Summary of A nonparametric approach to calculating value-at-risk

Oscar Johansson
oscar.johansson@student.kuleuven.com

1 Introduction

One of the first things statisticians learn when observing and measuring real data is that it often does often not behave according to our probabilistic assumptions. Often the assumptions made in parametric statistics are not fulfilled and we have to approach and look at the problem in a new way. A standard example of this is heavy tailed distributions(Cooke and Nieboer, 2011) commonly found finance(McNeil, 1998), insurance risk(Bolancé et al., 2012) and many other applied areas (Maillart and Sornette, 2010; Cooke and Nieboer, 2011; Maillart and Sornette, 2010). In this case we have that the density distributions have a large "tail". Logically we can argue this from an example, seen in figure 1 we have the standardized log return on the S&P500 15 years. In this we have certain extreme events like the 2008 crash on 13th of October and the covid-19 crash on the 24th of March which caused severe jumps in the S&P500. These rare, but not uncommon, extreme events makes the distribution have more extreme tails than the normal distribution estimate as can be seen in the figure. You also observe that the distribution is much sharper than the normal approximation

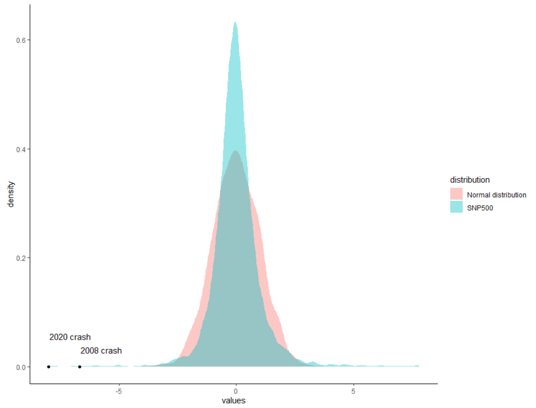

Figure 1: Distribution of S&P500 compared to a normal distribution. (gaussian kernel density)

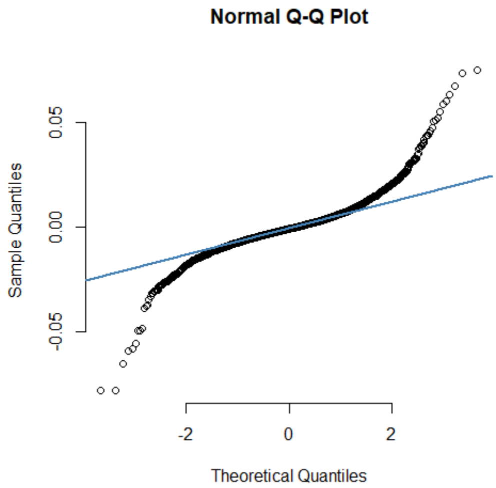

Figure 2: QQplot between empirical cdf S&P500 logreturn and normal distribution.

These heavy tailed are usually more seen studying the QQ-plot of our distribution as can be seen in figure 2. Here we compare the empirical distribution of our sample to the cdf of the normal distribution. As can be seen the data seem to follow the the normal cdf reasonable but deviates towards the ends of the cdf. This implies, and is often used as the evidence for, a heavy-tailed distribution.

There are a couple of approaches to solve the problem of heavy-tailed distributions. We have parametric distributions(McNeil, 1998), extreme value theory(Hill, 1975) or non-parametric solutions(Jones and Zitikis, 2007). These solutions are options for solving the problem of modeling the heavy tail, but some-times you are not interested in the distribution but usually properties of the distribution. A common measure to quantify risk is the so called *value at risk* measure, defined as

$$VaR_\alpha(X) = inf\{x : F_X(x) \geq \alpha\} = F_X^{-1}(\alpha). \quad (1)$$

One can see the VaR_α as the most extreme out value to occur given a safety margin α. Note the the VaR_α is primarily used for measures on one-sided distribution, estimating the maximum value below the α percentile. The interpretation is that we should expect that x will be below $VaR_\alpha(x)$ to a $100 * \alpha\%$ probability. As can be understood the VaR_α is very sensitive towards heavy-tails, and having a high accuracy towards the tails is very important to get an accurate measure of the VaR_α. In the case of the S&P data VaR_α it is often used to the α percentile of loss made on an investment in the S&P500. In table 1 you see the difference from calculating the VaR_α for the empirical distribution of S&P500 data and the approximate normal approximation for a series of :s. As seen the results vary, but that the heavy tail factor makes a normal approximation likely to undershoot the actual VaR_α, even if we never may know the true value and if it exists.

	Empirical S&P500	Normal
$\alpha = 0.1$	-0.96	-1.28
$\alpha = 0.05$	-1.35	-1.64
$\alpha = 0.01$	-2.54	-2.33
$\alpha = 0.001$	-5.99	-3.09

Table 1: Calculation of VaR_α for different levels of α and using the empirical cdf of our S&P500 data compared to normal distribution.

From this quick introduction you observe that to accurately calculate the VaR you need to accurately depict the cumulative distribution function. Well actually, you do not need to accurately depict the cdf accurately globally, you need to accurately approximate it around the area of $F^{-1}(\alpha)$. This is where the paper "A non-parametric approach to calculating value-at-risk"(Alemany et al., 2013) comes in. In this paper Alemany et al. present a method of to calculate the VaR_α using a double transformed kernel estimation (DTKE) method. They also aim to provide a rule of thumb bandwidth parameter selection for this kernel estimation. This method will be shortly explained in the following section and later the results will be presented in the result section. This will be followed by a discussion on the paper, what it does good and what parts may need further explanation andclarification. Finally the summary will use a practical example of implementing this algorithm and conclude its implication in terms of what it would imply in a real application.

2 Summary

In the paper, Alemany et al. largely base their approach on a paper by **??**, where Bolancé et al. present an approach of transforming the variables before estimating the cdf. The purpose of doing these transformations is to reduce variance, but this often comes with consequences of an increased bias in the estimate**??**. To further this explanation, Alemany et al seek to minimize the approximate mean integrated square error (AMISE) in order obtain a smoothing parameter that is asymptotically optimal it is sufficient to minimize

$$\begin{aligned} AMISE =& \frac{1}{n}\int F_X(x)[1-F_X] \\ &- \frac{1}{n}b(t)[1-K(t)]dt \\ &+ \frac{1}{4}b^4 \int [f'_X(x)]^2 dx \left(\int t^2 k(t)dt \right)^2 \end{aligned} \quad (2)$$

with respect to $\int [f'_X(x)]^2$ through a transformation $M : X \to Y$ that minimizes $[f'_Y(x)]^2$ (note the subindex of Y). From previus research by Terral(Terrell, 1990) it has been shown that the density of a $Beta(3,3)$ distribution scaled to the interval $[-1,1]$ minimize $\int [f'_Y(x)]^2$ in the set of all densities with known variance. The cdf and pdf of this distribution is described as:

$$m(x) = \frac{15}{16}(1 - x^2)^2$$
$$M(x) = \frac{1}{16}(15x - 5x^3 + x^5 + 11) \tag{3}$$

Here I actually found an error in the paper where they got the cdf by integrated the pdf from 0 to x and not $-\infty$ to x yielding

$$M(x) = \frac{3}{16}x^5 - \frac{5}{8}x^3 + \frac{15}{16}x + \frac{1}{2}. \tag{4}$$

Observe that the interval $[-1, 1]$ is in this case used based on the paper (Terrell, 1990), and that we simply scale the betadistriubtion on $[0, 1]$ by a variable transformation $t = 2x - 1$. I plotted the different functions in figure 3 and as seen the error cause the cdf to have negative values at $x = -1$ as the incorrect.

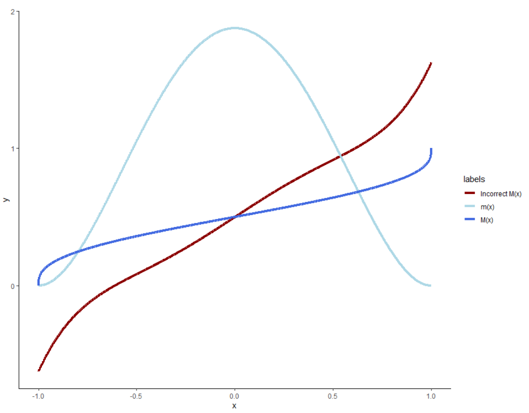

Figure 3: Correct cdf, pdf and the incorrect cdf of the $Beta(3, 3)$ distribution used as transformation.

The error is only calculation wise, and does not affect the theory underlying the assumption. The error is also found in the original paper(Buch-Larsen et al., 2005). The results are not impacted by this as a negative cdf would result in invalid results. Since the $Beta(3, 3)$ distribution is defined on the interval $[0, 1]$ Alemany et al. used an other transformation, called the generalized Champernowne distribution proposed by Buch Larsen et al.(Buch-Larsen et al., 2005) in 2005 defined as

$$T(x) = \frac{(x + c)^\delta - c^\delta}{(x + c)^\delta + (M + c)^\delta - 2c^\delta}. \tag{5}$$

The Champernowne distribution is a generalization of the logistical distribution. Observe that $T(M) = 0.5$ meaning that we estimate M as the median, then c, δ are to be estimated through MLE as described in (Buch-Larsen et al., 2005). In figure 4 we observe the distribution for different parameter choices. We observe that using this transformation our input is mapped to a $[0, 1]$ interval which can be used for the $Beta(3, 3)$ distribution. Observe that the MLE is done through estimating the parameters in regards to the empirical cdf of $\hat{F}(X)$. Basically what this transformation is a fitting of the empirical distribution towards a Champernowne distribution, this step can thereby be seen as an assumption that the cdf is Champernowne distributed. This is not explicitly said in the paper, but something to take note off.

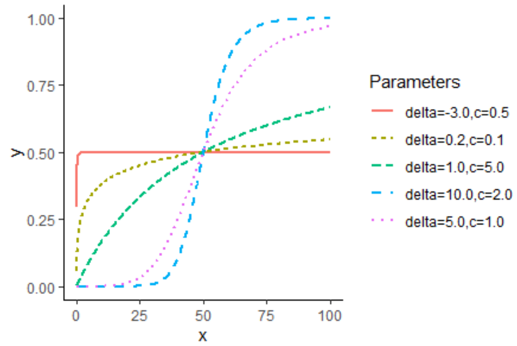

Figure 4: Illustration of Champernowne distribution for different parameter choices.

So now knowing how the transform the data we define our transformed variable as $Y_i = M^{-1}(T(X_i))$. For this variable we will do a kernel estimation with an Epanechnikov kernel K as

$$\hat{F_X}(x) = \frac{1}{n}\sum_{i=1}^{n} K(\frac{x - Y_i}{b}). \tag{6}$$

The goal is from this formulation calculate an optimal bandwidth parameter b. Knowing the distribution being $Beta(3,3)$ we can calculate $\int[f'_Y(x)]^2dx$ and can calculate an optimal bandwidth according to AMISE with

$$\begin{aligned} b^* &= \left(\frac{\int K(t)[1-K(t)]dt}{\int[f'_Y(x)]^2dx(\int t^2k(t)dt)^2}\right)^{\frac{1}{3}} n^{-\frac{1}{3}} \\ &= 3^{\frac{1}{3}}n^{-\frac{1}{3}} \end{aligned} \tag{7}$$

Alemany et al. also suggest to better weight the tails of the distribution minimize towards a weighted mean integrated square error (WISE) they define through the second moment.

$$Wise\{\hat{F_Y}(x)\} = E\{\int[F_Y(x)-\hat{F_Y}(x)]^2x^2dx\} \tag{8}$$

which results in the optimal bandwidth

$$\begin{aligned} b^{**} &= \left(\frac{\int f_Y(x)x^2dx\int K(t)[1-K(t)]dt}{\int[f'_Y(x)]^2dx(\int t^2k(t)dt)^2}\right)^{\frac{1}{3}} n^{-\frac{1}{3}} \\ &= (\frac{9}{7})^{\frac{1}{3}}n^{-\frac{1}{3}}. \end{aligned} \tag{9}$$

But as the goal of this method is to esstimate the VaR_α Alemany et al. do a clever approach and minimize the bandwidth around the α treshhold using the non integrated MSE at $x = x_\alpha = M^{-1}(\alpha)$ by Reiss and Azzalini(Reiss, 1981) approximated as:

$$\begin{aligned} &E\{\hat{F_Y}(x_\alpha) - F_Y(x_\alpha)\} \\ \approx &\frac{F_Y(x_\alpha)[1-F_Y(x_\alpha)]}{n} \\ &- f_Y(x_\alpha)\frac{b}{n}(1-\int K(t)^2dt) \\ &+ b^4\Big(\frac{1}{2}f'_Y(x_\alpha)\int t^2k(t)dt\Big)^2 \end{aligned} \tag{10}$$

Observe that this is using the taylor expansion of our expression. Using this specific MSE value one attains the optimal bandwidth as

$$\begin{aligned} b_\alpha^{clas} &= \left(\frac{f_Y(x_\alpha)\int K(t)[1-K(t)]dt}{(f'_Y(x_\alpha)\int t^2k(t)dt}\right)^{\frac{1}{3}} n^{-\frac{1}{3}} \\ &= (\frac{9}{35}\frac{m(x_\alpha)}{\frac{1}{25}m'(x_\alpha)})^{-\frac{1}{3}}n^{-\frac{1}{3}}. \end{aligned} \tag{11}$$

This is the optimal bandwidth at the point $x = x_\alpha$ according to our assumption that our cdf is approximately Champernowne distributed and using an Epanechnikov kernel. Observe that since the $Beta(3,3)$ being limited between $[0,1]$ our kernel estimate does not integrate to 1. To adjust for this, Alemany et al. do a boundary condition of cutting off the function on $[0,1]$ and normalize

$$\hat{F_Y}(x) = \frac{\bar{F_Y}(x)}{\int_0^1 \bar{F_Y}(x)dx}, x\in[0,1]. \tag{12}$$

The full algorithm is defined in algorithm 1 and as can be seen it only have a couple of steps.

Algorithm 1 The DTKE algorithm

Require: $\mathbf{X}, n = len(X), \alpha$
  $Z \leftarrow T(X)$
  $Y \leftarrow M^{-1}(Z)$
  $x_\alpha \leftarrow M^{-1}(\alpha)$
  $b_\alpha^{Clas} \leftarrow (\frac{9}{35}\frac{m(x_\alpha)}{\frac{1}{25}m'(x_\alpha)})^{-\frac{1}{3}}n^{-\frac{1}{3}}$
  $\bar{F_Y}(x) = \frac{1}{n}\sum_{i=1}^n K(\frac{x-Y_i}{b_\alpha^{Clas}})$
  $\hat{F}_x = \bar{F_Y}(x)/\int_0^1 \bar{F_Y}(x)dx, x\in[0,1]$
  q=$\hat{F}_X^{-1}(\alpha)$
  Return: $VaR_\alpha = T^{-1}(M(q))$

Note that some of the steps in this algorithm as $Z \leftarrow T(x)$ require MLE estimation of the distribution sample and that calculating the VaR_α uses a simple newton raphson for calculation of the inverse. I will now present the result and discuss if the error made with the beta(3,3) distribution affect the results present in the thesis.

3 Paper results

The paper present it results by comparing the esstimate towards a series of different parametric distributions with known VaR_α. They compare their results through a series of other common measures of the VaR_α as classical kernel esstimation , kernel quantile estimation(Sheather and Marron, 1990) and transformed transformed kernel estimation(Buch-Larsen et al., 2005). To not write out a massive amount of results, I will just provide a concise description of the results using $DTKE$ compared to classic kernel esstimate CKE and TKE being the single transform kernel esstimates in percentages of the results given by the empirical distribution. The sub index w imply optimal bandwidth with respect to the $AWISE$ and sub index x imply optimal bandwidth with respect to the explicit value. In the paper they only use the TKE_w results and not for the explicit single value MSE. The results are calculated on the parametric distribution $W = 0.3LN(0, 0.5)+ 0.7P(1, 1)$ where LN is a lognormal distribution and P the pareto distribution. The resulting estimates as fraction of the empirical cdf value can be seen in table2. As can be classical kernel estimation seem to converge, or is very similar to, the empirical kernel estimate meanwhile $DKTE$ seem to vary from these estimates with TKE being more varied along the axis. They have also similar tables for many distribution, but the conclusion is similar between this results.

VaR_α	$\alpha = 0.95$	$\alpha = 0.99$	$\alpha = 0.999$
$n = 500$			
CKE_w	0.91	1.00	-
CKE_x	0.93	1.00	-
TKE_w	0.49	0.65	
$DTKE_w$	0.89	1.90	-
$DTKE_x$	0.87	1.12	-
$n = 5000$			
CKE_w	0.98	1.00	1.00
CKE_x	0.99	1.00	1.00
TKE_w	0.86	12.04	1.58
$DTKE_w$	0.93	0.98	1.07
$DTKE_x$	0.93	0.88	0.66

Table 2: Resulting esstimates in fractions of the analytical true value.

Analysing the standard deviation, seen in table3, of these method on the same distributions. We observe that $DKTE$ having similar variance to the classical kernel estimate meanwhile TKE has a much lower estimated variance. This is argued that the $Beta(3, 3)$ transformation increase the variance but that it becauseof this should come with a reduced bias.

$STD(VaR_\alpha)$	$\alpha = 0.95$	$\alpha = 0.99$	$\alpha = 0.999$
$n = 500$			
CKE_w	2.786	148.574	-
CKE_x	2.815	148.574	-
TKE_w	2.016	2.543	-
$DTKE_w$	2.629	186.024	-
$DTKE_x$	2.626	149.128	-
$n = 5000$			
CKE_w	0.828	30.613	495.937
CKE_x	0.829	30.615	495.937
TKE_w	0.775	1.842	1.885
$DTKE_w$	0.792	28.233	456.538
$DTKE_x$	0.793	27.959	390.843

Table 3: Resulting standard deviation results of esstimate.

Finally looking at the biases of the method, seen in table 4 we so basically how much the estimate differs from the analytical value we observe that the classical kernel estimate have a lower bias compared to the alternative methods. Specifically we see that TKE have a larger bias for large α and that $DTKE$ provides similar biases towards the classical kernel estimate. We should also note that most estimates are positive, meaning that we would overestimate the analyticalVaR_α.

$Bias(VaR_\alpha)$	$\alpha = 0.95$	$\alpha = 0.99$	$\alpha = 0.999$
$n = 500$			
CKE_w	0.257	32.379	-
CKE_x	0.278	32.378	-
TKE_w	0.365	122.615	-
$DTKE_w$	0.855	96.517	-
$DTKE_x$	0.792	60.718	-
$n = 5000$			
CKE_w	0.022	5.031	187.030
CKE_x	0.021	5.047	186.883
TKE_w	0.007	107.893	667.352
$DTKE_w$	0.158	12.144	306.070
$DTKE_x$	0.145	8.081	177.801

Table 4: Resulting biases from the esstimates.

As a conclusion Alemany et al. talk positive about their algorithm. They say it is quick an easy to implement in applied setting. They also mention that it could be used to calculate other properties of heavy-tailed distributions. They also mention it works better the larger sample sizes. Finally they observe that non-parametric properties are very useful for empirical data.

4 Discussion

As we now have gone through the methodology and the a summary of the paper we should discuss the content of the paper. In short you can say that the method presented an extension towards the approach suggested by (Buch-Larsen et al., 2005) where they use an extra transformation with the $Beta(3, 3)$ on top of the Champernowne transformation to further minimize the AMISE,MSE and AWISE. The paper is good at giving background research and other methods in the area but is fairly vague in how the their approach is done. Specifically I find that some assumption and some parts of the methodology very vaguely defined. There is also some problems with how the results are presented. I will thereby highlight some of the main criticism I have with the paper, beginning with the assumptions made.

Assumptions

One of the factors which is not discussed and almost totally dismissed are the assumptions made by this method. Specifically I find that the initial step of fitting a Champernowne distribution through MLE is actually an assumption on the distribution of the data. In this step you actually assume that the cdf is distributed as a inverse Champernowne distribution. which is not a very non-parametric approach towards the problem. It also means that the algorithm can only be defined for strictly positive probability distributions, making it unable to be applied in for example esstimating the VaR_α of the S&P500 as shown in the beginning. You also assume in that the distribution is even, and symmetric around the median M, something which maybe often not happens in terms of heavy tails.

In the following step of using the inverse of an $Beta(3, 3)$ distribution you further assume that the Champernowne transformed data is uniform on $[0, 1]$. Being uniform is follows from the first assumption of the cdf being inverse Champernowne distributed should make the transformed data uniform, so there are actually no additional assumptions made here.. It is reasonable that the $Beta(3, 3)$ inverse transformation could be used to further minimize $\int [f''(x)]^2$, but there is not, or enough, analysis on how much of an improvement this makes. This is refereed to the research done by (Terrell, 1990) butI belive it could beexplained further in the paper. They do compare it with the single transform but this comparison leads to a problem with the thesis in how the results are presented.

Results

A fair criticism of the paper is how the results are presented, in the main results of table 2 they have used the empirical distribution as a benchmark. I can understand why it gives an easy comparison to some base case method, but in terms of accuracy it is impossible to interpret which method is the best. Since they use parametric data they could as well use the analytical values of the VaR_α and quote the results in terms in percentage of the analytical value. It is possible to compare to the analytical value in the bias table 4, but I believe that the results can be explained more clearly. Using the mixture model between the lognormal and pareto distribution and sample size $n = 5000$ we have an analytical value of $VaR_{0.99} = 139.0034, VaR_{0.999} = 699.0001$. As such I actually present the real results for the accuracy of the method plus and minus the standard deviation for the $n = 5000$ case in table5, repeated 100 times, which is to be compared to 2000 used in the paper. I think putting it in a single table make it easier to interpret the results, and put the many different aspects into consideration.

As can be seen this table convey the same information as all tables 243 in the case of $n = 5000$ and $\alpha = 0.99$ and $\alpha = 0.999$. We do observe that we have higher variance, likely due to lower amount of iterations, oth-

VaR_α	$\alpha = 0.99$	$\alpha = 0.999$
$Analytical$	139.0	699.0
CKE_w	159 ± 129	244 ± 205
CKE_x	162 ± 199	487 ± 244
$DTKE_w$	243 ± 100	359 ± 48.0
$DTKE_x$	200 ± 84	561 ± 49.1
TKE_w	195 ± 7.4	762 ± 159

Table 5: Difference in variance from the different distribution.

erwise estimates seem in line to the results found in the study, implying that the results are somewhat consistent. We do observe the strength of each method, and it is pretty easy to compare different methods using this table.

Methodology

An other criticism which relates to the assumption is that there is no clear description of the algorithm in the paper. They discuss the background and previous research in the area, saying that they will use some parts from different papers but not really how. One of these areas is the scaled $Beta(3, 3)$ distribution to $[-1, 1]$ which is not explained and motivated well at all. It took some analysis, but I presume that this choice is based on the paper by (Terrell, 1990). An other criticism is how they end up with final VaR_α, I presume they use the inverse transform of the quantile found, i.e $VaR_\alpha = T^{-1}(M(q))$, but this is not explicitly said in the paper.

Also in many equations they do not mention the definitions of the variables, and what limitation they have. This is specifically for the Champernowne which where they don't mention that the parameters should be strictly positive, something which caused issues when I implemented it. This is also the case of the boundary condition adjustment they make, I had to assume by their wording that they explicitly cut and standardize the cdf on the interval $[-1, 1]$ but more than that I cannot say.

4.1 Conclusion

To conclude the discussion I will begin to discuss their conclusion of the paper where I do not understand their optimism. The results imply that classical kernel methods have lower bias and similar variance even for large α but they still argue that their algorithm have many properties of being non-parametric, and being easily adapted to other tail risk measure makes it better. But classical kernel esstimation is also seemingly non-parametric and performs better I would argue. There algorithm is also not fully adaptable as it require the probability distribution to be strictly positive, which is according to me pretty limiting.

There are some critizism of this paper, and there are some gaps in the assumptions and theory. But overall they do present and motivate why this choice of method may be of interest and I can understand why they may be correct in assuming so. Therefor I will now explore the approach on a real data set to see and get an intuition for what the transformations do to the data and what the implication of using this method would mean if it were to be applied. This is all covered in the following section of exploration.

5 Exploration

Having gone through the paper and having a broad discussion on the paper I want to highlight and analyse the method in an applied setting. I will begin by studying the transformations in the data, get an understanding of what the transformation do, and how they reasonable fit with the data. I will analyse what difference the extra transformation of $Beta(3, 3)$ do in terms of bias and variance. I will then display how the resulting kernel density looks and how the result inverts to estimate the VaR_α. In this case I will use an classic example of insurance data for car claim insurance from Kaggle(Shah, 2017). This was inspired by inspired by (Cooke and Nieboer, 2011), and there where other options such as flooding claims, health bills such as flooding and similar but this data was sadly classified by the institutes.

The data can be seen in figure 5, where we have plotted the Gaussian kernel density and compared it to a normal fit. As can be seen we have a pretty sharp mean around $7000-8000$\$ claim followed by a sharp decline decline and an wiggly little tail. This sharp loss in claims is that many companies put a 10000 maximum for most common type of claims, mostly explaining the sharp shift in claim amount. In terms of low claims there are some exceptions on this part as well.

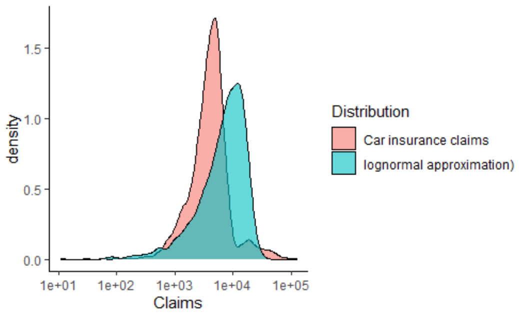

Figure 5: Guassian kernel density of car insurance claims compared to an approximate normal distribution.

Studying the qqplot, seen in figure 6, we see more clearly this shift and heavy tails of our distribution. We again observe that we also have a heavy tail towards low values as well.

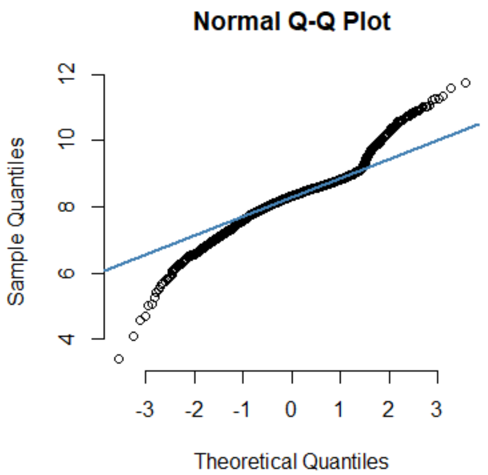

Figure 6: QQplot from claims and fitted lognormal density.

So imagining that we are an insurance company, and that this is our data of car insurance claims. We are considering a new pricing model and what to budget this pricing model towards some unlikely bad scenario. We have a good insight into how many claims we can expect in a year, but are interested in a bad case scenario where these claims are suprisingly large. In our case we want to budget towards the 1% percentile of worst case scenario of insurance claims. Budgeting towards this claim we only have a 1% risk of loosing money in paying out claims assuming there are no other risks. We are therefor interested in calculating the VaR_α with $\alpha = 0.99$ to get the claim amount for this worst case scenario. To do so we want to evaluate the VaR_α through a series of different methods and understand the implications of using the different methods.

In a practical situation it is not uncommon for insurance companies to put the roof of claim amount for this set value of VaR_α. This to eliminate the risk of massive claim amounts leading to losses according to the pricing model. If doing this we are only running the risk of number of claims, which in most scenarios is preferable. I will now go through the $DTKE$ estimation of the VaR_α looking at how well our assumptions are towards the data. We will begin by looking at the Champernowne transformation.

5.1 Champernowne transformation

In figure 7 we see the MLE fitted Champernowne distribution fitted towards the empirical distribution of our claim amounts. As can be seen we have a pretty good curve fitting, having good fit for majority of values around the $0-5000\$$ amount but then we have a worse fit for the higher claim, less dense, part of the data. This is generally a problem I found with the method, usually exceptional events, or heavy tail events are rare, and fitting the data through an equal weighted MLE the Champernowne approximation is often incorrect in these areas compared towards the more dense areas. As can be seen here it would imply that the high tail risk is under-estimated by the assumption, which may not be good for calculating the VaR_α

Looking at the resulting density in figure 8 we observe that the fit is not uniform distribution, an important property of the Champernowne assumption. This is also a usual problem I found with the method, and again we see for the less dense areas that we would underestimate the tail of the data. In comparison of the bandwidth between b_w and b_x we observe that b_x is less smooth, mainly since it is a less

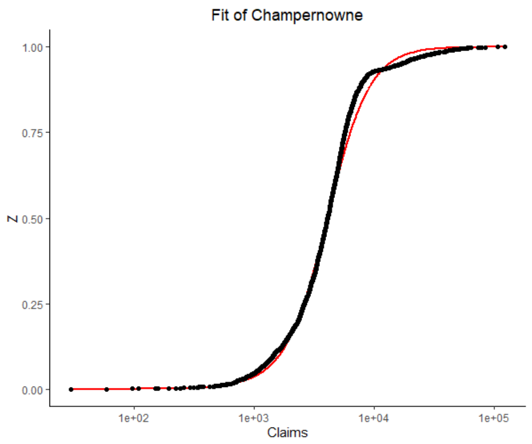

Figure 7: MLE fitted Champernowne distribution compared to empirical data.

stable estimate.

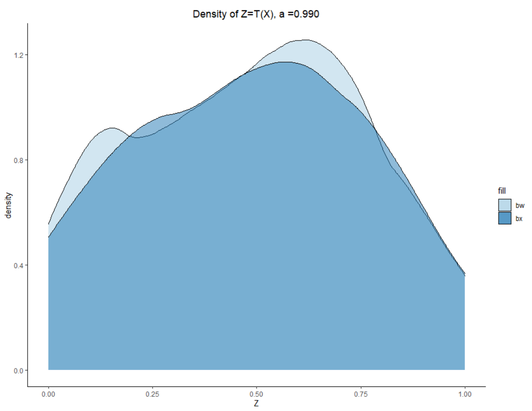

Figure 8: Distribution of the MLE fitted Champernowne transformation on the data with different bandiwths for $DTKE_w$ and $DTKE_x$ for $\alpha = 0.99$.

To conclude here we have that if we where to use the TKE method we would base that on the input seen here, take the 99% percentile and inverse that value to get our VaR_α. As been seen following the results this would likely underestimate the VaR_α.

5.2 Beta transformation

Using the approach suggested by Alemany et al. of using an additional $Beta(3, 3)$ transform we see in figure 9 the resulting densities. As can be seen it is not very symmetric and do not seem to follow the $Beta(3, 3)$ distribution fully, but this is the fault of Z not being uniform and not on any incorrect assumption here. You also observe that in this case b_w and b_x provide very similar densities and that the bandwidth likely is very similar.

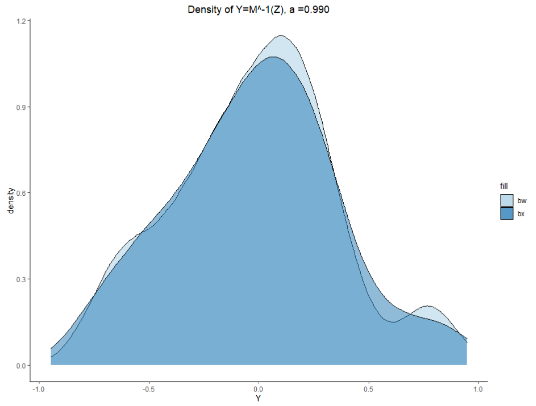

Figure 9: Distribution of the $Beta(3, 3)$ transformation on the Champernowne transformed data using the density for $DTKE_w$ and $DTKE_x$ with $\alpha = 0.99$.

In terms of how using cdf of this distribution to calculate the VaR_α we get that the 99% percentile equals $Var_\alpha = 0.988$ using the b_w bandwidth and $Var_\alpha = 0.992$ using the b_x bandwidth. Compare this value to the percentile we would have gotten from the Champernowne distribution which would have been $Var_\alpha = 0.998$ we have that this transformation adjusts the underestimation of the Champernowne transformation. Even if the difference is small, it does make a big difference once the inverse transformations are used, which will presented now in the final evaluation.

5.3 Evaluation on actual data

In table 6 we have the resulting estimate of the VaR_α for $\alpha = 0.99$ where we present the mean plus minus the standard deviation of our estimate. The variance and mean was evaluated by participating the data in 5 parts and then evaluate each method on each of these. Generally we see that classical kernel estimate a much larger VaR_α compared to the transformation methods. This follows from the analysis said based on the fit of the transformations should result in a lower esstimate. We also observe that the transformation methods are more stable, resulting in a much smaller variance estimate compared to the CKE.

We should mention that the normal approxi-

Method	Esstimate
CKE_w	37129 ± 24399
CKE_x	47338 ± 11850
TKE	18147 ± 1407
$DTKE_w$	25280 ± 6673
$DTKE_x$	24748 ± 5237

Table 6: Resulting esstimate of the VaR_α for $\alpha = 0.99$.

mation, which is used in the CKE method, is not a good approximation here, this is likely why we see these estimates being extremely high with the CKE_w actually estimating somewhere around the maximum value of the data. In regards to TKE we observe that it is much lower than the classical methods, this is likely an undershoot since the fit, as said, is under estimating large values, leading to a smaller inverse of high percentiles. In the case of the $DTKE$ we observe that both choices of bandwidth are very similar and that the resulting estimate compensate the possible underestimate of the TKE.

It is not possible to determine the best algorithm based on this results since the true value is unknown, but we should discuss these results in terms of the implications for our example.

5.4 Implications

So back to our insurance company and where we would like to put the VaR_α to calculate the next pricing model. Based on the different choice of methods we would get vastly different levels of pricing. If you assume that this is our only source of cost, budgeting for the $DKTE$ estimates would imply that we would budget towards half the cost if using the CKE_w approach. This would imply that our insurance would need to be priced at 2 times the level using CKE_w compared to the $DKTE$. The general problem with these approaches is that we do not know if they underestimate or overestimate the true value of the VaR_α , meaning that we don't know if we take on more risk or less risk using these methods. Based on our fit, it seems likely that the transformation methods undershoot the true VaR_α, even if $DTKE$ adjusts somewhat for this. But this would imply that we would underestimate our risk, which is not a good thing if we want to be cautious. In general we want to overestimate the VaR_α so we assume higher risk than what we take on, this way our margin will be of our calculations. In many ways the classical method is reasonable as we are pretty sure it overestimate the true VaR_α which as a consequence means that we would have a positive balance based on our budget.

In the end the choice of method has very different implications. In an applied setting we would usually look to model claims in more detail, putting other covariates into consideration and model the claims in higher detail.. We should also not take the VaR_α of a single claim and multiply that with the amount of claims n, but instead study the distribution of total claims and calculate the VaR_α on that distribution. This is more sensible in many ways, but would not serve as a good example of the method. Even more sensible would be to set a roof on the claims, removing any risk associated with claims payed out. This approach was chosen as demonstration for how our method could be applied in a real situation.

References

Ramon Alemany, Catalina Bolancé, and Montserrat Guillen. 2013. A nonparametric approach to calculating value-at-risk. *Insurance: Mathematics and Economics*, 52(2):255–262.

Catalina Bolancé, Mercedes Ayuso, and Montserrat Guillén. 2012. A nonparametric approach to analyzing operational risk with an application to insurance fraud. *The Journal of Operational Risk*, 7(1):57.

Tine Buch-Larsen, Jens Perch Nielsen, Montserrat Guillén, and Catalina Bolancé. 2005. Kernel density estimation for heavy-tailed distributions using the champernowne transformation. *Statistics*, 39(6):503–516.

Roger M Cooke and Daan Nieboer. 2011. Heavy-tailed distributions: Data, diagnostics, and new developments. *Resources for the Future Discussion Paper*, (11-19).

Bruce M Hill. 1975. A simple general approach to inference about the tail of a distribution. *The annals of statistics*, pages 1163–1174.

Bruce L Jones and Ričardas Zitikis. 2007. Risk measures, distortion parameters, and their empirical estimation. *Insurance: Mathematics and Economics*, 41(2):279–297.

Thomas Maillart and Didier Sornette. 2010. Heavy-tailed distribution of cyber-risks. *The European Physical Journal B*, 75(3):357–364.

Alexander J McNeil. 1998. Statistical analysis of extreme values: From insurance, finance, hydrology and other fields.

R-D Reiss. 1981. Nonparametric estimation of smooth distribution functions. *Scandinavian Journal of Statistics*, pages 116–119.

Bunty Shah. 2017. Auto insurance claims data. https://www.kaggle.com/buntyshah/auto-insurance-claims-data.

Simon J Sheather and James Stephen Marron. 1990. Kernel quantile estimators. *Journal of the American Statistical Association*, 85(410):410–416.

George R Terrell. 1990. The maximal smoothing principle in density estimation. *Journal of the American statistical association*, 85(410):470–477.