Maximum Kernel Likelihood Estimation

Thomas Jaki
Department of Mathematics and Statistics
Lancaster University
Lancaster LA1 4YF
United Kingdom

R. Webster West
Department of Statistics
Texas A&M University
College Station, TX 77843
USA

This is a preprint of an article accepted for publication in the
Journal of Computational and Graphical Statistics
http://www.amstat.org/publications/jcgs/index.cfm?fuseaction=main
Abstract

We introduce an estimator for the population mean based on maximizing likelihoods formed by parameterizing a kernel density estimate. Due to these origins, we have dubbed the estimator the maximum kernel likelihood estimate (mkle). A speedy computational method to compute the mkle based on binning is implemented in a simulation study which shows that the mkle at an optimal bandwidth is decidedly superior in terms of efficiency to the sample mean and other measures of location for heavy tailed symmetric distributions. An empirical rule and a computational method to estimate this optimal bandwidth are developed and used to construct bootstrap confidence intervals for the population mean. We show that the intervals have approximately nominal coverage and have significantly smaller average width than the standard $t$ and $z$ intervals. Lastly, we develop some mathematical properties for a very close approximation to the mkle called the kernel mean. In particular, we demonstrate that the kernel mean is indeed unbiased for the population mean for symmetric distributions.

**Keywords:** Bandwidth; Binning; Kernel density estimation; Maximum likelihood estimation;
1 Introduction

? and ? were the first to introduce the kernel density estimator (kde)

\[ \hat{f}(x|h, X_1, \ldots, X_n) = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{x - X_i}{h} \right), \]

from a random sample, \(X_1, \ldots, X_n\), as an estimator of the unknown underlying probability density function (pdf), \(f(x)\). In the above formulation, \(K(.)\) is the kernel function, usually assumed to be a probability density function that is symmetric about 0, and \(h\) is the bandwidth. Numerous authors (e. g. ?) have studied the properties of the kde. In particular, much work has been done on choosing an optimal bandwidth for the kde. In this regard, it is well known that choosing a bandwidth that is too small results in a kde that is too noisy while choosing the bandwidth too large results in an oversmoothed estimator.

In this note, we explore the use of a parameterized version of the kde to develop an estimated likelihood for the location model where \(f(x|\theta) = f_0(x - \theta)\). In this model, we assume that the mean associated with \(f_0\) is zero which of course implies that the mean of \(X_i\) is \(\theta\). For this location model, the likelihood for \(\theta\) given the random sample is

\[ L(\theta|X_1, \ldots, X_n) = \prod_{i=1}^{n} f(X_i|\theta) = \prod_{i=1}^{n} f_0(X_i - \theta). \]

Given a value of \(\theta\), the kde of the unknown root density, \(f_0(.)\), is simply \(\hat{f}_0(.) = \hat{f}(., h, X_1 - \theta, \ldots, X_n - \theta)\). Substituting this estimate into the above likelihood, results in the cancellation of \(\theta\), so instead we evaluate \(\hat{f}_0\) at the surrogate values \(X_i - \bar{X}\) rather than \(X_i - \theta\) where \(\bar{X}\) represents the sample mean. In other words, we are replacing \(\theta\) with an unbiased estimate. Our estimate of
the likelihood function is then

\[ \hat{L}(\theta|X_1, \ldots, X_n) = \prod_{i=1}^{n} \hat{f}_0(X_i - \bar{X}) = \prod_{i=1}^{n} \hat{f}(X_i - (\bar{X} - \theta)|h, X_1, \ldots, X_n). \] (3)

The last equality in the above expression follows from the substitution of the appropriate quantities into the kde. Since this likelihood estimate is based on a kde, we refer to it as the kernel likelihood. When constructing this kernel likelihood, we are essentially using the data twice; once to estimate the shape of the underlying density with the kde and once to construct the likelihood at the sample values horizontally shifted by \( \bar{X} - \theta \).

Proceeding in a straightforward fashion, we maximize the kernel likelihood over \( \theta \) and refer to the resulting estimate, \( \hat{\theta}_h = \arg \max_\theta \hat{L}(\theta|X_1, \ldots, X_n) \), as the maximum kernel likelihood estimate (mkle) of \( \theta \). In so doing, we are picking the value of \( \theta \) which shifts the observed data in order to maximize contributions to the estimated likelihood. The maximization of the kernel likelihood does not yield a closed form solution for any of the standard kernel functions. In this article, we shall consider the Gaussian kernel, \( K(x) = \exp(-x^2/2)/\sqrt{2\pi} \). For this kernel function, the resulting estimating equation for the mkle from differentiating the log of the kernel likelihood is

\[ \hat{\theta}_h = \frac{1}{n} \frac{1}{n} \sum_{j=1}^{n} \frac{X_i K((X_j - X_i - \bar{X} + \hat{\theta}_h)/h)}{\sum_{k=1}^{n} K((X_j - X_k - \bar{X} + \hat{\theta}_h)/h)}. \] (4)

Based on this rather complicated form, we show that \( \lim_{h \to 0} \hat{\theta}_h = \bar{X} \) and \( \lim_{h \to \infty} \hat{\theta}_h = \bar{X} \) in Lemma ?? in Appendix ??, so that for limiting bandwidths the mkle takes on the value of the sample mean. We chose to write the kernel likelihood only as a function of \( \theta \), since maximization over both \( \theta \) and \( h \) simultaneously yields the \( h = 0 \) solution which places a point mass of \( \frac{1}{n} \) on top of each
data point. Hence, the bandwidth parameter, $h$, is viewed as a tuning parameter in the context of estimating $\theta$. In the remainder of this manuscript, we will discuss a method for fast computation of $\hat{\theta}_h$ (Section ??). The efficiency of the estimator as a function of $h$ will be explored in Section ?? for a variety of symmetric distributions, the maximal gains in efficiency at an optimal bandwidth will be discussed and compared to other commonly used location estimates. In Section ?? two methods to select the optimal bandwidth are proposed and evaluated. The widths of bootstrap confidence intervals for the population mean based on these selection schemes are then compared to standard confidence interval approaches. We also develop and briefly explore an approximation to the mkle called the kernel mean which has a nice intuitive interpretation.

2 Computation of the mkle

The basic method described in the previous section can be easily implemented using standard optimization routines available for R (?). One simply needs to code the kernel likelihood function as shown above and then pass it to an optimization routine such as `optim` for maximization. To use standard optimization routines, a starting value for $\theta$ is required. Given the formulation of the mkle and its behavior for limiting bandwidths, we use $\bar{X}$ as our starting value. As an initial test of the methodology, we analyzed the SAT scores of 100 randomly selected students at Cleveland State University from 2000-2004. The kde is shown in Figure ??.

Figure 1: Kernel density estimate for the SAT scores of students at Cleveland State University from 2000-2004.
To better understand the shape of the kernel log likelihood function over different bandwidths, Figure ?? displays the kernel log likelihood versus $\theta$ over four increasingly larger values of $h$. The solid line represents $\bar{X}$ and the dashed line represents the mkle at the value of $h$. For larger bandwidths, the kernel likelihood is a unimodal function with a clear optimal value. The kernel likelihood at $h = 2$ however is bumpy, but it does appear that $\bar{X}$ is a good starting value as the likelihood is maximized in its neighborhood. The value of $h = 193.1$ was considered because it provided the mkle which was most different from the sample mean for the SAT data set. For any $h$ value the mkle is larger than $\bar{X}$ for this dataset.

While a standard approach to optimize the likelihood works well for a single small data set, it proved to be excruciatingly slow for conducting the simulation studies that follow or even a single computation with a large data set. The issue with this approach is that the density estimate must be recomputed numerous times during the optimization. To decrease computation time, we suggest an alternative approach. We begin by computing the density estimate once over an equidistant grid of points covering the range of the data using fast Fourier transforms as in ?? . We then compute the kernel likelihood function by evaluating this density at data points shifted based on each value of $\theta$. The density is evaluated by interpolating between the grid points containing the shifted values.

The kernel likelihood is then maximized using a grid search algorithm implemented in R with the
function `optimize` which uses a combination of golden section search and successive parabolic interpolation. The range of the grid search is restricted to be within two bandwidths of the sample mean as the density outside of 2 bandwidths is almost zero with the Gaussian kernel. This alternative approach yielded values that were identical to the fourth decimal place to the initial approach in a number of test cases if at least $2^{11}$ grid points were used. Even though it is slightly slower than the standard approach for sample sizes below 50, the new approach is several magnitudes faster for larger sample sizes.

3 Efficency of the mkle

We now investigate the properties of the mkle via simulation. In particular, the properties of $\hat{\theta}_h$ as an estimator for $\theta$ are compared to the sample mean and other measures of location. We define the efficiency for $\hat{\theta}_h$ relative to the sample mean as $
abla F_h = \frac{MSE(\bar{X})}{MSE(\hat{\theta}_h)}$, as a function of the tuning parameter $h$, where $MSE(.)$ is the mean squared error of the estimator in parenthesis. So, in this case, efficiencies greater than one indicate better performance of the mkle. We report results for the standard double exponential, $t$-distribution with 5 degrees of freedom ($t(5)$), standard logistic and standard normal distributions with each distribution centered at $\theta = 0$. These distributions were selected to represent a variety of different tail behaviors with the normal distribution providing light tails and the remaining distributions providing heavier tails. More distributions were also considered in our study with some further comments in the Discussion section. For each of the sample sizes of 10, 100, and 1000, the mkle was computed for 100,000 different samples from the distribution at each value of $h$ over a grid from 0 to 10 by 0.2. The results show considerable gains in efficiency can be obtained by the mkle for heavy tailed distributions across a wide
The maximum gains at the optimal $h$, that is the bandwidth that gives the largest efficiency for $\hat{\theta}_h$, are up to 40% for the double exponential distribution, up to 20% for the $t$-distribution, and roughly up to 10% for a logistic distribution. For the standard normal distribution, the efficiency of the mkle is one or close to it for the majority of the bandwidths considered with a loss in efficiency of only 8% at the worst possible choice of $h$. For the distributions with heavier tails, the efficiency of the mkle increases with sample size with values at $n = 10$ that are well below those at $n = 100$ and $n = 1000$. When breaking down the MSE of the mkle, the MSE appears to be entirely composed of variance with no significant bias for any of the distributions considered. The sampling distribution of the mkle also appeared to be approximately normal for every sample size and every bandwidth.

In terms of the optimal bandwidth, a couple of unique characteristics can be seen. One striking feature shown in the graphs is the wide range of optimal or near optimal bandwidths. For the logistic distribution, for example, any $h$ between 1.5 and 4 yields a near optimal gain in efficiency when $n = 100$. For all heavy tailed distributions, the optimal $h$ falls between 2 and 4 with very little sensitivity to sample size. Upon closer examination, we see that the optimal choice for $h$ appears to be roughly twice the standard deviation of the heavy tailed distributions. The optimal $h$ values are in the neighborhood of 2.83, 2.58 and 3.63 for the double exponential distribution, the $t$-distribution with 5 degrees of freedom and the logistic distribution, respectively. For the normal distribution, the mkle efficiency at a bandwidth of twice the standard deviation is approaching one with a slightly larger bandwidth being more desirable. The optimal bandwidth for the mkle therefore is very different to the optimal bandwidth in density estimation on two levels. Firstly, unlike in density estimation, the optimal $h$ does not appear to depend substantially on sample size and
secondly, the ideal bandwidths for the mkle are almost absurdly large compared to those of the kde. For a $t$-distribution with 5 degrees of freedom and $n = 100$, for example, the optimal bandwidth of the kde is 0.425 while it is 2.582 for the mkle. Both of these differences can perhaps be attributed to the goal of the mkle procedure. While in density estimation the goal is to accurately depict the underlying density in detail, the aim of the mkle is to capture the broad location feature of the density.

Figure 3: The plots display $EFF_h$ over $h$ for different distributions for 100,000 simulations.

To further establish the quality of the mkle as an estimator of $\theta$, Tables ?? - ?? compare the efficiencies of commonly used nonparametric measures of center for the distributions considered above. In addition to the mkle, the estimators investigated, based on ?, are the trimmed mean ($\bar{X}_\alpha$), the Hodges-Lehman (HL) estimator (?), Hoggs $T_1$ (?) and the outmean ($\bar{X}^*_0.25$). The median is omitted since it is a special case of the trimmed mean. We used smaller sample sizes of $n = 10, 25, 50$ and 100 for these comparisons due to computational limitations associated with the estimators used. The efficiency values were estimated based on evaluating the estimators for 100,000 separate samples from each distribution. The efficiencies displayed in the tables are relative to the estimator with smallest MSE at that sample size. The best estimator then has an efficiency of 1 in each case. For the trimmed mean, the value, $\alpha^*$, shown in parentheses is the trimming percentage that resulted in the maximal efficiency as estimated via simulation. The equivalent approach was used for the mkle with efficiency results given at estimates with the optimal $h$ shown parenthetically.

For the distributions previously considered, none of the measures is universally superior to the
others across all distributions. For the double exponential distribution, the trimmed mean proved to be uniformly better than the remaining estimators at all sample sizes with the Hodges-Lehman estimator and the mkle finishing second and third, respectively. The advantage of the trimmed mean over the other location measures increases as sample size increases. For the $t(5)$ distribution, the trimmed mean and the Hodges-Lehman estimator are neck and neck across the range of sample sizes with the mkle following closely in third. For the logistic distribution, the mkle has the best efficiency uniformly across all sample sizes with the trimmed mean and the Hodges-Lehman estimator following closely behind. For the light tailed normal distribution, it is known that the sample mean has the smallest mean squared error among unbiased estimators. The fact that the mkle has higher efficiency in this case is an artifact of the simulation as the MSE displayed is the smallest across an entire range of bandwidths. The optimal bandwidth is relatively stable across sample size for all but the normal distribution where the noise in estimated efficiency across $h$ dominates the selection of the bandwidth. This result is in accordance with Figure ?? where essentially every bandwidth larger than 3.5 yields an efficiency close to one for the standard normal distribution.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{X}$</td>
<td>0.7047</td>
</tr>
<tr>
<td>$\bar{X}_{\alpha}$</td>
<td>1.0000</td>
</tr>
<tr>
<td>(0.30)</td>
<td>(0.40)</td>
</tr>
<tr>
<td>HL</td>
<td>0.9091</td>
</tr>
<tr>
<td>$T_1$</td>
<td>0.4484</td>
</tr>
<tr>
<td>$\bar{X}_{0.25}$</td>
<td>0.4115</td>
</tr>
<tr>
<td>$\hat{\theta}_h$</td>
<td>0.8498</td>
</tr>
<tr>
<td>(2.6)</td>
<td>(2.4)</td>
</tr>
</tbody>
</table>

Table 1: Simulated efficiencies relative to the estimator with the smallest MSE for double exponential distributed data. Numbers in parenthesis are the optimal trimming percentage for the trimmed mean and the optimal bandwidth for the mkle.
Table 2: Simulated efficiencies relative to the estimator with the smallest MSE for $t(5)$-distributed data. Numbers in parenthesis are the optimal trimming percentage for the trimmed mean and the optimal bandwidth for the mkle.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>n=10</th>
<th>n=25</th>
<th>n=50</th>
<th>n=100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{X}$</td>
<td>0.8539</td>
<td>0.8237</td>
<td>0.7968</td>
<td>0.8040</td>
</tr>
<tr>
<td>$\bar{X}_{\alpha^*}$</td>
<td>1.0000</td>
<td>1.0000</td>
<td>0.9916</td>
<td>1.0000</td>
</tr>
<tr>
<td>$HL$</td>
<td>0.9928</td>
<td>0.9925</td>
<td>1.0000</td>
<td>0.9895</td>
</tr>
<tr>
<td>$T_1$</td>
<td>0.5935</td>
<td>0.5308</td>
<td>0.5087</td>
<td>0.5003</td>
</tr>
<tr>
<td>$\bar{X}_{0.25}^c$</td>
<td>0.5417</td>
<td>0.4847</td>
<td>0.4757</td>
<td>0.4642</td>
</tr>
<tr>
<td>$\hat{\theta}_{h^*}$</td>
<td>0.9604</td>
<td>0.9557</td>
<td>0.9524</td>
<td>0.9561</td>
</tr>
</tbody>
</table>

Table 3: Simulated efficiencies relative to the estimator with the smallest MSE for logistic distributed data. Numbers in parenthesis are the optimal trimming percentage for the trimmed mean and the optimal bandwidth for the mkle.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>n=10</th>
<th>n=25</th>
<th>n=50</th>
<th>n=100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{X}$</td>
<td>0.9283</td>
<td>0.9112</td>
<td>0.9245</td>
<td>0.9068</td>
</tr>
<tr>
<td>$\bar{X}_{\alpha^*}$</td>
<td>0.9825</td>
<td>0.9820</td>
<td>0.9924</td>
<td>0.9905</td>
</tr>
<tr>
<td>$HL$</td>
<td>0.9719</td>
<td>0.9875</td>
<td>0.9985</td>
<td>0.9964</td>
</tr>
<tr>
<td>$T_1$</td>
<td>0.7014</td>
<td>0.6389</td>
<td>0.6314</td>
<td>0.6163</td>
</tr>
<tr>
<td>$\bar{X}_{0.25}^c$</td>
<td>0.6767</td>
<td>0.6260</td>
<td>0.6195</td>
<td>0.6085</td>
</tr>
<tr>
<td>$\hat{\theta}_{h^*}$</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

4 Selecting the bandwidth

The simulation results of the previous section show that the mkle at the optimal bandwidth is a competitive measure of location for the heavier tailed distributions considered. In this section, two simple approaches to select the optimal bandwidth will be discussed. Both methods will be used to construct bootstrap confidence intervals for $\theta$. Their coverage probabilities and average lengths will be compared to methods based on the trimmed mean as well as standard $z$ and $t$ intervals.
Table 4: Simulated efficiencies relative to the estimator with the smallest MSE for standard normal
distributed data. Numbers in parenthesis are the optimal trimming percentage for the trimmed
mean and the optimal bandwidth for the mkle.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Efficiency</th>
<th>n=10</th>
<th>n=25</th>
<th>n=50</th>
<th>n=100</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\bar{X}$</td>
<td>0.9961</td>
<td>0.9960</td>
<td>0.9906</td>
<td>0.9913</td>
<td></td>
</tr>
<tr>
<td>$\bar{X}_{\alpha^*}$</td>
<td>0.9954</td>
<td>0.9937</td>
<td>0.9901</td>
<td>0.9910</td>
<td></td>
</tr>
<tr>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.00)</td>
<td>(0.01)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HL</td>
<td>0.9306</td>
<td>0.9421</td>
<td>0.9515</td>
<td>0.9513</td>
<td></td>
</tr>
<tr>
<td>$T_1$</td>
<td>0.8829</td>
<td>0.8407</td>
<td>0.8369</td>
<td>0.8349</td>
<td></td>
</tr>
<tr>
<td>$\bar{X}_{5,25}$</td>
<td>0.8809</td>
<td>0.8373</td>
<td>0.8405</td>
<td>0.8297</td>
<td></td>
</tr>
<tr>
<td>$\hat{\theta}_{h^*}$</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td>1.0000</td>
<td></td>
</tr>
<tr>
<td>(5.0)</td>
<td>(4.8)</td>
<td>(7.0)</td>
<td>(9.4)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The trimmed mean was chosen as a baseline for comparison given its high level of performance in
the simulation study from the previous section.

The idea of finding the optimal value for a tuning parameter using bootstrap techniques has been
discussed by ? in the context of selecting the optimal trimming percentage for the trimmed mean.

For a range of trimming percentages, $B$ bootstrap resamples are obtained and the trimmed mean
computed. The MSE at a particular $\alpha$ is estimated by summing the squared deviations of each
bootstrapped trimmed mean from the sample mean and dividing by $B$. The trimming percentage
that results in the smallest estimated MSE is used as the optimal trimming percentage, $\alpha^*$. The
bootstrapped trimmed means at this trimming percentage are then used to construct bootstrap
confidence intervals.

We followed the standard method described above for the trimmed mean to estimate the opti-
mal bandwidth for the mkle. Over a range of bandwidths, the mkle is computed for $B$ bootstrap
resamples. The MSE at a particular $h$ is estimated by summing the squared deviations of each
bootstrapped mkle from the sample mean and dividing by $B$. The bandwidth that results in the smallest estimated MSE is used as the optimal bandwidth, $h^*$. Notice that this bootstrap approach ignores the uncertainty in the determination of the optimal tuning parameter when constructing the intervals. Including this uncertainty would require a double bootstrap (see ?) which was not computationally feasible for the simulation study that follows below. On a very fast computer, the single bootstrap computations for a dataset of size 100 take about half an hour to complete, making the even more intensive double bootstrap approach infeasible in practice.

We also consider an alternative to this standard single bootstrap approach. It was pointed out in Section ?? that the optimal bandwidth appears to be roughly twice the standard deviation of the underlying distribution regardless of sample size. We conducted a much more extensive simulation study evaluating this rough approximation for a variety of symmetric distributions shifted and scaled with different means, standard deviations and kurtosis values. These additional simulations confirmed our previous observation. The only significant predictor among those considered was the standard deviation with the coefficient on this predictor being almost exactly two. Thus, the optimal bandwidth can also be estimated as $\hat{h} = 2\hat{\sigma}$, where $\hat{\sigma}$ is the sample standard deviation of the data set. After finding $\hat{h}$, we select $B$ bootstrap resamples for which we compute the mkle at this bandwidth. The set of bootstrapped mkles are then used to construct the bootstrap confidence interval for $\theta$.

Using sample sizes of $n = 10, 25, 50$ and 100 from each of the distributions discussed earlier, we constructed bootstrap Wald and bootstrap percentile intervals (e. g. ?) for each of the methods described above for the trimmed mean and the mkle. A confidence level of 95% was used, and the
number of bootstrap replications was \( B = 1000 \) for each method. Simulating 1000 samples at each sample size, we studied the empirical coverage probabilities and average interval widths shown in Tables ?? through ???. The results for the standard \( z \) and \( t \) intervals are also provided in these tables. Using a standard \( z \) test for proportions, empirical coverage probabilities smaller than 0.939 are considered to be significantly below the nominal level of 0.95.

For the trimmed mean, the Wald intervals are preferred since this method has much better coverage characteristics than the percentile method for every distribution and sample size considered. In this case, the Wald intervals are almost always on target at the desired nominal level with the percentile intervals approaching the nominal level for larger sample sizes. The same preference may be stated for the plug in mkle method with the coverage probabilities for the Wald method uniformly closer to nominal than the percentile method. In this case, both coverage probabilities are slightly below nominal for samples of size 10 and in some cases for samples of size 25. The preference for bootstrap method, however, does not repeat itself for the standard mkle method. While coverage probabilities for the Wald method are closer to nominal than the percentile method for every sample size for the double exponential and \( t(5) \) distributions, the opposite is true for the logistic and normal distributions. Comparing the various approaches, it appears that the standard \( t \) interval has the best coverage probability characteristics followed very closely by the trimmed mean and the mkle methods.

The mkle methods, however, have distinctly smaller average widths than the standard \( z \) and \( t \) intervals as well as the trimmed mean methods for the heavy tailed distributions. This is most surprising for the double exponential distribution where the results from the previous section showed
that the potential gains in efficiency were larger for the trimmed mean. This may be attributed to
the wide range of reasonable choices of bandwidth for the mkle which makes the selection of a de-
sirable $h$ less difficult. For the heavier tailed distributions, the average width of the mkle interval is
roughly 20% smaller than the average width of the standard $t$ interval when $n = 10$. These gains in
average width for the mkle methods diminish somewhat to 10% as sample size increases to $n = 100$.

Comparing the different mkle methods, it appears that the simple plug in rule is more desirable
than the standard bootstrap approach. The plug in rule has better coverage characteristics than
the standard approach with very similar average widths. The plug in rule, therefore, provides a
much faster and more reliable way to estimate the optimal bandwidth while the standard approach
requires many more bootstrap resamples in order to obtain equivalent or worse results.

<table>
<thead>
<tr>
<th>Method</th>
<th>$n=10$</th>
<th>$n=25$</th>
<th>$n=50$</th>
<th>$n=100$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wald $\hat{\theta}_h$</td>
<td>0.921 (1.559)</td>
<td>0.953 (0.986)</td>
<td>0.952 (0.700)</td>
<td>0.953 (0.495)</td>
</tr>
<tr>
<td>Perc $\hat{\theta}_h$</td>
<td>0.901 (1.552)</td>
<td>0.943 (0.985)</td>
<td>0.943 (0.699)</td>
<td>0.950 (0.493)</td>
</tr>
<tr>
<td>Wald $\hat{\theta}_h^*$</td>
<td>0.914 (1.508)</td>
<td>0.929 (0.987)</td>
<td>0.940 (0.695)</td>
<td>0.929 (0.494)</td>
</tr>
<tr>
<td>Perc $\hat{\theta}_h^*$</td>
<td>0.905 (1.504)</td>
<td>0.913 (0.984)</td>
<td>0.935 (0.694)</td>
<td>0.925 (0.493)</td>
</tr>
<tr>
<td>Wald $\bar{X}_\alpha$</td>
<td>0.939 (1.825)</td>
<td>0.949 (1.101)</td>
<td>0.959 (0.786)</td>
<td>0.961 (0.555)</td>
</tr>
<tr>
<td>Perc $\bar{X}_\alpha$</td>
<td>0.882 (1.562)</td>
<td>0.923 (1.040)</td>
<td>0.938 (0.764)</td>
<td>0.940 (0.546)</td>
</tr>
<tr>
<td>$z$-interval</td>
<td>0.916 (1.645)</td>
<td>0.947 (1.077)</td>
<td>0.948 (0.777)</td>
<td>0.946 (0.557)</td>
</tr>
<tr>
<td>$t$-interval</td>
<td>0.952 (1.898)</td>
<td>0.959 (1.135)</td>
<td>0.956 (0.796)</td>
<td>0.951 (0.564)</td>
</tr>
</tbody>
</table>

Table 5: Empirical coverage of different confidence intervals for $\theta$ based on double exponential
distributed data using 1000 simulations. Number in parenthesis is the average interval width.

5 Discussion

In this manuscript, we have introduced a new estimation technique based on kernel density esti-
mation which has been implemented in the R package MKLE (?). Unfortunately, a closed form
expression for our estimator does not exist, so establishing the mathematical properties of the esti-
mator has been difficult. This is true even in terms of large sample properties as the estimator does
not fall nicely into any of the standard classes of asymptotic statistics. In this regard, however,
we have developed an approximate solution to the estimating equation for the mkle via a Taylor
series expansion of the right hand side of equation (5) of order zero about \( \theta_h = \bar{X} \). In so doing, the
approximation becomes

\[
\hat{\theta}_h \approx \frac{1}{n} \sum_{j=1}^{n} \frac{\sum_{i=1}^{n} X_i \exp\left(-\frac{(X_j - X_i)^2}{2h^2}\right)}{\sum_{k=1}^{n} \exp\left(-\frac{(X_j - X_k)^2}{2h^2}\right)}.
\]

(5)
This approximation could also be derived as the first sequential value if one considered solving the estimating equation in an iterative fashion starting at $\bar{X}$. This approximation, which will be referred to as the kernel mean, is particularly interesting since it is the average Nadaraya-Watson smoother (??) applied to each $X_j$. Through numerical studies, we have found that the kernel mean is extremely close to the mkle, and the efficiency profiles for the kernel mean match those of the mkle almost perfectly (see Figure ?? for an example). Furthermore, we show in Lemma ?? of Appendix ?? that the kernel mean is unbiased for $\theta$ when the underlying distribution is symmetric. Recall that we found the same to be true of the mkle in our simulation studies. In the future, closer examination of the kernel mean may provide some insights into the theoretical properties of the mkle. Ironically, the kernel mean approximation is more difficult to compute than the mkle as it requires an $n^2$ computation.

Figure 4: The plot displays $EFF_h$ of the mkle and kernel mean over $h$ for the logistic distribution for 10,000 simulations with $n = 100$.

Another point of interest is the choice of the kernel function. While only results using a Gaussian

Table 8: Empirical coverage of different confidence intervals for $\theta$ based on standard normal distributed data using 1000 simulations. Number in parenthesis is the average interval width.

<table>
<thead>
<tr>
<th>Method</th>
<th>n=10</th>
<th>n=25</th>
<th>n=50</th>
<th>n=100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wald $\hat{\theta}_h$</td>
<td>0.915 (1.178)</td>
<td>0.944 (0.779)</td>
<td>0.948 (0.557)</td>
<td>0.959 (0.398)</td>
</tr>
<tr>
<td>Perc $\hat{\theta}_h$</td>
<td>0.913 (1.168)</td>
<td>0.941 (0.774)</td>
<td>0.942 (0.554)</td>
<td>0.955 (0.396)</td>
</tr>
<tr>
<td>Wald $\hat{\theta}_{h^*}$</td>
<td>0.888 (1.093)</td>
<td>0.924 (0.722)</td>
<td>0.925 (0.516)</td>
<td>0.945 (0.370)</td>
</tr>
<tr>
<td>Perc $\hat{\theta}_{h^*}$</td>
<td>0.905 (1.136)</td>
<td>0.937 (0.753)</td>
<td>0.937 (0.539)</td>
<td>0.955 (0.388)</td>
</tr>
<tr>
<td>Wald $\bar{X}_{\alpha^*}$</td>
<td>0.924 (1.313)</td>
<td>0.946 (0.803)</td>
<td>0.956 (0.561)</td>
<td>0.941 (0.394)</td>
</tr>
<tr>
<td>Perc $\bar{X}_{\alpha^*}$</td>
<td>0.885 (1.126)</td>
<td>0.937 (0.758)</td>
<td>0.956 (0.544)</td>
<td>0.942 (0.387)</td>
</tr>
<tr>
<td>$z$-interval</td>
<td>0.918 (1.210)</td>
<td>0.942 (0.777)</td>
<td>0.946 (0.549)</td>
<td>0.954 (0.392)</td>
</tr>
<tr>
<td>$t$-interval</td>
<td>0.949 (1.397)</td>
<td>0.956 (0.819)</td>
<td>0.949 (0.563)</td>
<td>0.955 (0.396)</td>
</tr>
</tbody>
</table>
kernel are presented in this note, we have also investigated a variety of standard kernel functions. For all kernel functions studied, the profiles of efficiency over \( h \) showed a significant gain in efficiency and generally looked very similar to the one’s presented here (see Figure ?? for the profiles for the biweight kernel). The potential gains with finite support kernels, however, generally appeared to be lower than the maximum gains in efficiency for the Gaussian kernel. Furthermore, the finite support results in a discontinuous kernel likelihood which makes the maximization of the kernel likelihood much more difficult than for a Gaussian kernel. Lastly, the optimal bandwidths for finite support kernels are much larger than for the Gaussian kernel.

Figure 5: The plots display \( EFF_h \) over \( h \) for different distributions for 100,000 simulations using a biweight kernel.

The successful combination of kernel density techniques with the maximum likelihood approach bring other possible applications to mind. The mkle proves most valuable for symmetric distributions, yet we have not taken the symmetric nature of the underlying density into account when constructing the kde. Different authors, such as ?, present a symmetric kde which assumes a given point of symmetry around which every observed data point is mirrored to construct the kde. When combining this symmetric kde with the kernel likelihood approach to estimate the point of symmetry, our initial results indicate even further gains in efficiency due to the effectively doubled sample size.

In this seminal manuscript on the mkle method, we would be somewhat remiss not to mention potential applications outside of simple mean estimation. Applying the mkle approach to the area
of regression also looks quite promising based on our initial research. While we have considered a simple location model in this manuscript, consider the simple linear regression equation, \( Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \). The error terms \( \epsilon_i \) are assumed to be independent with an unspecified symmetric distribution, \( f(\cdot) \), for which the \( \text{kde} \)

\[
\hat{f}(\epsilon) = \frac{1}{nh} \sum_{i=1}^{n} K \left( \frac{\epsilon - \epsilon_i}{h} \right),
\]

(6)

that is based on the unobserved error terms, can be used. By replacing the unobserved error terms with the values \( Y_i - \beta_0 - \beta_1 X_i \) and constructing the product of \( \hat{f} \) over these same values, our kernel likelihood function becomes

\[
\hat{L}(\beta_1|Y_1,\ldots,Y_n, X_1,\ldots, X_n) = \prod_{i=1}^{n} \hat{f}(Y_i - \beta_0 - \beta_1 X_i).
\]

(7)

It is interesting to note that the intercept term from the regression model disappears from the kernel likelihood so alternative methods would be required to estimate this model parameter. Using methods similar to those provided above, however, one can maximize the kernel likelihood to estimate the slope parameter. Once again, preliminary simulations have shown great increases in efficiency over the standard ordinary least squares estimator for almost all bandwidths over a wide range of heavy tailed error distributions. Figure ?? shows the efficiency of the mkle for the slope relative to least squares estimate of the slope for a sample size of 100 with the error distribution taking the form of the distributions considered above.

Figure 6: The plot displays the efficiency of the regression mkle against the ordinary least squares estimate over \( h \) for different distributions for \( n = 100 \) based on 10,000 simulations.
The plot shows that there are potential losses in efficiency for very small bandwidths but also large gains in efficiency for larger bandwidths. As the tails of the distribution become heavier, the gains in efficiency become more pronounced. As with the simple mean estimate, the efficiency of the mkle seems to converge to one from above as the bandwidth goes to infinity. Unlike the mkle in simple mean estimation, however, the optimal bandwidth in this case does not appear to be close to two standard deviations for the error distribution. Determining the optimal bandwidth for this and other more complex models is a matter for future research.

A Lemmas and proofs

Lemma 1

\[
\lim_{h \to 0} \hat{\theta}_h = \bar{X} \quad \text{and} \quad \lim_{h \to \infty} \hat{\theta}_h = \bar{X}
\]

\[\square\]

Proof: First we can rewrite the estimating equation as

\[
\theta = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} X_i \exp \left( -\frac{(X_j - X_i - \bar{X} + \theta)^2}{2h^2} \right) 
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} X_i \sum_{j=1}^{n} \exp \left( -\frac{(X_j - X_i - \bar{X} + \theta)^2}{2h^2} \right)
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} X_i \sum_{j=1}^{n} w(X_1, \ldots, X_n, \theta, h).
\]
If we now consider

\[
\lim_{h \to 0} w(X_1, \ldots, X_n, \theta, h) = \lim_{h \to 0} \frac{1}{\sum_{i=1}^{n} \exp \left( -\frac{(X_j - X_i - \bar{X} + \theta)^2}{2h^2} \right)}
\]

\[
\exp \left( -\frac{(X_j - X_i - \bar{X} + \theta)^2}{2h^2} \right)
\]

\[
= 1 + \lim_{h \to 0} \sum_{k \neq i} \exp \left( \frac{-(X_j - X_k - \bar{X} + \theta)^2 - (X_j - X_i - \bar{X} + \theta)^2}{2h^2} \right)^{-1}
\]

\[
= \begin{cases} 
1 & \text{if } (X_j - X_k - \bar{X} + \theta)^2 - (X_j - X_i - \bar{X} + \theta)^2 > 0 \forall k \neq i \\
0 & \text{otherwise}
\end{cases}
\]

it follows that \( \sum_{j=1}^{n} \lim_{h \to 0} \frac{\exp \left( -\frac{(X_i - X_j - \bar{X} + \theta)^2}{2h^2} \right)}{\sum_{k=1}^{n} \exp \left( -\frac{(X_j - X_k - \bar{X} + \theta)^2}{2h^2} \right)} = 1 \) since the summand is only one iff \( X_i = \min(X_1, \ldots, X_n) \) and hence \( \hat{\theta}_h \xrightarrow{h \to 0} \bar{X} \). Also,

\[
\lim_{h \to \infty} \theta = \frac{1}{n} \sum_{i=1}^{n} X_i \sum_{j=1}^{n} \lim_{h \to \infty} \exp \left( -\frac{(X_j - X_i - \bar{X} + \theta)^2}{2h^2} \right)
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} X_i \sum_{j=1}^{n} \frac{1}{n}
\]

\[
= \bar{X}.
\]

\[\Box\]

**Lemma 2** Let \( \tilde{\theta}_h = \frac{1}{n} \sum_{j=1}^{n} \frac{\sum_{i=1}^{n} X_i \exp \left( -\frac{(X_j - X_i)^2}{2h^2} \right)}{\sum_{k=1}^{n} \exp \left( -\frac{(X_j - X_k)^2}{2h^2} \right)} \) be the kernel mean. If \( E(X) = \mu \) and \( f(x) \) is symmetric about \( \mu \), then

\( E(\tilde{\theta}_h) = \mu. \)

\[\Box\]

**Proof:** The kernel mean can be written as \( \frac{1}{n} \sum_{j=1}^{n} S_j \) where \( S_j \) is the Naradaya-Watson smoother
evaluated at $X_j$, defined as

$$S_j = \frac{\sum_{i=1}^{n} X_i K(X_j - X_i)}{\sum_{i=1}^{n} K(X_j - X_i)} = \frac{X_j + \sum_{i \neq j, i=1}^{n} X_i K(X_j - X_i)}{1 + \sum_{i \neq j} K(X_j - X_i)}$$

where $K(x) = \exp\left(-\frac{1}{2\theta} x^2\right)$ in this application. Since the collection $\{S_1, \ldots, S_n\}$ are identically distributed (but not independent), $E(\tilde{\theta}_h) = E(S_j)$. Without loss of generality we will let $j = 1$ and consider the difference

$$E(\tilde{\theta}_h) - \mu = E(S_1 - x_1)$$

$$= E(E(S_1 - X_1 | X_1))$$

$$= \int_{-\infty}^{\infty} g(x) f(x) dx$$

where $g(x) = E(S_1 - X_1 | X_1 = x) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \sum_{i=2}^{n} \frac{(x_i - x)}{1 + \sum_{i=2}^{n} K(x_i - x)} f(x_2) \cdots f(x_n) dx_2 \cdots dx_n$. We then find

$$g(\mu + \epsilon) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \sum_{i=2}^{n} \frac{(x_i - \mu - \epsilon)}{1 + \sum_{i=2}^{n} K(\mu + \epsilon - x_i)} f(x_2) \cdots f(x_n) dx_2 \cdots dx_n$$

let $z_i = x_i - \mu$

$$= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \sum_{i=2}^{n} \frac{(z_i - \epsilon)}{1 + \sum_{i=2}^{n} K(\epsilon - z_i)} f(z_2 + \mu) \cdots f(z_n + \mu) dz_2 \cdots dz_n$$
and similarly

\[ g(\mu - \epsilon) = \int_{-\infty}^\infty \ldots \int_{-\infty}^\infty \frac{\sum_{i=2}^{n}(x_i - \mu + \epsilon)K(\mu - \epsilon - x_i)}{1 + \sum_{i=2}^{n} K(\mu - \epsilon - x_i)} f(x_2) \ldots f(x_n) dx_2 \ldots dx_n \]

by symmetricity of \( K(x) \) about 0

\[ = \int_{-\infty}^\infty \ldots \int_{-\infty}^\infty \frac{\sum_{i=2}^{n}(x_i - \mu + \epsilon)K(\epsilon -(\mu - x_i))}{1 + \sum_{i=2}^{n} K(\epsilon -(\mu - x_i))} f(x_2) \ldots f(x_n) dx_2 \ldots dx_n \]

let \( z_i = \mu - x_i \)

\[ = \int_{-\infty}^\infty \ldots \int_{-\infty}^\infty \frac{\sum_{i=2}^{n}(-z_i + \epsilon)K(\epsilon - z_i)}{1 + \sum_{i=2}^{n} K(\epsilon - z_i)} f(\mu - z_2) \ldots f(\mu - z_n) dz_2 \ldots dz_n. \]

Note that the Jacobian for this transformation is \((-1)^{n-1}\) which in absolute value is 1. If we assume now that \( f(x) \) is symmetric about \( \mu \),

\[ g(\mu - \epsilon) = \int_{-\infty}^\infty \ldots \int_{-\infty}^\infty \frac{\sum_{i=2}^{n}(-z_i + \epsilon)K(\epsilon - z_i)}{1 + \sum_{i=2}^{n} K(\epsilon - z_i)} f(\mu + z_2) \ldots f(\mu + z_n) dz_2 \ldots dz_n \]

\[ = -g(\mu + \epsilon), \]

and hence \( g(.) \) is an odd function about \( \mu \) and therefore

\[ \int_{-\infty}^\infty g(x)f(x)dx = 0 \]

which implies that \( E(\tilde{\theta}_h) = \mu \), completing the proof.