{1}{T^{'}(x)}) + 0(h^2)$$

$$Variance = V(f(x)) = \frac{1}{nh}R(K)T^{'}(x)f(x) + 0(\frac{1}{nh})$$

where $\mu_2(K) = \int(u^2K(u)du$ and $R(K) = \int(K^2(u)du$. Now, as for the classical density estimation the classical density estimator follows a normal distribution asymptotically as $n \longrightarrow \infty$:

$$\sqrt(nh)(\hat{g(z)} - E(g(z)) \sim N(0, R(K)g(y))$$

then as $\hat{f}(x) = \hat{T^{'}(x)}g(x)$ with $z = T(x)$. Then

$$\sqrt(nh)(f(x) - E(f(x)) \sim N(0, R(K)T^{'}(x)f(x))$$

where $z_i = T(x_i)$ i.e. transformed variable have distribution g and $\hat{g(z)}$ is the classical kernel density estimate of $g(z)$.
The proofs regarding the asymptotic theory of classical kernel density estimator is done in the book by E.L.Lehmann and the proof for the transformed kernel density estimate is in the paper by Buch-Larsen et al.

Comparison between Untransformed and Transformed Kernel Density Estimates using Simulation and a Real Data

Kullback-Leibler Divergence In mathematical statistics the KL divergence(D_{kl}) is a measure of how one probability distribution is different from another probability distribution. But in our problem we interpret this value in a different way. In the context of machine learning $D_{kl}(P||Q)$ is often called the information gain achieved if density P would be used instead of density Q. The mathematical expression of KL Divergence is,
for discrete probability distribution P and Q

$$D_{kl} = \sum_{x\in\chi} P(x)\log(\frac{P(x)}{Q(x)})$$

for Continuous probability distribution P and Q

$$D_{kl} = \int_{x\in\chi} P(x)\log(\frac{P(x)}{Q(x)})dx$$

The Integrated squared error (**ISE**) is defined as,

$$ISE = \int_0^\infty (f_h\hat{(x)} - f(x))^2\,dx$$

Here $\hat{f(x)}$ is estimated density and $f(x)$ is true density. ISE weighs errors of the estimator near 0 and in the tail equally.

The Weighted integrated squared error (WISE) is defined as,

$$WISE = \int_0^\infty (\hat{f_h(x)} - f(x))^2 x^2 \, dx$$

Here $\hat{f(x)}$ is estimated density and $f(x)$ is true density. The Weighted integrated squared error (**WISE**) is generalization of ISE where, the weighting function is x^2 Therefore, WISE is giving more importance to deviations in the tail. This also has a practical interpretation, since the x values correspond to financial data or payments to be made by the insurer.

The comparison is based on the data simulated from the two different distributions (Log normal and Wiebull) and for three sample sizes n= 250, n= 500 and n= 1000 . Each combination of distribution is replicated. In the table below the error measurements and KL divergence are given based on the replicated data:

Sample size	Error	Probability Density			
		Weibull		Log Normal	
		Untransformed	Transformed	Untransformed	Transformed
n = 250	ISE	0.2828	0.0022	0.3373	0.0036
	WISE	0.2744	0.0043	0.3675	0.0020
	KL	0.0799	0.5072	0.0170	0.3742
n = 500	ISE	0.2745	0.0015	0.3373	0.0036
	WISE	0.2756	0.0021	0.3675	0.0020
	KL	0.0844	0.5209	0.0171	0.3750
n = 1000	ISE	0.2651	0.0009	0.3374	0.0036
	WISE	0.2721	0.0014	0.3676	0.0021
	KL	0.0832	0.5312	0.0171	0.3752

For real life data case as we don't know the original distribution $f(x)$ so we can't calculate ISE and WISE. So here we only calculate KL divergence value.

The value of KL divergence for untransformed and transformed data is respectively 0.000002537 and 0.000102.

Conclusion

Real life data sets can not be approximated by Gaussian form in every situation. So, using transformation can give a better estimate of the density in many problems. We clearly conclude from our analysis that for both the cases simulated as well as real life data(Car Insurance Data), transformation gives us better results for estimation of density.

References

[1] Clements, A.E., Hurn, A.S. and Lindsay, K.A.., 2003, Möbius-like Mappings and their use in kernel density estimation.*Journal of the American Statistical Association*,98, 993-1000.

[2] Buch-Larsen, T., Nielsen, J.P., Guillén, M. and Bolancé, C., 2005, Kernel density estimation for heavy-tailed distributions using the Champernowne transformation, *Statistics*,39, 503-518

[3] Lehmann, E.L., 1998, Elements of Large-Sample Theory(Springer)

[4] Some reference from wikipedia on Kernel density estimator.