Applied Mathematical Sciences, Vol. 7, 2013, no. 142, 7053 - 7063
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ams.2013.311625

Simulation Study for Boundary Effect Density Estimation

Kartiko

Mathematics Department, Sebelas Maret University, Surakarta, Indonesia
Graduate Student Math. Dept., Gadjahmada University, Yogyakarta, Indonesia

Suryo Guritno, Dedi Rosadi and Abdurakhman

Mathematics Department, Gadjahmada University, Yogyakarta, Indonesia

Abstract

Density estimation based on data is often use in engineering, finance, biomedical and social science data. People use parametric fit which specify the form of the density in advance, which very often is incorrect. Along with the development of computing software people can easily do the nonparametric methods, that rely heavily on computing but free from model assumptions.

Kernel smoothing refers to a general class of techniques for nonparametric estimation of density functions. For certain univariate set of data that one wants to display graphically, using kernel function as the weight, this method can be used and is known as kernel density estimation.

Estimation in the boundary points suffer a large bias, however a special treatment is needed. Simulation study is conducted to see that champernowne transformation can do the job properly.

Mathematics Subject Classification: 62G08

Keywords: kernel, boundary points, Champernowne

1 Introduction

Buch-Larsen et.al.[4] used modified Champernowne distribution to estimate loss distributions in insurance which is categorically heavy-tailed distributions. Some time it is difficult to find a parametric model which is simple and fit for all values of claim in the insurance industry. Gustafsson et.al.[6] used asymmetric kernel density estimation to estimate actuarial loss distributions. The estimator is obtained by transforming the data using generalized Champernowne distribution. Then by using local asymmetric kernel methods the density of the transformed data is estimated to obtain superior estimation properties in the tails. Kromann [2] introduced a new tail-dependent parameter estimation method for the Champernowne distribution, computed by conditional maximum likelihood, and showed that, by using this new method, they obtained an estimator that in general outperforms the benchmark estimators with respect to tail performance.

Jones [7] estimated a probability density function which has bounded support using kernel density estimation often overspill the boundaries and the estimator are biased at and near these edges. He consider a simple unified framework is provided which covers a number of straightforward methods and allows for their comparison: generalized jackknifing generates a variety of simple boundary kernel formula. In his paper Karunamuni [9] proposed a new general method of boundary correction for univariate kernel density estimation. The proposed method generates a class of boundary corrected estimators. They all possess desirable properties such as local adaptivity and non-negativity.

Chen [5] proposes gamma kernel smoothers for estimating curves with compact support since gamma kernel is non negative and have natural varying shape. Sayah et.al.[10] produce a kernel quantile estimator for heavy-tailed distributions using a modification of the Champernowne distribution.

Kromann [4] approach a liability data base on Champernowne distribution modification and applying the result to an actual data set.

Kernel density estimator which is of the form

$$\hat{f}(x) = (nh)^{-1} \sum_{i=1}^{n} K\{(x - X_i)/h\}, \tag{1.1}$$

In this paper a step by step discussion, which was done by Buch-Larsen et.al.[4], will be shown by constructing some lemmas and theorems, and finally simulation will be performed for estimating heavy tail distribution.

2 Champernowne Distribution

Buch-Larsen et.al.[4], the original Champernowne distribution has density

$$f(x) = \frac{c}{x\left((1/2)(x/M)^{-\alpha} + \lambda + (1/2)(x/M)^{\alpha}\right)}, \quad x \geq 0$$

where c is a normalizing constant and α, λ and M are parameters.The distribution was mentioned for the first time in 1936 by D.G. Champernowne when he spoke on The Theory of Income Distribution at the Oxford Meeting of the Econometric Society. Then, he gave more details about the distribution, and its application to economics. When λ equals to one and the normalizing constant c equals $(1/2)\alpha$, the density of the original distribution is simply called the Champernowne. Champernowne cumulative distribution function is defined on $x \geq 0$ and has the form

$$T_{\alpha,M}(x) = \frac{x^{\alpha}}{x^{\alpha} + M^{\alpha}} \tag{2.2}$$

with parameter parameter $\alpha > 0$,$M > 0$, and density function is of the form

$$t_{\alpha,M}(x) = \frac{\alpha M^{\alpha} x^{\alpha-1}}{(x^{\alpha} + M^{\alpha})^2}. \tag{2.3}$$

Lemma 2.1. *M is the median of the Champernowne distribution, therefore its estimator is the empirical median.*

Proof:

$$T_{\alpha,M}(x) = \frac{x^{\alpha}}{x^{\alpha} + M^{\alpha}}$$

let $X = M$ then

$$\begin{aligned} T_{\alpha,M}(M) &= \frac{M^{\alpha}}{M^{\alpha} + M^{\alpha}} = \frac{M^{\alpha}}{2M^{\alpha}} = \frac{1}{2} \\ P(X \leq M) &= \frac{1}{2}, \text{ we may conclude that } M \text{ is median} \end{aligned}$$

■

Champernowne probability density function has thick tail. Figure 1 shows several probability density function for different α and fix M. It can be easily proved that M is the median. The picture also shows that smaller α gives thicker tail. While Figure 2 gives their corresponding cumulative distribution function. It can be seen from Figure 2 that

$$\text{for } \alpha < \alpha' \begin{cases} T_{\alpha,M}(x) > T_{\alpha',M}(x) & \text{for } 0 \leq x < M \\ T_{\alpha,M}(x) = T_{\alpha',M}(x) & \text{for } x = M \\ T_{\alpha,M}(x) < T_{\alpha',M}(x) & \text{for } M < x < \infty \end{cases}$$

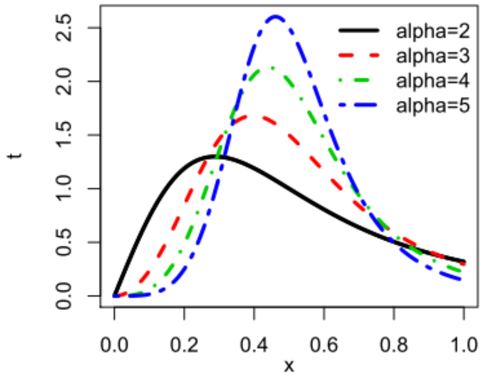

Figure 1: Champernowne Probability Distribution Function, M=0,5

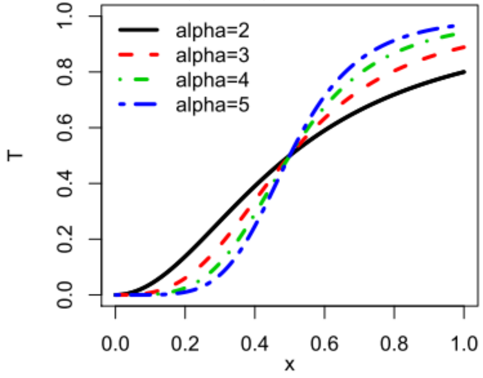

Figure 2: Champernowne Cumulative Distribution Function, M=0,5

Actually α is not a scale parameter, but has a property like a scale parameter, as can be seen from Figure 1, that for $\alpha > 1$, increasing α gives a more steep derivation of the *cdf* in the value of M, the mode move to the right, and the tail become lighter. For $\alpha < 1$, increasing α gives less steep shape of the density near 0.

The effect of parameter M is that for $M < M'$ resulting $T_{\alpha,M}(x) > T_{\alpha',M'}(x)$, or increasing M lower the *cdf* and for $\alpha > 1$ the mode moves to the right and getting smaller.

Modified Champernowne distribution is defined on $x \geq 0$ and formulated as

$$T_{\alpha,M,c}(x) = \frac{(x+c)^\alpha - c^\alpha}{(x+c)^\alpha + (M+c)^\alpha - 2c^\alpha} \tag{2.4}$$

with parameter $\alpha > 0, M > 0$ and $c \geq 0$ and its density is

$$t_{\alpha,M}(x) = \frac{\alpha(x+c)^{\alpha-1}((M+c)^\alpha - c^\alpha}{((x+c)^\alpha + (M+c)^\alpha - 2c^\alpha)^2} \tag{2.5}$$

It is shown on Figure 3 and 4,

$$\text{for} \quad c < c' \quad \text{and for} \quad \alpha > 1 \begin{cases} T_{\alpha,M,c}(x) < T_{\alpha',M,c'}(x), & \text{for} \quad 0 \leq x < M \\ T_{\alpha,M,c}(x) = T_{\alpha',M,c'}(x), & \text{for} \quad x = M \\ T_{\alpha,M,c}(x) > T_{\alpha',M,c'}(x), & \text{for} \quad M < x < \infty \end{cases}$$

$$\text{while for} \quad \alpha < 1 \qquad \begin{cases} T_{\alpha,M,c}(x) > T_{\alpha',M,c'}(x), & \text{for} \quad 0 \leq x < M \\ T_{\alpha,M,c}(x) = T_{\alpha',M,c'}(x), & \text{for} \quad x = M \\ T_{\alpha,M,c}(x) < T_{\alpha',M,c'}(x), & \text{for} \quad M < x < \infty \end{cases}$$

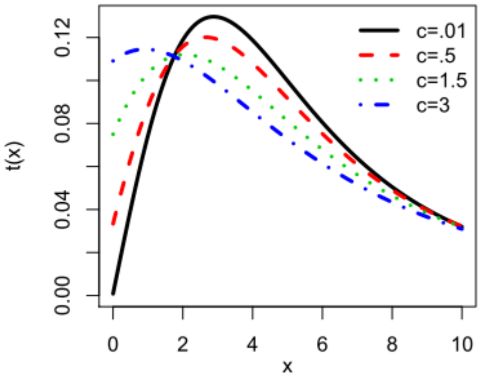

Figure 3: Modified Champernowne Probability Distribution Function, M=5, $\alpha = 2$

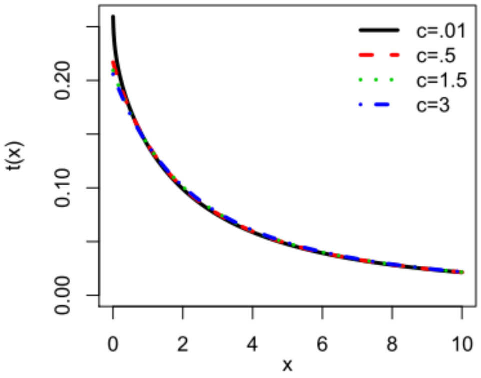

Figure 4: Modified Champernowne Probability Distribution Function, M=5, $\alpha = 0.95$

When $\alpha \neq 1$, c acts as "scale parameter", different value of c gives changes to the tail density. In particular for $\alpha < 1$, increasing c produces thinner tail. The opposite for $\alpha > 1$. The greater c moves the mode to the right. While for $\alpha = 1$, c does not give any effects.

3 Density Estimation Using Champernowne Transformation

Consider a sample random of size n, $X_1, X_2 \ldots, X_n$, from unknown *cdf*, F or (*pdf*, f). As mention in section 2 the estimator of M is the empirical median. Likelihood function of the *pdf* given by equation (2.5) is

$$\begin{aligned} l(\alpha, c) &= n \ \log(\alpha) + n \ \log((M+c)^\alpha - c^\alpha) + (\alpha - 1)\sum_{i=1}^{n} \log(X_i + c) \\ &\quad -2\sum_{i=1}^{n} \log((X_i+c)^\alpha + (M+c)^\alpha - 2c^\alpha) \end{aligned} \tag{3.6}$$

Lemma 3.1. *The likelihood estimator of α and c are the values which maximize likelihood function*

Lemma 3.2. *Using transformation $y = T(x)$ (simplified notation for $T_{\alpha,M,c}(x)$) then $f_Y(y) = f_X(T^{-1}(y))\frac{1}{|t(T^{-1}(y))|}$ and $f_X(x) = f_Y(T(x))t(x) = f_Y(T(x))\frac{1}{|(T^{-1})'(x)|}$ where $t(x) = T'(x)$.*

Proof: For $y = T(x)$, $x = T^{-1}(y)$ and $t(x) = \frac{dT(x)}{dx}$,

$$\begin{aligned} f_Y(y) &= f_X(T^{-1}(y))|\frac{dT^{-1}(y)}{dy}| \\ &= f_X(T^{-1}(y))\frac{1}{|\frac{dT(x)}{dx}|} \\ &= f_X(T^{-1}(y))\frac{1}{|t(T^{-1}(y))|} \end{aligned}$$

For $x = T^{-1}(y)$, $y = T(x)$,

$$\begin{aligned} f_X(x) &= f_Y(T(x))|\frac{dT(x)}{dx}| \\ &= f_Y(T(x))|t(x)| \\ &= f_Y(T(x))\frac{1}{|(T^{-1})'(x)|} \end{aligned}$$

■

Theorem 3.3. *Given a set of data $X_1, X_2 \ldots, X_n$, cdf $T_{\alpha,M,c}$, Modified Champernowne distribution, then*

$$Z_i = T_{\alpha,M,c}(X_i), \quad i = 1, \ldots, n.$$

are new variable, Z_i is in the interval $(0, 1)$ and uniform distributed. and the kernel density estimation for transforms data

$$\hat{f}_{trans}(z) = f_X(T^{-1}_{\alpha,M,c}(z))\frac{1}{|t(T^{-1}_{\alpha,M,c}(z))|}, \tag{3.7}$$

And

$$\hat{f}_{trans}(x) = f_Z(T_{\alpha,M,c}(x))\frac{1}{|(T^{-1}_{\alpha,M,c})'(x)|} \tag{3.8}$$

Formulation of the kernel density estimation for transform data $\{Z_i\}_{i=1}^n$ is

$$\hat{f}_{trans}(z) = \frac{1}{nk_z}\sum_{i=1}^{n} K_h(z - Z_i),$$

where $K_h = (1/h)K(.)$ and $K(.)$ is kernel function. Boundary correction, k_z is needed since z are in the interval $(0, 1)$ so that we have to divide by the area under the kernel that lies in this interval, which defined by

$$k_z = \int_{max(-1,-z/h)}^{max(1,(1-z)/h)} K(u)du. \tag{3.9}$$

Using Theorem 3.3 kernel density estimation for data $X_i,\ i = 1, \ldots, n$ is;

$$\hat{f}(x) = \frac{\hat{f}_{trans}(T_{\hat{\alpha},\hat{M},\hat{c}}(x))}{\left|(T^{-1}_{\hat{\alpha},\hat{M},\hat{c}})'(x)\right|}$$

In the $KMCE$ the formula of kernel density estimation is

$$\begin{aligned}\hat{f}(x) &= \frac{1}{nk_{T_{\hat{\alpha},\hat{M},\hat{c}}(x)}}\sum_{i=1}^{n} K_h(T_{\hat{\alpha},\hat{M},\hat{c}}(x) - T_{\hat{\alpha},\hat{M},\hat{c}}(X_i))T'_{\hat{\alpha},\hat{M},\hat{c}}(x) &(3.10)\\ &= \frac{1}{nk(T(x))}\sum_{i=1}^{n} K_h(T(x) - T(X_i))T'(x) &(3.11)\end{aligned}$$

Buch-Larsen et.al.[4], Karunamuni et. al. [9] presented a theorem about the asymptotic theory of the transformation kernel density estimator in general (asymptotic bias and variance).

Theorem 3.4. *Let $X_i,\ i = 1, \ldots, n$ be independent identically distributed with density f and $\hat{f}(x)$ is the transform kernel density estimator of $f(x)$*

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

with $T(.)$ is the transform function. Then the bias and variance of $\hat{f}(x)$ is

$$E[\hat{f}(x)] - f(x) = \frac{1}{2}\mu_2(K)b^2\left(\left(\frac{f(x)}{T'(x)}\right)'\frac{1}{T'(x)}\right)' + o(h^2).$$

$$V[\hat{f}(x)] = \frac{1}{nh}R(K)T'(x)f(X) + o\left(\frac{1}{nh}\right)$$

for $n \to \infty$, $\mu_2(K) = \int u^2K(u)du$ and $R(K) = \int K^2(u)du$.

4 Simulation Study

Simulation study is done based on log normal distribution

$$f(x) = \frac{1}{x\sigma\sqrt{2\pi}}\exp(-(\ln\ x - \mu)^2/2\sigma^2) \tag{4.12}$$

with mean $\mu = 0.01$ and $\sigma = 0.5$ several observations added to thicker the right tail. Figure 5 present the log normal pdf with mean 0,01 and variance 0.25. Sample of size 200 is taken from this distribution. The histogram and kernel density estimation are plotted on the same axes.

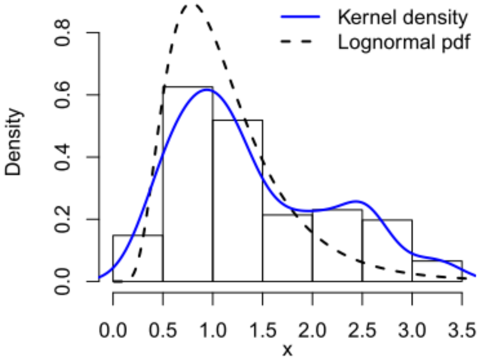

Figure 5: Log normal pdf, histogram of sample size 200, kernel density estimation

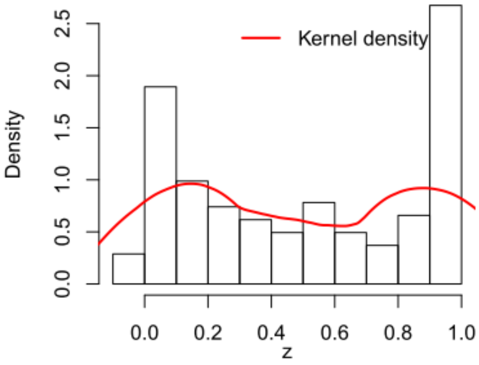

Figure 6: Kernel density estimation of transformed variable

Figure 6 shows the histogram and density estimation of the transformed data, while Figure 7 shows the histogram and density estimation of the transformed data with correction factor.

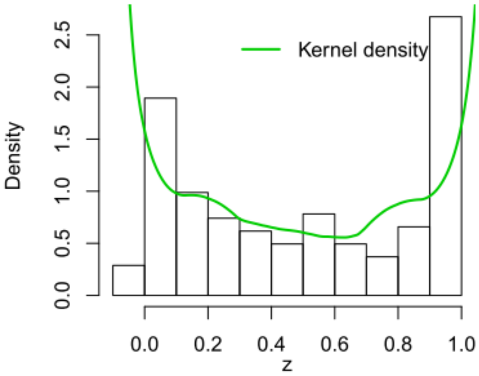

Figure 7: Kernel density estimation of transformed variable using correction factor

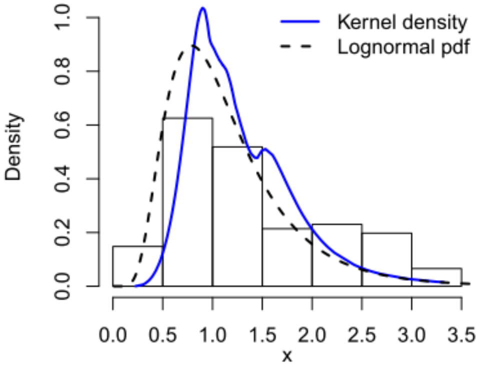

Figure 8: Kernel density estimation of inverse transformed variable

Figure 8 shows the inverse transform of the density estimation plotted on the same axis with the histogram of the original data.

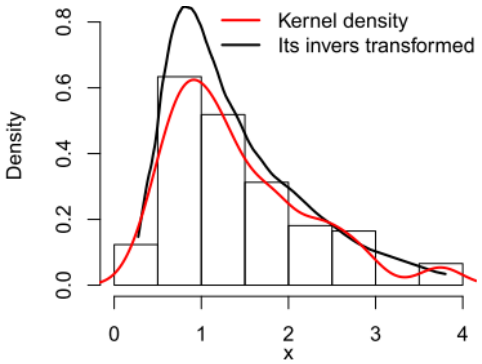

Figure 9: Kernel density estimation vs its inverse transformed

References

[1] Ackley, D.H., Hinton, G.E. and Sejnowski, T.J. (1985), A learning algorithm for Boltzmann machine, *Cognitive Science,* **9**, 147 - 169.

[2] Buch-Kromann, T. (2009), Comparison of tail performance of the Champernowne transformed kernel density estimator, the generalized Pareto distribution and the g-and-h distribution ,*The Journal of Operational Risk,* **4(2)**, 43-67.

[3] Buch-Kromann T. (2006), Estimation of large insurance losses: A case study, *Journal of Actuarial Practice,* **12,** 191-211.

[4] Buch-Larsen, T.,Guillen, M.,Neilson, J.P. and Bolance,C. (2005), Kernel density estimation for heavy-tailed distributions using the Chempernowne transformation, *Statistics,* **39**, 503-518.

[5] Chen, S. X. (2000), Probability density function using Gamma kernel, *Ann. Inst. Statist. Math,***52(3)**, 471-480.

[6] Gustafsson, J.M., Hagmann, J.P. Nielsen, Scallet, O., Local transformation kernel density estimation of loss distributions, *National Centre of Competence in research Financial Valuation and Risk Management*(2007)

[7] Jones, M. C. (1993), Simple Boundary Correction for Density Estimation, *Statistics and Computing,***3,**135-146.

[8] Jones, M. C. and Foster, P. J. (1996),A simple nonnegative boundary correction method for kernel density estimation, *Statistica Sinica,***6,** 1005-1013.

[9] Karunamuni, R.J. and Alberts T. (2004), On boundary correction in kernel density estimation, *Fifth Biennial IISA International Conference on Statistics, Probability and Related Areas,* **May**, 14-16.

[10] Sayah, A., Yahia, D. and Necir, A. (2010), Champernowne transformation in kernel quantile estimation for heavy-tailed distributions, *Journal Afrika Statistika,* **5(12)**,288-296.

Received: November 1, 2013